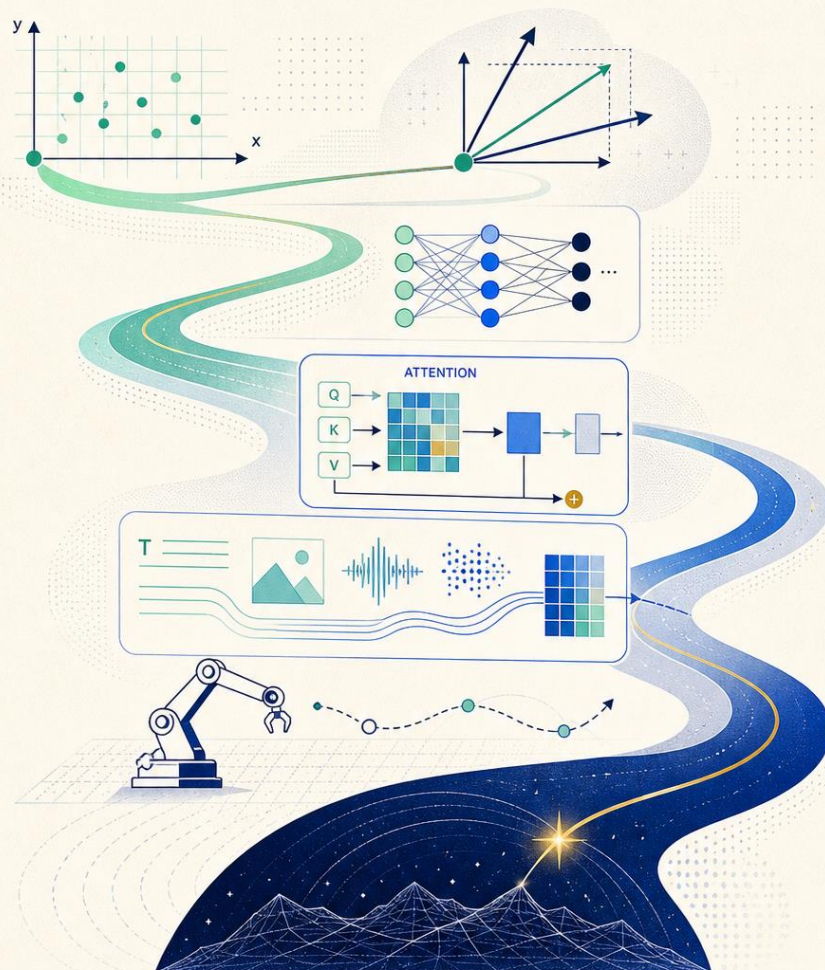


从零开始学AI

普通人与高中生的24个月互动教材

2026.08 Beta



 NEWVAR

第二卷：经典 AI 与机器学习

版本、边界与使用方法

结论先行：这是一套面向普通人与高中生的系统教材，不是速成承诺。当前为 **2026.08 Beta**；只有技术评审、事实核查和领域专家签字全部完成后，才能进入 stable。

项目	内容
版本	2026.08 Beta
卷册	第二卷：经典 AI 与机器学习
本卷章节	24
全书结构	4 卷 · 24 单元 · 96 章
实践资产	96 个互动实验 · 24 个项目 · 288 道练习
正文规模	约 689,651 个汉字
核实日期	2026-08-06
版权	Copyright © 2026 NEWVAR LAB. All rights reserved.

能力终点

本书以独立开发、部署、论文阅读、实验复现、失败分析和研究问题提出为终点能力；不承诺仅靠一本书自动成为世界顶尖研究者。

全面覆盖的可审计边界

“全面”指对主要 AI 知识域、能力框架、工程链路和研究方法进行可审计覆盖，不等于收录互联网上每一页资料。

每章学习契约

- 先看结论、价值和完成后能做什么。
- 完成先修检查，用生活类比获得方向，再主动找出类比失效处。
- 一次只改变一个核心变量；网页提供键盘操作和静态替代说明。
- 按直觉、数学、最小实现、PyTorch、实验、失败、迁移的顺序推进。
- 工程线与研究线至少选择一条；三级练习先独立完成，再看答案。
- 所有结果必须标成实际测量、教学模拟或文献引用；不能混写。
- 遇到高风险决策、隐私数据或证据不足时停止自动化并交给给人。

隐私与材料规则

全书案例只使用虚构或公开数据。不要在练习、截图、Notebook 或网站笔记中输入姓名、账号、联系方式、密钥、医疗记录、成绩等个人信息。网页学习记录仅保存在当前设备，可匿名导入导出；不建账号、不做行为追踪。

目录

目录由 Word 字段生成。打开文件后若页码尚未刷新，请按 Ctrl+A，再按 F9。

版本、边界与使用方法..... 2

 能力终点..... 2

 全面覆盖的可审计边界..... 3

 每章学习契约..... 3

 隐私与材料规则..... 3

目录..... 4

本卷 24 章路线图..... 34

 月 7 · 搜索、约束与规划..... 34

 月 8 · 逻辑、知识表示和概率推理..... 34

 月 9 · 监督学习..... 34

 月 10 · 无监督与自监督学习..... 35

 月 11 · 评估、因果、公平与鲁棒性..... 35

 月 12 · 神经网络、自动微分与 PyTorch..... 36

第 25 章 状态空间与 BFS、DFS、UCS..... 37

 1. 为什么重要，现在能做什么..... 37

 2. 先修检查与生活类比..... 38

 3. 互动实验：只改变一个变量..... 38

 4. 直觉到数学..... 39

 5. 最小实现..... 39

6. 五轮系统化理解	40
1. 问题边界	40
3. 机制与因果链	41
7. 评估与不确定性	42
9. 工程与复现	43
11. 迁移与研究	45
7. PyTorch 实现路线	46
8. 实验设计与结果记录	46
9. 常见失败与排查顺序	47
10. 迁移任务	47
11. 三级练习	47
12. 本章能力检查	48
13. 来源与核实	48
第 26 章 启发式与 A*：用估计减少盲目探索	49
1. 为什么重要，现在能做什么	49
2. 先修检查与生活类比	50
3. 互动实验：只改变一个变量	50
4. 直觉到数学	51
5. 最小实现	51
6. 五轮系统化理解	52
1. 问题边界	52

3. 机制与因果链.....	53
7. 评估与不确定性	54
9. 工程与复现	55
11. 迁移与研究	56
7. PyTorch 实现路线	58
8. 实验设计与结果记录	58
9. 常见失败与排查顺序	58
10. 迁移任务.....	59
11. 三级练习.....	59
12. 本章能力检查.....	60
13. 来源与核实	60
第 27 章 约束满足、回溯与 SAT	61
1. 为什么重要，现在能做什么.....	61
2. 先修检查与生活类比	62
3. 互动实验：只改变一个变量.....	62
4. 直觉到数学	63
5. 最小实现.....	63
6. 五轮系统化理解	64
1. 问题边界.....	64
3. 机制与因果链.....	65
7. 评估与不确定性	66

9. 工程与复现	67
11. 迁移与研究	69
7. PyTorch 实现路线	70
8. 实验设计与结果记录	70
9. 常见失败与排查顺序	71
10. 迁移任务	71
11. 三级练习	71
12. 本章能力检查	72
13. 来源与核实	72
第 28 章 规划、博弈树与对手	73
1. 为什么重要，现在能做什么	73
2. 先修检查与生活类比	74
3. 互动实验：只改变一个变量	74
4. 直觉到数学	75
5. 最小实现	75
6. 五轮系统化理解	76
1. 问题边界	76
3. 机制与因果链	77
7. 评估与不确定性	78
9. 工程与复现	79
11. 迁移与研究	80

7. PyTorch 实现路线	81
8. 实验设计与结果记录	82
9. 常见失败与排查顺序	82
10. 迁移任务	83
11. 三级练习	83
12. 本章能力检查	83
13. 来源与核实	84
月 7 阶段项目：实现一个校园路线与排课求解器	85
交付物	85
验收方法	85
第 29 章 命题逻辑与一阶逻辑	86
1. 为什么重要，现在能做什么	86
2. 先修检查与生活类比	87
3. 互动实验：只改变一个变量	87
4. 直觉到数学	88
5. 最小实现	88
6. 五轮系统化理解	89
1. 问题边界	89
3. 机制与因果链	90
7. 评估与不确定性	91
9. 工程与复现	92

11. 迁移与研究	93
7. PyTorch 实现路线	94
8. 实验设计与结果记录	95
9. 常见失败与排查顺序	95
10. 迁移任务	96
11. 三级练习	96
12. 本章能力检查	96
13. 来源与核实	97
第 30 章 知识图谱、本体与语义关系	98
1. 为什么重要，现在能做什么	98
2. 先修检查与生活类比	99
3. 互动实验：只改变一个变量	99
4. 直觉到数学	100
5. 最小实现	100
6. 五轮系统化理解	101
1. 问题边界	101
3. 机制与因果链	102
7. 评估与不确定性	103
9. 工程与复现	104
11. 迁移与研究	105
7. PyTorch 实现路线	107

8. 实验设计与结果记录	107
9. 常见失败与排查顺序	107
10. 迁移任务	108
11. 三级练习	108
12. 本章能力检查	109
13. 来源与核实	109
第 31 章 贝叶斯网、马尔可夫网与 HMM	110
1. 为什么重要，现在能做什么	110
2. 先修检查与生活类比	111
3. 互动实验：只改变一个变量	111
4. 直觉到数学	112
5. 最小实现	112
6. 五轮系统化理解	113
1. 问题边界	113
3. 机制与因果链	114
7. 评估与不确定性	115
9. 工程与复现	116
11. 迁移与研究	117
7. PyTorch 实现路线	119
8. 实验设计与结果记录	119
9. 常见失败与排查顺序	119

10. 迁移任务	120
11. 三级练习	120
12. 本章能力检查	121
13. 来源与核实	121
第 32 章 决策论与概率程序	122
1. 为什么重要，现在能做什么	122
2. 先修检查与生活类比	123
3. 互动实验：只改变一个变量	123
4. 直觉到数学	124
5. 最小实现	124
6. 五轮系统化理解	125
1. 问题边界	125
3. 机制与因果链	126
7. 评估与不确定性	127
9. 工程与复现	128
11. 迁移与研究	130
7. PyTorch 实现路线	131
8. 实验设计与结果记录	131
9. 常见失败与排查顺序	132
10. 迁移任务	132
11. 三级练习	132

12. 本章能力检查	133
13. 来源与核实	133
月 8 阶段项目：制作一个带解释的虚构植物诊断知识库	135
交付物	135
验收方法	135
第 33 章 从问题定义到数据切分	136
1. 为什么重要，现在能做什么	136
2. 先修检查与生活类比	137
3. 互动实验：只改变一个变量	137
4. 直觉到数学	138
5. 最小实现	138
6. 五轮系统化理解	139
1. 问题边界	139
3. 机制与因果链	140
7. 评估与不确定性	141
9. 工程与复现	142
11. 迁移与研究	144
7. PyTorch 实现路线	145
8. 实验设计与结果记录	145
9. 常见失败与排查顺序	146
10. 迁移任务	146

11. 三级练习	146
12. 本章能力检查	147
13. 来源与核实	147
第 34 章 线性回归与逻辑回归	149
1. 为什么重要，现在能做什么	149
2. 先修检查与生活类比	150
3. 互动实验：只改变一个变量	150
4. 直觉到数学	151
5. 最小实现	151
6. 五轮系统化理解	152
1. 问题边界	152
3. 机制与因果链	153
7. 评估与不确定性	154
9. 工程与复现	155
11. 迁移与研究	156
7. PyTorch 实现路线	157
8. 实验设计与结果记录	158
9. 常见失败与排查顺序	158
10. 迁移任务	159
11. 三级练习	159
12. 本章能力检查	159

13. 来源与核实	160
第 35 章 决策树、随机森林与梯度提升.....	161
1. 为什么重要，现在能做什么.....	161
2. 先修检查与生活类比	162
3. 互动实验：只改变一个变量.....	162
4. 直觉到数学	163
5. 最小实现.....	163
6. 五轮系统化理解	164
1. 问题边界.....	164
3. 机制与因果链.....	165
7. 评估与不确定性	166
9. 工程与复现	167
11. 迁移与研究	168
7. PyTorch 实现路线	170
8. 实验设计与结果记录	170
9. 常见失败与排查顺序	170
10. 迁移任务.....	171
11. 三级练习.....	171
12. 本章能力检查.....	172
13. 来源与核实	172
第 36 章 间隔、核方法与泛化.....	173

1. 为什么重要，现在能做什么	173
2. 先修检查与生活类比	174
3. 互动实验：只改变一个变量	174
4. 直觉到数学	175
5. 最小实现	175
6. 五轮系统化理解	176
1. 问题边界	176
3. 机制与因果链	177
7. 评估与不确定性	178
9. 工程与复现	179
11. 迁移与研究	180
7. PyTorch 实现路线	181
8. 实验设计与结果记录	182
9. 常见失败与排查顺序	182
10. 迁移任务	183
11. 三级练习	183
12. 本章能力检查	183
13. 来源与核实	184
月 9 阶段项目：建立一个可解释的学习资源分类器	185
交付物	185
验收方法	185

第 37 章 聚类：发现结构但不发明真相.....	186
1. 为什么重要，现在能做什么.....	186
2. 先修检查与生活类比	187
3. 互动实验：只改变一个变量.....	187
4. 直觉到数学	188
5. 最小实现.....	188
6. 五轮系统化理解	189
1. 问题边界.....	189
3. 机制与因果链.....	190
7. 评估与不确定性	191
9. 工程与复现	192
11. 迁移与研究	193
7. PyTorch 实现路线	195
8. 实验设计与结果记录	195
9. 常见失败与排查顺序	195
10. 迁移任务.....	196
11. 三级练习.....	196
12. 本章能力检查.....	197
13. 来源与核实	197
第 38 章 密度估计与异常检测.....	198
1. 为什么重要，现在能做什么.....	198

2. 先修检查与生活类比	199
3. 互动实验：只改变一个变量.....	199
4. 直觉到数学	200
5. 最小实现.....	200
6. 五轮系统化理解	201
1. 问题边界.....	201
3. 机制与因果链.....	202
7. 评估与不确定性	203
9. 工程与复现	204
11. 迁移与研究	205
7. PyTorch 实现路线	207
8. 实验设计与结果记录	207
9. 常见失败与排查顺序	207
10. 迁移任务.....	208
11. 三级练习.....	208
12. 本章能力检查.....	209
13. 来源与核实	209
第 39 章 降维、流形与可视化陷阱.....	210
1. 为什么重要，现在能做什么.....	210
2. 先修检查与生活类比	211
3. 互动实验：只改变一个变量.....	211

4. 直觉到数学	212
5. 最小实现.....	212
6. 五轮系统化理解	213
1. 问题边界.....	213
3. 机制与因果链.....	214
7. 评估与不确定性	215
9. 工程与复现	216
11. 迁移与研究	217
7. PyTorch 实现路线	219
8. 实验设计与结果记录	219
9. 常见失败与排查顺序	219
10. 迁移任务.....	220
11. 三级练习.....	220
12. 本章能力检查.....	221
13. 来源与核实	221
第 40 章 自监督任务与表示学习	222
1. 为什么重要，现在能做什么.....	222
2. 先修检查与生活类比	223
3. 互动实验：只改变一个变量.....	223
4. 直觉到数学	224
5. 最小实现.....	224

6. 五轮系统化理解	225
1. 问题边界	225
3. 机制与因果链	226
7. 评估与不确定性	227
9. 工程与复现	228
11. 迁移与研究	229
7. PyTorch 实现路线	231
8. 实验设计与结果记录	231
9. 常见失败与排查顺序	231
10. 迁移任务	232
11. 三级练习	232
12. 本章能力检查	233
13. 来源与核实	233
月 10 阶段项目：探索一组公开图书或音乐元数据	234
交付物	234
验收方法	234
第 41 章 指标、阈值、校准与不确定性	235
1. 为什么重要，现在能做什么	235
2. 先修检查与生活类比	236
3. 互动实验：只改变一个变量	236
4. 直觉到数学	237

5. 最小实现.....	237
6. 五轮系统化理解	238
1. 问题边界.....	238
3. 机制与因果链.....	239
7. 评估与不确定性	240
9. 工程与复现	241
11. 迁移与研究	242
7. PyTorch 实现路线	244
8. 实验设计与结果记录	244
9. 常见失败与排查顺序	244
10. 迁移任务.....	245
11. 三级练习.....	245
12. 本章能力检查.....	246
13. 来源与核实	246
第 42 章 实验设计与因果推断.....	247
1. 为什么重要，现在能做什么.....	247
2. 先修检查与生活类比	248
3. 互动实验：只改变一个变量.....	248
4. 直觉到数学	249
5. 最小实现.....	249
6. 五轮系统化理解	250

1. 问题边界	250
3. 机制与因果链	251
7. 评估与不确定性	252
9. 工程与复现	253
11. 迁移与研究	254
7. PyTorch 实现路线	256
8. 实验设计与结果记录	256
9. 常见失败与排查顺序	256
10. 迁移任务	257
11. 三级练习	257
12. 本章能力检查	258
13. 来源与核实	258
第 43 章 公平、偏差与数据分布变化	259
1. 为什么重要，现在能做什么	259
2. 先修检查与生活类比	260
3. 互动实验：只改变一个变量	260
4. 直觉到数学	261
5. 最小实现	261
6. 五轮系统化理解	262
1. 问题边界	262
3. 机制与因果链	263

7. 评估与不确定性	264
9. 工程与复现	265
11. 迁移与研究	266
7. PyTorch 实现路线	268
8. 实验设计与结果记录	268
9. 常见失败与排查顺序	268
10. 迁移任务	269
11. 三级练习	269
12. 本章能力检查	270
13. 来源与核实	270
第 44 章 鲁棒性、对抗样本与分布外检测	271
1. 为什么重要，现在能做什么	271
2. 先修检查与生活类比	272
3. 互动实验：只改变一个变量	272
4. 直觉到数学	273
5. 最小实现	273
6. 五轮系统化理解	274
1. 问题边界	274
3. 机制与因果链	275
7. 评估与不确定性	276
9. 工程与复现	277

11. 迁移与研究	278
7. PyTorch 实现路线	279
8. 实验设计与结果记录	280
9. 常见失败与排查顺序	280
10. 迁移任务	281
11. 三级练习	281
12. 本章能力检查	281
13. 来源与核实	282
月 11 阶段项目：建立一个模型审计小组报告	283
交付物	283
验收方法	283
第 45 章 感知机与多层网络	284
1. 为什么重要，现在能做什么	284
2. 先修检查与生活类比	285
3. 互动实验：只改变一个变量	285
4. 直觉到数学	286
5. 最小实现	286
6. 五轮系统化理解	287
1. 问题边界	287
3. 机制与因果链	288
7. 评估与不确定性	289

9. 工程与复现	290
11. 迁移与研究	291
7. PyTorch 实现路线	293
8. 实验设计与结果记录	293
9. 常见失败与排查顺序	293
10. 迁移任务	294
11. 三级练习	294
12. 本章能力检查	295
13. 来源与核实	295
第 46 章 反向传播与自动微分实现	296
1. 为什么重要，现在能做什么	296
2. 先修检查与生活类比	297
3. 互动实验：只改变一个变量	297
4. 直觉到数学	298
5. 最小实现	298
6. 五轮系统化理解	299
1. 问题边界	299
3. 机制与因果链	300
7. 评估与不确定性	301
9. 工程与复现	302
11. 迁移与研究	303

7. PyTorch 实现路线	305
8. 实验设计与结果记录	305
9. 常见失败与排查顺序	305
10. 迁移任务	306
11. 三级练习	306
12. 本章能力检查	307
13. 来源与核实	307
第 47 章 初始化、归一化、正则化与训练诊断	308
1. 为什么重要，现在能做什么	308
2. 先修检查与生活类比	309
3. 互动实验：只改变一个变量	309
4. 直觉到数学	310
5. 最小实现	310
6. 五轮系统化理解	311
1. 问题边界	311
3. 机制与因果链	312
7. 评估与不确定性	313
9. 工程与复现	315
11. 迁移与研究	316
7. PyTorch 实现路线	317
8. 实验设计与结果记录	317

9. 常见失败与排查顺序	318
10. 迁移任务	318
11. 三级练习	319
12. 本章能力检查	319
13. 来源与核实	319
第 48 章 PyTorch 工程：数据、检查点与推理	321
1. 为什么重要，现在能做什么	321
2. 先修检查与生活类比	322
3. 互动实验：只改变一个变量	322
4. 直觉到数学	323
5. 最小实现	323
6. 五轮系统化理解	324
1. 问题边界	324
3. 机制与因果链	325
7. 评估与不确定性	326
9. 工程与复现	327
11. 迁移与研究	329
7. PyTorch 实现路线	330
8. 实验设计与结果记录	330
9. 常见失败与排查顺序	331
10. 迁移任务	331

11. 三级练习	331
12. 本章能力检查	332
13. 来源与核实	332
月 12 阶段项目：训练并发布一个 CPU 可运行的小型神经网络.....	333
交付物	333
验收方法.....	333
独立练习答案.....	334
第 25 章答案 · 状态空间与 BFS、DFS、UCS.....	334
基础题参考	334
应用题参考	334
挑战题参考	334
自评量规.....	335
第 26 章答案 · 启发式与 A*：用估计减少盲目探索	335
基础题参考	335
应用题参考	335
挑战题参考	335
自评量规.....	336
第 27 章答案 · 约束满足、回溯与 SAT.....	336
基础题参考	336
应用题参考	336
挑战题参考	336

自评量规.....	337
第 28 章答案 · 规划、博弈树与对手	337
基础题参考	337
应用题参考	337
挑战题参考	337
自评量规.....	338
第 29 章答案 · 命题逻辑与一阶逻辑	338
基础题参考	338
应用题参考	338
挑战题参考	338
自评量规.....	339
第 30 章答案 · 知识图谱、本体与语义关系	339
基础题参考	339
应用题参考	339
挑战题参考	339
自评量规.....	340
第 31 章答案 · 贝叶斯网、马尔可夫网与 HMM	340
基础题参考	340
应用题参考	340
挑战题参考	340
自评量规.....	341

第 32 章答案 · 决策论与概率程序	341
基础题参考	341
应用题参考	341
挑战题参考	341
自评量规	342
第 33 章答案 · 从问题定义到数据切分	342
基础题参考	342
应用题参考	342
挑战题参考	342
自评量规	343
第 34 章答案 · 线性回归与逻辑回归	343
基础题参考	343
应用题参考	343
挑战题参考	343
自评量规	344
第 35 章答案 · 决策树、随机森林与梯度提升	344
基础题参考	344
应用题参考	344
挑战题参考	344
自评量规	345
第 36 章答案 · 间隔、核方法与泛化	345

基础题参考	345
应用题参考	345
挑战题参考	345
自评量规.....	346
第 37 章答案 · 聚类：发现结构但不发明真相	346
基础题参考	346
应用题参考	346
挑战题参考	346
自评量规.....	347
第 38 章答案 · 密度估计与异常检测	347
基础题参考	347
应用题参考	347
挑战题参考	347
自评量规.....	348
第 39 章答案 · 降维、流形与可视化陷阱	348
基础题参考	348
应用题参考	348
挑战题参考	348
自评量规.....	349
第 40 章答案 · 自监督任务与表示学习	349
基础题参考	349

应用题参考	349
挑战题参考	349
自评量规.....	350
第 41 章答案 · 指标、阈值、校准与不确定性	350
基础题参考	350
应用题参考	350
挑战题参考	350
自评量规.....	351
第 42 章答案 · 实验设计与因果推断	351
基础题参考	351
应用题参考	351
挑战题参考	351
自评量规.....	352
第 43 章答案 · 公平、偏差与数据分布变化.....	352
基础题参考	352
应用题参考	352
挑战题参考	352
自评量规.....	353
第 44 章答案 · 鲁棒性、对抗样本与分布外检测.....	353
基础题参考	353
应用题参考	353

挑战题参考	353
自评量规.....	354
第 45 章答案 · 感知机与多层网络	354
基础题参考	354
应用题参考	354
挑战题参考	354
自评量规.....	355
第 46 章答案 · 反向传播与自动微分实现	355
基础题参考	355
应用题参考	355
挑战题参考	355
自评量规.....	356
第 47 章答案 · 初始化、归一化、正则化与训练诊断	356
基础题参考	356
应用题参考	356
挑战题参考	356
自评量规.....	357
第 48 章答案 · PyTorch 工程：数据、检查点与推理	357
基础题参考	357
应用题参考	357
挑战题参考	357

自评量规.....	358
本卷来源台账.....	359
[S010] Dive into Deep Learning.....	359
[S015] Automatic Differentiation with torch.autograd	359
[S016] Artificial Intelligence: A Modern Approach	360
[S017] CS221: Artificial Intelligence	360
[S018] OWL 2 Web Ontology Language Document Overview.....	361
[S019] Probabilistic Graphical Models.....	362
[S020] scikit-learn User Guide.....	362
[S021] Machine Learning Crash Course	363
[S022] Causal Inference: The Mixtape.....	363
[S023] Fairlearn User Guide	364
[S024] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations	364
[S025] PyTorch Tutorials.....	365
[S030] Attention Is All You Need	365
[S071] Model Cards for Model Reporting.....	366
[S072] Datasheets for Datasets.....	366
全书术语表.....	368
版权、版本与勘误.....	378

本卷 24 章路线图

月 7 · 搜索、约束与规划

明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。

1. [第 25 章 状态空间与 BFS、DFS、UCS — 明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。](#)
2. [第 26 章 启发式与 A*：用估计减少盲目探索 — 理解可采纳、一致启发式和搜索效率之间的关系。](#)
3. [第 27 章 约束满足、回溯与 SAT — 把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。](#)
4. [第 28 章 规划、博弈树与对手 — 理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。](#)

月 8 · 逻辑、知识表示和概率推理

理解语法、语义、满足、量词和推理规则，区分真值与现实证据。

5. [第 29 章 命题逻辑与一阶逻辑 — 理解语法、语义、满足、量词和推理规则，区分真值与现实证据。](#)
6. [第 30 章 知识图谱、本体与语义关系 — 建模实体、关系、类别、约束和溯源，理解开放世界假设。](#)
7. [第 31 章 贝叶斯网、马尔可夫网与 HMM — 用图表达条件独立性，在序列中区分隐藏状态和观测。](#)
8. [第 32 章 决策论与概率程序 — 把不确定推断与偏好、代价、行动连接，理解抽样近似。](#)

月 9 · 监督学习

把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。

9. [第 33 章 从问题定义到数据切分 — 把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。](#)
10. [第 34 章 线性回归与逻辑回归 — 理解线性组合、最小二乘、对数几率、正则化和系数解释边界。](#)

11. [第 35 章 决策树、随机森林与梯度提升 — 理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。](#)
12. [第 36 章 间隔、核方法与泛化 — 从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。](#)

月 10 · 无监督与自监督学习

比较 K 均值、层次聚类 and 密度方法，理解尺度、距离和簇数选择。

13. [第 37 章 聚类：发现结构但不发明真相 — 比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。](#)
14. [第 38 章 密度估计与异常检测 — 理解概率密度、邻域、孤立和阈值，区分罕见与有害。](#)
15. [第 39 章 降维、流形与可视化陷阱 — 比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。](#)
16. [第 40 章 自监督任务与表示学习 — 理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。](#)

月 11 · 评估、因果、公平与鲁棒性

根据错误代价选择指标，区分排序、分类、概率质量和置信表达。

17. [第 41 章 指标、阈值、校准与不确定性 — 根据错误代价选择指标，区分排序、分类、概率质量和置信表达。](#)
18. [第 42 章 实验设计与因果推断 — 区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。](#)
19. [第 43 章 公平、偏差与数据分布变化 — 识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。](#)
20. [第 44 章 鲁棒性、对抗样本与分布外检测 — 测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。](#)

月 12 · 神经网络、自动微分与 PyTorch

理解线性层、非线性激活、隐藏表示和通用近似的实际边界。

21. [第 45 章 感知机与多层网络 — 理解线性层、非线性激活、隐藏表示和通用近似的实际边界。](#)
22. [第 46 章 反向传播与自动微分实现 — 沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。](#)
23. [第 47 章 初始化、归一化、正则化与训练诊断 — 理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。](#)
24. [第 48 章 PyTorch 工程：数据、检查点与推理 — 掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。](#)

第 25 章 状态空间与 BFS、DFS、UCS

State-space search

项目	内容
学习单元	月 7 · 搜索、约束与规划
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 24 章 数值稳定、条件数与非凸地形](#)

结论先行：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“状态空间与 BFS、DFS、UCS”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 numerical-stability。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 25-1 状态空间与 BFS、DFS、UCS：单变量实验（教学模拟）

项目	内容
只改变量	分支因子，范围 2–10，步长 1
实验目标	只改变“分支因子”，观察结果、代价和风险如何一起变化，理解明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把分支因子分别设为 2、3 和 10，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“分支因子”，观察结果、代价和风险如何一起变化，理解明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。

静态替代说明：静态替代：把分支因子分别设为 2、3 和 10，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`frontier ← successors(state); cost(path)=Σ edge_cost`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 25-1 python 最小实现

```
from collections import deque
def bfs(graph, start, goal):
```

```

queue = deque([(start, [start])])
visited = {start}
while queue:
    node, path = queue.popleft()
    if node == goal:
        return path
    for nxt in graph.get(node, []):
        if nxt not in visited:
            visited.add(nxt)
            queue.append((nxt, path + [nxt]))
return None

graph = {"A": ["B", "C"], "B": ["D"], "C": ["D"], "D": []}
print(bfs(graph, "A", "D"))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“状态空间与 BFS、DFS、UCS”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“分支因子”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `frontier ← successors(state); cost(path)=Σ edge_cost`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当分支因子变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“状态空间与 BFS、DFS、UCS”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低

可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“分支因子”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `frontier ← successors(state); cost(path)=Σ edge_cost`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当分支因子变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“状态空间与 BFS、DFS、UCS”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“分支因子”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier} \leftarrow \text{successors}(\text{state}); \text{cost}(\text{path}) = \sum \text{edge_cost}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当分支因子变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“状态空间与 BFS、DFS、UCS”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。。接着逐项标记观察到的事实、由公式

推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“分支因子”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{frontier} \leftarrow \text{successors}(\text{state}); \text{cost}(\text{path}) = \sum \text{edge_cost}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。只是入口，并给出一个类比失效的地方。如果解释中出现

“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当分支因子变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“状态空间与 BFS、DFS、UCS”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“分支因子”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `frontier ← successors(state); cost(path)=Σ edge_cost`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在迷宫里每个路口是状态，每次移动

是动作；搜索策略决定先查看哪条岔路。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当分支因子变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“分支因子”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

25. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
26. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
27. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
28. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
29. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
30. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
31. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择不含个人信息的教学模拟天气记录作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开天文观测的缩小样例，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

32. **基础题：** 用自己的话解释“状态空间与 BFS、DFS、UCS”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

33. **应用题：** 在合成的植物叶片特征中设计一个只改变“分支因子”的实验，列出基线、三档设置、指标和失败样例。
34. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/state-space-search.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](http://aima.cs.berkeley.edu))
- [S017] Stanford University, 《CS221: Artificial Intelligence》，[访问 stanford-cs221.github.io](https://stanford-cs221.github.io))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 26 章 启发式与 A*：用估计减少盲目探索

Heuristics and A*

项目	内容
学习单元	月 7 · 搜索、约束与规划
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 25 章 状态空间与 BFS、DFS、UCS](#)

结论先行：理解可采纳、一致启发式和搜索效率之间的关系。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“启发式与 A*：用估计减少盲目探索”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解可采纳、一致启发式和搜索效率之间的关系。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `state-space-search`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：导航会同时考虑已经走过的距离和到终点的估计距离。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 26-1 启发式与 A*：用估计减少盲目探索：单变量实验（教学模拟）

项目	内容
只改变量	启发式权重，范围 0–2，步长 0.1
实验目标	只改变“启发式权重”，观察结果、代价和风险如何一起变化，理解理解可采纳、一致启发式和搜索效率之间的关系。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把启发式权重分别设为 0、1 和 2，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“启发式权重”，观察结果、代价和风险如何一起变化，理解理解可采纳、一致启发式和搜索效率之间的关系。

静态替代说明：静态替代：把启发式权重分别设为 0、1 和 2，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $f(n)=g(n)+h(n)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。

若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 26-1 python 最小实现

```
from collections import deque

def bfs(graph, start, goal):
    queue = deque([(start, [start])])
    visited = {start}
    while queue:
        node, path = queue.popleft()
        if node == goal:
            return path
        for nxt in graph.get(node, []):
            if nxt not in visited:
                visited.add(nxt)
                queue.append((nxt, path + [nxt]))
    return None

graph = {"A": ["B", "C"], "B": ["D"], "C": ["D"], "D": []}
print(bfs(graph, "A", "D"))
```


运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“启发式与 A*：用估计减少盲目探索”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解可采纳、一致启发式和搜索效率之间的关系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。导航会同时考虑已经走过的距离和到终点的估计距离。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“启发式权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $f(n)=g(n)+h(n)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空

输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么导航会同时考虑已经走过的距离和到终点的估计距离。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当启发式权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“启发式与 A*：用估计减少盲目探索”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解可采纳、一致启发式和搜索效率之间的关系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。导航会同时考虑已经走过的距离和到终点的估计距离。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“启发式权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $f(n)=g(n)+h(n)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么导航会同时考虑已经走过的距离和到终点的估计距离。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当启发式权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“启发式与 A*：用估计减少盲目探索”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解可采纳、一致启发式和搜索效率之间的关系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。导航会同时考虑已经走过的距离和到终点的估计距离。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“启发式权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $f(n)=g(n)+h(n)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么导航会同时考虑已经走过的距离和到终点的估计距离。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当启发式权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“启发式与 A^* ：用估计减少盲目探索”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解可采纳、一致启发式和搜索效率之间的关系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。导航会同时考虑已经走过的距离和到终点的估计距离。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“启发式权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $f(n)=g(n)+h(n)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么导航会同时考虑已经走过的距离和到终点的估计距离。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当启发式权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“启发式与 A^* ：用估计减少盲目探索”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解可采纳、一致启发式和搜索效率之间的关系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。导航会同时考虑已经走过的距离和到终点的估计距离。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“启发式权重”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $f(n)=g(n)+h(n)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么导航会同时考虑已经走过的距离和到终点的估计距离。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当启发式权重变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“启发式权重”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

35. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

36. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

37. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

38. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

39. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

40. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

41. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开道路网络与虚构出行任务作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成音频波形与环境声标签，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

42. **基础题：** 用自己的话解释“启发式与 A*：用估计减少盲目探索”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

43. **应用题：** 在虚构游戏角色的状态日志中设计一个只改变“启发式权重”的实验，列出基线、三档设置、指标和失败样例。

44. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/heuristics-and-astar.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](https://aima.cs.berkeley.edu))
- [S017] Stanford University, 《CS221: Artificial Intelligence》，[访问 stanford-cs221.github.io](https://stanford-cs221.github.io))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 27 章 约束满足、回溯与 SAT

Constraint satisfaction and SAT

项目	内容
学习单元	月 7 · 搜索、约束与规划
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 26 章 启发式与 A*：用估计减少盲目探索](#)

结论先行： 把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“约束满足、回溯与 SAT”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `heuristics-and-astar`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。类比的 价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在 笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证 据。

3. 互动实验：只改变一个变量

实验 27-1 约束满足、回溯与 SAT：单变量实验（教学模拟）

项目	内容
只改变量	约束密度，范围 0–100%，步长 5
实验目标	只改变“约束密度”，观察结果、代价和风险如何一起变化，理解把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把约束密度分别设为 0%、40%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性

项目	内容
	能或现实世界因果效果。

实验目标：只改变“约束密度”，观察结果、代价和风险如何一起变化，理解把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。

静态替代说明：静态替代：把约束密度分别设为 0%、40%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{solution} \in \text{domain}^{\text{variables}}$ subject to all constraints。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 27-1 python 最小实现

```
from collections import deque

def bfs(graph, start, goal):
    queue = deque([(start, [start])])
    visited = {start}
    while queue:
```

```

    node, path = queue.popleft()
    if node == goal:
        return path
    for nxt in graph.get(node, []):
        if nxt not in visited:
            visited.add(nxt)
            queue.append((nxt, path + [nxt]))
    return None

graph = {"A": ["B", "C"], "B": ["D"], "C": ["D"], "D": []}
print(bfs(graph, "A", "D"))

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“约束满足、回溯与 SAT”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“约束密度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{solution} \in \text{domain}^{\text{variables}}$ subject to all constraints，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当约束密度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“约束满足、回溯与 SAT”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“约束密度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{solution} \in \text{domain}^{\text{variables}}$ subject to all constraints，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当约束密度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“约束满足、回溯与 SAT”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“约束密度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{solution} \in \text{domain}^{\text{variables}}$ subject to all constraints，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当约束密度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“约束满足、回溯与 SAT”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。。接着逐项标记观察到的事实、由公式推导的结果、来

自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“约束密度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{solution} \in \text{domain}^{\text{variables}}$ subject to all constraints，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当约束密度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“约束满足、回溯与 SAT”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“约束密度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{solution} \in \text{domain}^{\text{variables}}$ subject to all constraints，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。只是入口，并给出一个类比失效的地方。如果解释中出

现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当约束密度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“约束密度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 45. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 46. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 47. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 48. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 49. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 50. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 51. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成的植物叶片特征作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园能源使用的匿名汇总数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 52. **基础题：** 用自己的话解释“约束满足、回溯与 SAT”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

53. **应用题：**在公开天文观测的缩小样例中设计一个只改变“约束密度”的实验，列出基线、三档设置、指标和失败样例。

54. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/csp-and-sat.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](http://aima.cs.berkeley.edu))
- [S017] Stanford University, 《CS221: Artificial Intelligence》，[访问 stanford-cs221.github.io](https://stanford-cs221.github.io))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 28 章 规划、博弈树与对手

Planning and games

项目	内容
学习单元	月 7 · 搜索、约束与规划
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 27 章 约束满足、回溯与 SAT](#)

结论先行：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“规划、博弈树与对手”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 csp-and-sat。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：下棋不仅要想自己下一步，还要模拟对手最合理的回应。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 28-1 规划、博弈树与对手：单变量实验（教学模拟）

项目	内容
只改变量	搜索深度，范围 1–12 层，步长 1
实验目标	只改变“搜索深度”，观察结果、代价和风险如何一起变化，理解理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把搜索深度分别设为 1 层、4 层和 12 层，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“搜索深度”，观察结果、代价和风险如何一起变化，理解理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。

静态替代说明：静态替代：把搜索深度分别设为 1 层、4 层和 12 层，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $V(s) = \max_a \min_b V(T(s, a, b))$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 28-1 python 最小实现

```
from collections import deque

def bfs(graph, start, goal):
    queue = deque([(start, [start])])
    visited = {start}
    while queue:
        node, path = queue.popleft()
        if node == goal:
            return path
        for nxt in graph.get(node, []):
            if nxt not in visited:
                visited.add(nxt)
                queue.append((nxt, path + [nxt]))
    return None
```



```
graph = {"A": ["B", "C"], "B": ["D"], "C": ["D"], "D": []}
print(bfs(graph, "A", "D"))
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化管理理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“规划、博弈树与对手”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。下棋不仅要想自己下一步，还要模拟对手最合理的回应。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“搜索深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $V(s)=\max_a \min_b V(T(s,a,b))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么下棋不仅要想自己下一步，还要模拟对手最合理的回应。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当搜索深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“规划、博弈树与对手”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。下棋不仅要想自己下一步，还要模拟对手最合理的回应。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“搜索深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $V(s)=\max_a \min_b V(T(s,a,b))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么下棋不仅要想自己下一步，还要模拟对手最合理的回应。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当搜索深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“规划、博弈树与对手”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。下棋不仅要想自己下一步，还要模拟对手最合理的回应。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“搜索深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $V(s)=\max_a \min_b V(T(s,a,b))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么下棋不仅要想自己下一步，还要模拟对手最合理的回应。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当搜索深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“规划、博弈树与对手”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。下棋不仅要想自己下一步，还要模拟对手最合理的回应。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“搜索深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $V(s)=\max_a \min_b V(T(s,a,b))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么下棋不仅要想自己下一步，还要模拟对手最合理的回应。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当搜索深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“规划、博弈树与对手”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。下棋不仅要想自己下一步，还要模拟对手最合理的回应。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“搜索深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $V(s) = \max_a \min_b V(T(s, a, b))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么下棋不仅要想自己下一步，还要模拟对手最合理的回应。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当搜索深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“搜索深度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

55. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

56. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

57. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

58. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

59. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

60. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

61. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构游戏角色的状态日志作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开古诗文本与版权已明确的材料，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

62. **基础题**：用自己的话解释“规划、博弈树与对手”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

63. **应用题**：在合成音频波形与环境声标签中设计一个只改变“搜索深度”的实验，列出基线、三档设置、指标和失败样例。

64. **挑战题**：设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/planning-and-games.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。

- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](http://aima.cs.berkeley.edu))
- [S017] Stanford University, 《CS221: Artificial Intelligence》，[访问 stanford-cs221.github.io](https://stanford-cs221.github.io))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 7 阶段项目：实现一个校园路线与排课求解器

项目结论： 分别用图搜索和约束求解处理路径与排课，比较正确性、代价和扩展节点数。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

65. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
66. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
67. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
68. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 29 章 命题逻辑与一阶逻辑

Logic

项目	内容
学习单元	月 8 · 逻辑、知识表示和概率推理
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 28 章 规划、博弈树与对手](#)

结论先行：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“命题逻辑与一阶逻辑”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `planning-and-games`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。类比的价
值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 29-1 命题逻辑与一阶逻辑：单变量实验（教学模拟）

项目	内容
只改变量	规则数量，范围 1–100 条，步长 1
实验目标	只改变“规则数量”，观察结果、代价和风险如何一起变化，理解理解语法、语义、满足、量词和推理规则，区分真值与现实证据。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把规则数量分别设为 1 条、12 条和 100 条，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“规则数量”，观察结果、代价和风险如何一起变化，理解理解语法、语义、满足、量词和推理规则，区分真值与现实证据。

静态替代说明：静态替代：把规则数量分别设为 1 条、12 条和 100 条，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $KB \models \alpha$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 29-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“命题逻辑与一阶逻辑”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“规则数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $KB \models \alpha$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当规则数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“命题逻辑与一阶逻辑”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“规则数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $KB \models \alpha$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当规则数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“命题逻辑与一阶逻辑”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“规则数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $KB \models \alpha$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当规则数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“命题逻辑与一阶逻辑”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“规则数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $KB \models \alpha$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当规则数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“命题逻辑与一阶逻辑”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解语法、语义、满足、量词和推理规则，区分真值与现实证据。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，

最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“规则数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $KB \models \alpha$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当规则数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“规则数量”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

69. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

70. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

71. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

72. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

73. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

74. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

75. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开天文观测的缩小样例作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构机器人在二维网格中的轨迹，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

76. **基础题**：用自己的话解释“命题逻辑与一阶逻辑”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

77. **应用题**：在校园能源使用的匿名汇总数据中设计一个只改变“规则数量”的实验，列出基线、三档设置、指标和失败样例。

78. **挑战题**：设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/propositional-first-order.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。

- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](http://aima.cs.berkeley.edu))
- [S018] W3C, 《OWL 2 Web Ontology Language Document Overview》，[访问 w3.org](http://www.w3.org))
- [S019] Daphne Koller, Nir Friedman, 《Probabilistic Graphical Models》，[访问 mitpress.mit.edu](http://mitpress.mit.edu))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

第 30 章 知识图谱、本体与语义关系

Knowledge graphs and ontologies

项目	内容
学习单元	月 8 · 逻辑、知识表示和概率推理
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系： [第 29 章 命题逻辑与一阶逻辑](#)

结论先行： 建模实体、关系、类别、约束和溯源，理解开放世界假设。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“知识图谱、本体与语义关系”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：建模实体、关系、类别、约束和溯源，理解开放世界假设。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 propositional-first-order。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 30-1 知识图谱、本体与语义关系：单变量实验（教学模拟）

项目	内容
只改变量	图连接率，范围 0–100%，步长 5
实验目标	只改变“图连接率”，观察结果、代价和风险如何一起变化，理解建模实体、关系、类别、约束和溯源，理解开放世界假设。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把图连接率分别设为 0%、30%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“图连接率”，观察结果、代价和风险如何一起变化，理解建模实体、关系、类别、约束和溯源，理解开放世界假设。

静态替代说明：静态替代：把图连接率分别设为 0%、30%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：triple = (subject, predicate, object)。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 30-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“知识图谱、本体与语义关系”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建模实体、关系、类别、约束和溯源，理解开放世界假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“图连接率”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{triple} = (\text{subject}, \text{predicate}, \text{object})$ ，逐个说明符号的对象、形状、单位和取值范围；

把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当图连接率变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“知识图谱、本体与语义关系”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建模实体、关系、类别、约束和溯源，理解开放世界假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“图连接率”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 triple = (subject, predicate, object)，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当图连接率变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“知识图谱、本体与语义关系”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建模实体、关系、类别、约束和溯源，理解开放世界假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“图连接率”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 triple = (subject, predicate, object)，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当图连接率变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“知识图谱、本体与语义关系”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建模实体、关系、类别、约束和溯源，理解开放世界假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“图连接率”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{triple} = (\text{subject}, \text{predicate}, \text{object})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当图连接率变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“知识图谱、本体与语义关系”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：建模实体、关系、类别、约束和溯源，理解开放世界假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳

定能力。图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“图连接率”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{triple} = (\text{subject}, \text{predicate}, \text{object})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当图连接率变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“图连接率”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

79. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

80. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

81. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

82. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

83. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

84. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

85. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成音频波形与环境声标签作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开科研论文的标题与摘要元数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

86. **基础题：** 用自己的话解释“知识图谱、本体与语义关系”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

87. **应用题：** 在公开古诗文本与版权已明确的材料中设计一个只改变“图连接率”的实验，列出基线、三档设置、指标和失败样例。

88. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/knowledge-graphs.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](https://aima.cs.berkeley.edu))
- [S018] W3C, 《OWL 2 Web Ontology Language Document Overview》，[访问 w3.org](https://www.w3.org))
- [S019] Daphne Koller, Nir Friedman, 《Probabilistic Graphical Models》，[访问 mitpress.mit.edu](https://mitpress.mit.edu))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 31 章 贝叶斯网、马尔可夫网与 HMM

Graphical models and HMMs

项目	内容
学习单元	月 8 · 逻辑、知识表示和概率推理
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 30 章 知识图谱、本体与语义关系](#)

结论先行：用图表达条件独立性，在序列中区分隐藏状态和观测。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“贝叶斯网、马尔可夫网与 HMM”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：用图表达条件独立性，在序列中区分隐藏状态和观测。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 knowledge-graphs。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 31-1 贝叶斯网、马尔可夫网与 HMM：单变量实验（教学模拟）

项目	内容
只改变量	转移噪声，范围 0–1，步长 0.05
实验目标	只改变“转移噪声”，观察结果、代价和风险如何一起变化，理解用图表达条件独立性，在序列中区分隐藏状态和观测。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把转移噪声分别设为 0、0.15 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“转移噪声”，观察结果、代价和风险如何一起变化，理解用图表达条件独立性，在序列中区分隐藏状态和观测。

静态替代说明：静态替代：把转移噪声分别设为 0、0.15 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $P(x_1, \dots, x_n) = \prod P(x_i | \text{parents}(x_i))$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 31-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“贝叶斯网、马尔可夫网与 HMM”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：用图表达条件独立性，在序列中区分隐藏状态和观测。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“转移噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x_1 \cdots x_n) = \prod P(x_i | \text{parents}(x_i))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、

极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当转移噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“贝叶斯网、马尔可夫网与 HMM”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：用图表达条件独立性，在序列中区分隐藏状态和观测。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“转移噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x_1 \cdots x_n) = \prod P(x_i | \text{parents}(x_i))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当转移噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“贝叶斯网、马尔可夫网与 HMM”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：用图表达条件独立性，在序列中区分隐藏状态和观测。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“转移噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x_1 \cdots x_n) = \prod P(x_i | \text{parents}(x_i))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当转移噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“贝叶斯网、马尔可夫网与 HMM”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：用图表达条件独立性，在序列中区分隐藏状态和观测。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“转移噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x_1 \cdots x_n) = \prod P(x_i | \text{parents}(x_i))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当转移噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“贝叶斯网、马尔科夫网与 HMM”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：用图表达条件独立性，在序列中区分隐藏状态和观测。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳

定能力。隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“转移噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P(x_1 \cdots x_n) = \prod P(x_i | \text{parents}(x_i))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当转移噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“转移噪声”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

89. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。

90. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。

91. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

92. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

93. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

94. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

95. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园能源使用的匿名汇总数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成电商目录但不含真实用户行为，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

96. **基础题：** 用自己的话解释“贝叶斯网、马尔可夫网与 HMM”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

97. **应用题：** 在虚构机器人在二维网格中的轨迹中设计一个只改变“转移噪声”的实验，列出基线、三档设置、指标和失败样例。

98. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/graphical-models-hmm.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](http://aima.cs.berkeley.edu))
- [S018] W3C, 《OWL 2 Web Ontology Language Document Overview》，[访问 w3.org](http://www.w3.org))
- [S019] Daphne Koller, Nir Friedman, 《Probabilistic Graphical Models》，[访问 mitpress.mit.edu](http://mitpress.mit.edu))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 32 章 决策论与概率程序

Decision theory and probabilistic programming

项目	内容
学习单元	月 8 · 逻辑、知识表示和概率推理
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 31 章 贝叶斯网、马尔可夫网与 HMM](#)

结论先行： 把不确定推断与偏好、代价、行动连接，理解抽样近似。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“决策论与概率程序”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：把不确定推断与偏好、代价、行动连接，理解抽样近似。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 graphical-models-hmm。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 32-1 决策论与概率程序：单变量实验（教学模拟）

项目	内容
只改变量	误判代价，范围 0–100，步长 5
实验目标	只改变“误判代价”，观察结果、代价和风险如何一起变化，理解把不确定推断与偏好、代价、行动连接，理解抽样近似。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把误判代价分别设为 0、40 和 100，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性

项目	内容
	能或现实世界因果效果。

实验目标：只改变“误判代价”，观察结果、代价和风险如何一起变化，理解把不确定推断与偏好、代价、行动连接，理解抽样近似。

静态替代说明：静态替代：把误判代价分别设为 0、40 和 100，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $a^* = \operatorname{argmax}_a E[U(a, \theta) | \text{data}]$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 32-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
```

```

score = sum(x * w for x, w in zip(features, weights)) + bias
return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“决策论与概率程序”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把不确定推断与偏好、代价、行动连接，理解抽样近似。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“误判代价”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $a^* = \operatorname{argmax}_a E[U(a, \theta) | \text{data}]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当误判代价变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“决策论与概率程序”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把不确定推断与偏好、代价、行动连接，理解抽样近似。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“误判代价”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $a^* = \operatorname{argmax}_a E[U(a, \theta) | \text{data}]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当误判代价变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“决策论与概率程序”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把不确定推断与偏好、代价、行动连接，理解抽样近似。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“误判代价”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $a^* = \operatorname{argmax}_a E[U(a, \theta) | \text{data}]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当误判代价变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“决策论与概率程序”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把不确定推断与偏好、代价、行动连接，理解抽样近似。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。医生式决策

不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“误判代价”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $a^* = \operatorname{argmax}_a E[U(a, \theta) | \text{data}]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当误判代价变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“决策论与概率程序”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把不确定推断与偏好、代价、行动连接，理解抽样近似。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“误判代价”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $a^* = \operatorname{argmax}_a E[U(a, \theta) | \text{data}]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。只是入口，并给出一个类比失效的地方。如果解释中

出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当误判代价变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“误判代价”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 99. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 100. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 101. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 102. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 103. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 104. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 105. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开古诗文本与版权已明确的材料作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园图书馆的公开借阅类别，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 106. **基础题：** 用自己的话解释“决策论与概率程序”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

107. **应用题：**在公开科研论文的标题与摘要元数据中设计一个只改变“误判代价”的实验，列出基线、三档设置、指标和失败样例。

108. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/decision-and-probabilistic-programming.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S016] Stuart Russell, Peter Norvig, 《Artificial Intelligence: A Modern Approach》，[访问 aima.cs.berkeley.edu](http://aima.cs.berkeley.edu))
- [S018] W3C, 《OWL 2 Web Ontology Language Document Overview》，[访问 w3.org](http://www.w3.org))
- [S019] Daphne Koller, Nir Friedman, 《Probabilistic Graphical Models》，[访问 mitpress.mit.edu](http://mitpress.mit.edu))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 8 阶段项目：制作一个带解释的虚构植物诊断知识库

项目结论： 使用公开植物规则与教学模拟概率，展示规则链、冲突、不确定性和无法判断的情况。
当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

109. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
110. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
111. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
112. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 33 章 从问题定义到数据切分

Problem framing and splits

项目	内容
学习单元	月 9 · 监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系： [第 32 章 决策论与概率程序](#)

结论先行： 把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“从问题定义到数据切分”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 decision-and-probabilistic-programming。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 33-1 从问题定义到数据切分：单变量实验（教学模拟）

项目	内容
只改变量	测试集比例，范围 5–50%，步长 5
实验目标	只改变“测试集比例”，观察结果、代价和风险如何一起变化，理解把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把测试集比例分别设为 5%、20%和 50%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性

项目	内容
	能或现实世界因果效果。

实验目标：只改变“测试集比例”，观察结果、代价和风险如何一起变化，理解把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。

静态替代说明：静态替代：把测试集比例分别设为 5%、20%和 50%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`generalization gap = test_error - train_error`。先不要急着推公式。逐个给符号贴

标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 33-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
```

```

score = sum(x * w for x, w in zip(features, weights)) + bias
return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“从问题定义到数据切分”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“测试集比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{generalization gap} = \text{test_error} - \text{train_error}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当测试集比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“从问题定义到数据切分”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“测试集比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{generalization gap} = \text{test_error} - \text{train_error}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当测试集比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“从问题定义到数据切分”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均

值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“测试集比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{generalization gap} = \text{test_error} - \text{train_error}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当测试集比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“从问题定义到数据切分”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。

期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“测试集比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{generalization gap} = \text{test_error} - \text{train_error}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当测试集比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“从问题定义到数据切分”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“测试集比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{generalization gap} = \text{test_error} - \text{train_error}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。只是入口，并给出一个类比失效的地方。如果解释中出现“肯

定” “完全” “智能” “理解” 等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当测试集比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“测试集比例”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 113. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 114. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 115. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 116. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 117. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 118. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 119. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构机器人在二维网格中的轨迹作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择不含个人信息的教学模拟天气记录，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 120. **基础题：** 用自己的话解释“从问题定义到数据切分”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

121. **应用题：**在合成电商目录但不含真实用户行为中设计一个只改变“测试集比例”的实验，列出基线、三档设置、指标和失败样例。

122. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/problem-framing-and-splits.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》，[访问 scikit-learn.org](https://scikit-learn.org))
- [S021] Google, 《Machine Learning Crash Course》，[访问 developers.google.com](https://developers.google.com))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》，[访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 34 章 线性回归与逻辑回归

Linear and logistic regression

项目	内容
学习单元	月 9 · 监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：第 33 章 [从问题定义到数据切分](#)

结论先行：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“线性回归与逻辑回归”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `problem-framing-and-splits`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 34-1 线性回归与逻辑回归：单变量实验（教学模拟）

项目	内容
只改变量	决策阈值，范围 0–1，步长 0.05
实验目标	只改变“决策阈值”，观察结果、代价和风险如何一起变化，理解理解线性组合、最小二乘、对数几率、正则化和系数解释边界。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把决策阈值分别设为 0、0.5 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“决策阈值”，观察结果、代价和风险如何一起变化，理解理解线性组合、最小二乘、对数几率、正则化和系数解释边界。

静态替代说明：静态替代：把决策阈值分别设为 0、0.5 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\hat{y} = Xw + b$; $P(y=1|x) = \sigma(w^T x + b)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 34-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“线性回归与逻辑回归”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\hat{y}=Xw+b$; $P(y=1|x)=\sigma(w^Tx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、

极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“线性回归与逻辑回归”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\hat{y}=Xw+b$; $P(y=1|x)=\sigma(w^Tx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“线性回归与逻辑回归”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\hat{y}=Xw+b$; $P(y=1|x)=\sigma(wTx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“线性回归与逻辑回归”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\hat{y}=Xw+b$; $P(y=1|x)=\sigma(w^Tx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“线性回归与逻辑回归”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性组合、最小二乘、对数几率、正则化和系数解释边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、

均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\hat{y}=Xw+b$; $P(y=1|x)=\sigma(w^Tx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“决策阈值”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 123. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 124. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 125. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 126. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 127. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 128. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 129. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开科研论文的标题与摘要元数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开道路网络与虚构出行任务，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

130. **基础题：**用自己的话解释“线性回归与逻辑回归”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

131. **应用题：**在校园图书馆的公开借阅类别中设计一个只改变“决策阈值”的实验，列出基线、三档设置、指标和失败样例。

132. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/linear-logistic-regression.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。

- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》, [访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》, [访问 scikit-learn.org](https://scikit-learn.org))
- [S021] Google, 《Machine Learning Crash Course》, [访问 developers.google.com](https://developers.google.com))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 35 章 决策树、随机森林与梯度提升

Trees and ensembles

项目	内容
学习单元	月 9 · 监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 34 章 线性回归与逻辑回归](#)

结论先行：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“决策树、随机森林与梯度提升”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 linear-logistic-regression。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 35-1 决策树、随机森林与梯度提升：单变量实验（教学模拟）

项目	内容
只改变量	树深度，范围 1–30 层，步长 1
实验目标	只改变“树深度”，观察结果、代价和风险如何一起变化，理解理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把树深度分别设为 1 层、6 层和 30 层，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“树深度”，观察结果、代价和风险如何一起变化，理解理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。

静态替代说明：静态替代：把树深度分别设为 1 层、6 层和 30 层，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{ensemble}(x) = \text{aggregate}(f_1(x) \dots f_k(x))$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 35-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```


运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“决策树、随机森林与梯度提升”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“树深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{ensemble}(x) = \text{aggregate}(f_1(x) \cdots f_k(x))$ ，逐个说明符号的对象、形状、单位和取值范围；

把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当树深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“决策树、随机森林与梯度提升”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“树深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{ensemble}(x) = \text{aggregate}(f_1(x) \cdots f_k(x))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当树深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“决策树、随机森林与梯度提升”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“树深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{ensemble}(x) = \text{aggregate}(f_1(x) \cdots f_k(x))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当树深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“决策树、随机森林与梯度提升”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“树深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{ensemble}(x) = \text{aggregate}(f_1(x) \cdots f_k(x))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当树深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“决策树、随机森林与梯度提升”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误

当成稳定能力。多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“树深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{ensemble}(x) = \text{aggregate}(f_1(x) \cdots f_k(x))$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当树深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“树深度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

133. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
134. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
135. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

136. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

137. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

138. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

139. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成电商目录但不含真实用户行为作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成的植物叶片特征，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

140. **基础题：** 用自己的话解释“决策树、随机森林与梯度提升”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

141. **应用题：** 在不含个人信息的教学模拟天气记录中设计一个只改变“树深度”的实验，列出基线、三档设置、指标和失败样例。

142. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/trees-ensembles-boosting.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》，[访问 scikit-learn.org](https://scikit-learn.org))
- [S021] Google, 《Machine Learning Crash Course》，[访问 developers.google.com](https://developers.google.com))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》，[访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 36 章 间隔、核方法与泛化

Margins, kernels and generalization

项目	内容
学习单元	月 9 · 监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 35 章 决策树、随机森林与梯度提升](#)

结论先行：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“间隔、核方法与泛化”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 trees-ensembles-boosting。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：在两群人之间画分界线时，离双方都远的线通常更稳。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 36-1 间隔、核方法与泛化：单变量实验（教学模拟）

项目	内容
只改变量	正则强度，范围 0–1，步长 0.01
实验目标	只改变“正则强度”，观察结果、代价和风险如何一起变化，理解从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把正则强度分别设为 0、0.1 和 1，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“正则强度”，观察结果、代价和风险如何一起变化，理解从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。

静态替代说明：静态替代：把正则强度分别设为 0、0.1 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\min \frac{1}{2}||w||^2 + C\S\xi_0$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 36-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
```

```

    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“间隔、核方法与泛化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在两群人之间画分界线时，离双方都远的线通常更稳。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min 1/2\|w\|^2 + C\Sigma\xi_i$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在两群人之间画分界线时，离双方都远的线通常更稳。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“间隔、核方法与泛化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在两群人之间画分界线时，离双方都远的线通常更稳。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min 1/2\|w\|^2 + C\S\xi$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在两群人之间画分界线时，离双方都远的线通常更稳。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“间隔、核方法与泛化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在两群人之间画分界线时，离双方都远的线通常更稳。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min 1/2\|w\|^2 + C\S\xi$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在两群人之间画分界线时，离双方都远的线通常更稳。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“间隔、核方法与泛化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在两群人之间画分界线时，离双方都远的线通常更稳。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低

可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min 1/2\|w\|^2 + C\S\xi$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在两群人之间画分界线时，离双方都远的线通常更稳。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“间隔、核方法与泛化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。在两群人之间画分界线时，离双方都远的线通常更稳。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min \frac{1}{2}\|w\|^2 + C\S\xi$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么在两群人之间画分界线时，离双方都远的线通常更稳。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存

`state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“正则强度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

143. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
144. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
145. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
146. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
147. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
148. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
149. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园图书馆的公开借阅类别作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构游戏角色的状态日志，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

150. **基础题：**用自己的话解释“间隔、核方法与泛化”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
151. **应用题：**在公开道路网络与虚构出行任务中设计一个只改变“正则强度”的实验，列出基线、三档设置、指标和失败样例。
152. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/margins-kernels-generalization.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。

- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》, [访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》, [访问 scikit-learn.org](https://scikit-learn.org))
- [S021] Google, 《Machine Learning Crash Course》, [访问 developers.google.com](https://developers.google.com))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 9 阶段项目：建立一个可解释的学习资源分类器

项目结论：使用虚构标签或公开数据，完成问题定义、基线、训练/验证/测试、错误分析和模型卡。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

153. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
154. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
155. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
156. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 37 章 聚类：发现结构但不发明真相

Clustering

项目	内容
学习单元	月 10 · 无监督与自监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 36 章 间隔、核方法与泛化](#)

结论先行： 比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“聚类：发现结构但不发明真相”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 margins-kernels-generalization。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 37-1 聚类：发现结构但不发明真相：单变量实验（教学模拟）

项目	内容
只改变量	簇数量，范围 2-20 个，步长 1
实验目标	只改变“簇数量”，观察结果、代价和风险如何一起变化，理解比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把簇数量分别设为 2 个、4 个和 20 个，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“簇数量”，观察结果、代价和风险如何一起变化，理解比较 K 均值、层次聚类 and 密度方法，理解尺度、距离和簇数选择。

静态替代说明：静态替代：把簇数量分别设为 2 个、4 个和 20 个，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\min \sum ||x_i - \mu_{\{c_i\}}||^2$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 37-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“聚类：发现结构但不发明真相”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“簇数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min \sum ||x_i - \mu_{c_i}||^2$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、

空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当簇数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“聚类：发现结构但不发明真相”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“簇数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min \sum ||x_i - \mu_{\{c_i\}}||^2$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当簇数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“聚类：发现结构但不发明真相”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“簇数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min \sum ||x_i - \mu_{c_i}||^2$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当簇数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“聚类：发现结构但不发明真相”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“簇数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min \sum ||x_i - \mu_{\{c\}}||^2$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当簇数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“聚类：发现结构但不发明真相”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字

误当成稳定能力。整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“簇数量”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\min \sum ||x_i - \mu_{\{c_i\}}||^2$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当簇数量变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“簇数量”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

157. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
158. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
159. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

160. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

161. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

162. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

163. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择不含个人信息的教学模拟天气记录作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开天文观测的缩小样例，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

164. **基础题：** 用自己的话解释“聚类：发现结构但不发明真相”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

165. **应用题：** 在合成的植物叶片特征中设计一个只改变“簇数量”的实验，列出基线、三档设置、指标和失败样例。

166. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 content/ai-book/answers/clustering.md，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》, [访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》, [访问 scikit-learn.org](https://scikit-learn.org))
- [S030] Ashish Vaswani et al., 《Attention Is All You Need》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 38 章 密度估计与异常检测

Density and anomaly detection

项目	内容
学习单元	月 10 · 无监督与自监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：第 37 章 聚类：发现结构但不发明真相

结论先行：理解概率密度、邻域、孤立和阈值，区分罕见与有害。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“密度估计与异常检测”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解概率密度、邻域、孤立和阈值，区分罕见与有害。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 clustering。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 38-1 密度估计与异常检测：单变量实验（教学模拟）

项目	内容
只改变量	异常阈值，范围 0–100%，步长 5
实验目标	只改变“异常阈值”，观察结果、代价和风险如何一起变化，理解理解概率密度、邻域、孤立和阈值，区分罕见与有害。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把异常阈值分别设为 0%、80%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“异常阈值”，观察结果、代价和风险如何一起变化，理解理解概率密度、邻域、孤立和阈值，区分罕见与有害。

静态替代说明：静态替代：把异常阈值分别设为 0%、80%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{anomaly_score}(x) = -\log p(x)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 38-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“密度估计与异常检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解概率密度、邻域、孤立和阈值，区分罕见与有害。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“异常阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{anomaly_score}(x) = -\log p(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零

值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当异常阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“密度估计与异常检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解概率密度、邻域、孤立和阈值，区分罕见与有害。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“异常阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{anomaly_score}(x) = -\log p(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当异常阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“密度估计与异常检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解概率密度、邻域、孤立和阈值，区分罕见与有害。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设

是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“异常阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{anomaly_score}(x) = -\log p(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当异常阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“密度估计与异常检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解概率密度、邻域、孤立和阈值，区分罕见与有害。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“异常阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{anomaly_score}(x) = -\log p(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当异常阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“密度估计与异常检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解概率密度、邻域、孤立和阈值，区分罕见与有害。。接着逐项标记观察到的事实、由公式推导的结果、来自文献

的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“异常阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{anomaly_score}(x) = -\log p(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当异常阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“异常阈值”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

167. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
168. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
169. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

170. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

171. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

172. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

173. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开道路网络与虚构出行任务作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成音频波形与环境声标签，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

174. **基础题：** 用自己的话解释“密度估计与异常检测”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

175. **应用题：** 在虚构游戏角色的状态日志中设计一个只改变“异常阈值”的实验，列出基线、三档设置、指标和失败样例。

176. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/density-and-anomaly.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》，[访问 scikit-learn.org](https://scikit-learn.org))
- [S030] Ashish Vaswani et al., 《Attention Is All You Need》，[访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 39 章 降维、流形与可视化陷阱

Dimensionality reduction

项目	内容
学习单元	月 10 · 无监督与自监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 38 章 密度估计与异常检测](#)

结论先行： 比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“降维、流形与可视化陷阱”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 density-and-anomaly。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 39-1 降维、流形与可视化陷阱：单变量实验（教学模拟）

项目	内容
只改变量	嵌入维度，范围 2–64 维，步长 2
实验目标	只改变“嵌入维度”，观察结果、代价和风险如何一起变化，理解比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把嵌入维度分别设为 2 维、8 维和 64 维，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“嵌入维度”，观察结果、代价和风险如何一起变化，理解比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。

静态替代说明：静态替代：把嵌入维度分别设为 2 维、8 维和 64 维，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $z=f(x)$, $\dim(z)\ll\dim(x)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 39-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“降维、流形与可视化陷阱”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“嵌入维度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z=f(x)$, $\dim(z) \ll \dim(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大

值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当嵌入维度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“降维、流形与可视化陷阱”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“嵌入维度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z=f(x)$, $\dim(z) \ll \dim(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当嵌入维度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“降维、流形与可视化陷阱”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“嵌入维度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z=f(x)$, $\dim(z) \ll \dim(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当嵌入维度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“降维、流形与可视化陷阱”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均

值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“嵌入维度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z=f(x)$, $\dim(z) \ll \dim(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当嵌入维度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“降维、流形与可视化陷阱”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成

稳定能力。把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“嵌入维度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $z=f(x)$, $\dim(z) \ll \dim(x)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当嵌入维度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“嵌入维度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

177. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
178. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
179. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

- 180. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 181. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 182. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 183. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成的植物叶片特征作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园能源使用的匿名汇总数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 184. **基础题：** 用自己的话解释“降维、流形与可视化陷阱”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
- 185. **应用题：** 在公开天文观测的缩小样例中设计一个只改变“嵌入维度”的实验，列出基线、三档设置、指标和失败样例。
- 186. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/manifold-and-dimension.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》，[访问 scikit-learn.org](https://scikit-learn.org))
- [S030] Ashish Vaswani et al., 《Attention Is All You Need》，[访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 40 章 自监督任务与表示学习

Self-supervised representation learning

项目	内容
学习单元	月 10 · 无监督与自监督学习
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：第 39 章 降维、流形与可视化陷阱

结论先行：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“自监督任务与表示学习”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 manifold-and-dimension。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 40-1 自监督任务与表示学习：单变量实验（教学模拟）

项目	内容
只改变量	遮盖比例，范围 0–90%，步长 5
实验目标	只改变“遮盖比例”，观察结果、代价和风险如何一起变化，理解理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把遮盖比例分别设为 0%、15%和 90%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“遮盖比例”，观察结果、代价和风险如何一起变化，理解理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。

静态替代说明：静态替代：把遮盖比例分别设为 0%、15%和 90%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{loss} = \text{agreement}(\text{positive}) + \text{separation}(\text{negative})$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向 and 边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 40-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“自监督任务与表示学习”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“遮盖比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = \text{agreement}(\text{positive}) + \text{separation}(\text{negative})$ ，逐个说明符号的对象、形状、单

位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当遮盖比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“自监督任务与表示学习”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“遮盖比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = \text{agreement}(\text{positive}) + \text{separation}(\text{negative})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当遮盖比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“自监督任务与表示学习”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“遮盖比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = \text{agreement}(\text{positive}) + \text{separation}(\text{negative})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当遮盖比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“自监督任务与表示学习”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、

均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“遮盖比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = \text{agreement}(\text{positive}) + \text{separation}(\text{negative})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当遮盖比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“自监督任务与表示学习”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误

当成稳定能力。把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“遮盖比例”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{loss} = \text{agreement}(\text{positive}) + \text{separation}(\text{negative})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当遮盖比例变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“遮盖比例”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

187. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
188. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
189. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

190. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

191. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

192. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

193. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构游戏角色的状态日志作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开古诗文本与版权已明确的材料，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

194. **基础题：** 用自己的话解释“自监督任务与表示学习”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

195. **应用题：** 在合成音频波形与环境声标签中设计一个只改变“遮盖比例”的实验，列出基线、三档设置、指标和失败样例。

196. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/self-supervised-representation.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S020] scikit-learn developers, 《scikit-learn User Guide》，[访问 scikit-learn.org](https://scikit-learn.org))
- [S030] Ashish Vaswani et al., 《Attention Is All You Need》，[访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 10 阶段项目：探索一组公开图书或音乐元数据

项目结论：完成聚类、降维、异常检查和自监督表示实验，并说明结果不是天然真实类别。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

197. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
198. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
199. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
200. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 41 章 指标、阈值、校准与不确定性

Metrics and calibration

项目	内容
学习单元	月 11 · 评估、因果、公平与鲁棒性
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系： [第 40 章 自监督任务与表示学习](#)

结论先行： 根据错误代价选择指标，区分排序、分类、概率质量和置信表达。 **版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“指标、阈值、校准与不确定性”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：根据错误代价选择指标，区分排序、分类、概率质量和置信表达。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 self-supervised-representation。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。类比的值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 41-1 指标、阈值、校准与不确定性：单变量实验（教学模拟）

项目	内容
只改变量	决策阈值，范围 0–1，步长 0.05
实验目标	只改变“决策阈值”，观察结果、代价和风险如何一起变化，理解根据错误代价选择指标，区分排序、分类、概率质量和置信表达。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把决策阈值分别设为 0、0.5 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“决策阈值”，观察结果、代价和风险如何一起变化，理解根据错误代价选择指标，区分排序、分类、概率质量和置信表达。

静态替代说明：静态替代：把决策阈值分别设为 0、0.5 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$ ； $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$ 。先不要急着推公式。逐个给符号贴标签：

它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 41-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“指标、阈值、校准与不确定性”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：根据错误代价选择指标，区分排序、分类、概率质量和置信表达。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$; $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$ ，逐个说明符号的对象、形状、单位和取值

范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“指标、阈值、校准与不确定性”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：根据错误代价选择指标，区分排序、分类、概率质量和置信表达。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{precision} = \text{TP}/(\text{TP} + \text{FP})$; $\text{recall} = \text{TP}/(\text{TP} + \text{FN})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“指标、阈值、校准与不确定性”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：根据错误代价选择指标，区分排序、分类、概率质量和置信表达。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{precision} = \text{TP}/(\text{TP} + \text{FP})$; $\text{recall} = \text{TP}/(\text{TP} + \text{FN})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“指标、阈值、校准与不确定性”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：根据错误代价选择指标，区分排序、分类、概率质量和置信表达。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、

均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$; $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“指标、阈值、校准与不确定性”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：根据错误代价选择指标，区分排序、分类、概率质量和置信表达。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字

误当成稳定能力。火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“决策阈值”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$; $\text{recall} = \text{TP} / (\text{TP} + \text{FN})$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当决策阈值变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“决策阈值”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

201. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
202. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
203. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

204. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

205. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

206. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

207. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开天文观测的缩小样例作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构机器人在二维网格中的轨迹，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

208. **基础题：** 用自己的话解释“指标、阈值、校准与不确定性”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

209. **应用题：** 在校园能源使用的匿名汇总数据中设计一个只改变“决策阈值”的实验，列出基线、三档设置、指标和失败样例。

210. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/metrics-and-calibration.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S022] Scott Cunningham, 《Causal Inference: The Mixtape》, [访问 mixtape.scunning.com](https://mixtape.scunning.com))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 42 章 实验设计与因果推断

Experiments and causality

项目	内容
学习单元	月 11 · 评估、因果、公平与鲁棒性
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 41 章 指标、阈值、校准与不确定性](#)

结论先行： 区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“实验设计与因果推断”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 metrics-and-calibration。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 42-1 实验设计与因果推断：单变量实验（教学模拟）

项目	内容
只改变量	混杂强度，范围 0–100%，步长 5
实验目标	只改变“混杂强度”，观察结果、代价和风险如何一起变化，理解区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把混杂强度分别设为 0%、30%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“混杂强度”，观察结果、代价和风险如何一起变化，理解区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。

静态替代说明：静态替代：把混杂强度分别设为 0%、30%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $ATE = E[Y(1) - Y(0)]$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 42-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“实验设计与因果推断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“混杂强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $ATE=E[Y(1)-Y(0)]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、

空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当混杂强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“实验设计与因果推断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“混杂强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $ATE=E[Y(1)-Y(0)]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当混杂强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“实验设计与因果推断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“混杂强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $ATE=E[Y(1)-Y(0)]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当混杂强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“实验设计与因果推断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“混杂强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $ATE=E[Y(1)-Y(0)]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当混杂强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“实验设计与因果推断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成

稳定能力。带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“混杂强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $ATE=E[Y(1)-Y(0)]$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当混杂强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“混杂强度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

211. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
212. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
213. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

- 214. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 215. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 216. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 217. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成音频波形与环境声标签作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开科研论文的标题与摘要元数据，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 218. **基础题：** 用自己的话解释“实验设计与因果推断”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
- 219. **应用题：** 在公开古诗文本与版权已明确的材料中设计一个只改变“混杂强度”的实验，列出基线、三档设置、指标和失败样例。
- 220. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/experiments-and-causality.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S022] Scott Cunningham, 《Causal Inference: The Mixtape》, [访问 mixtape.scunning.com](https://mixtape.scunning.com))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 43 章 公平、偏差与数据分布变化

Fairness, bias and shift

项目	内容
学习单元	月 11 · 评估、因果、公平与鲁棒性
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 42 章 实验设计与因果推断](#)

结论先行： 识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。**版本边界：** 本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“公平、偏差与数据分布变化”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 experiments-and-causality。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 43-1 公平、偏差与数据分布变化：单变量实验（教学模拟）

项目	内容
只改变量	少数组占比，范围 0–100%，步长 5
实验目标	只改变“少数组占比”，观察结果、代价和风险如何一起变化，理解识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把少数组占比分别设为 0%、50%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“少数组占比”，观察结果、代价和风险如何一起变化，理解识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。

静态替代说明：静态替代：把少数组占比分别设为 0%、50%和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $P_{\text{train}}(X,Y) \neq P_{\text{deploy}}(X,Y)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 43-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“公平、偏差与数据分布变化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“少数组占比”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P_{\text{train}}(X,Y) \neq P_{\text{deploy}}(X,Y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零

值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当少数组占比变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“公平、偏差与数据分布变化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“少数组占比”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P_{\text{train}}(X,Y) \neq P_{\text{deploy}}(X,Y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当少数组占比变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“公平、偏差与数据分布变化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“少数组占比”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P_{\text{train}}(X,Y) \neq P_{\text{deploy}}(X,Y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当少数组占比变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“公平、偏差与数据分布变化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园能源使用的匿名汇总数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规

则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园图书馆的公开借阅类别作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“少数组占比”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P_{\text{train}}(X,Y) \neq P_{\text{deploy}}(X,Y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在校园能源使用的匿名汇总数据中，当少数组占比变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“公平、偏差与数据分布变化”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当

成稳定能力。统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用不含个人信息的教学模拟天气记录建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成音频波形与环境声标签作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“少数组占比”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $P_{\text{train}}(X,Y) \neq P_{\text{deploy}}(X,Y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在不含个人信息的教学模拟天气记录中，当少数组占比变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“少数组占比”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

221. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
222. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
223. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

224. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

225. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

226. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

227. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园能源使用的匿名汇总数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成电商目录但不含真实用户行为，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

228. **基础题：** 用自己的话解释“公平、偏差与数据分布变化”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

229. **应用题：** 在虚构机器人在二维网格中的轨迹中设计一个只改变“少数组占比”的实验，列出基线、三档设置、指标和失败样例。

230. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/fairness-and-shift.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S022] Scott Cunningham, 《Causal Inference: The Mixtape》, [访问 mixtape.scunning.com](https://mixtape.scunning.com))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 44 章 鲁棒性、对抗样本与分布外检测

Robustness and OOD

项目	内容
学习单元	月 11 · 评估、因果、公平与鲁棒性
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 43 章 公平、偏差与数据分布变化](#)

结论先行：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“鲁棒性、对抗样本与分布外检测”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 fairness-and-shift。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。 类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 44-1 鲁棒性、对抗样本与分布外检测：单变量实验（教学模拟）

项目	内容
只改变量	输入噪声，范围 0–100%，步长 5
实验目标	只改变“输入噪声”，观察结果、代价和风险如何一起变化，理解测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把输入噪声分别设为 0%、10%和 100%，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“输入噪声”，观察结果、代价和风险如何一起变化，理解测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。

静态替代说明：静态替代：把输入噪声分别设为 0%、10% 和 100%，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\max_{\{\|\delta\| \leq \epsilon\}} \text{loss}(f(x+\delta), y)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 44-1 python 最小实现

```
import math

def sigmoid(z: float) -> float:
    return 1.0 / (1.0 + math.exp(-max(-30.0, min(30.0, z))))

def predict(features, weights, bias=0.0):
    score = sum(x * w for x, w in zip(features, weights)) + bias
    return sigmoid(score)

probability = predict([0.4, 0.8, -0.2], [0.5, 1.2, -0.7], -0.1)
print({"probability": round(probability, 4), "teaching_simulation": True})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“鲁棒性、对抗样本与分布外检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“输入噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失

败样例。对于 $\max_{\{\|\delta\| \leq \epsilon\}} \text{loss}(f(x+\delta), y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当输入噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“鲁棒性、对抗样本与分布外检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“输入噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\max_{\{\|\delta\| \leq \varepsilon\}} \text{loss}(f(x+\delta), y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当输入噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“鲁棒性、对抗样本与分布外检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“输入噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\max_{\|\delta\| \leq \epsilon} \text{loss}(f(x+\delta), y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当输入噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“鲁棒性、对抗样本与分布外检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开古诗文本与版权已明确的材料建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把不含个人信息的教学模拟天气记录作为迁移案例，检查相同表示和假设

是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“输入噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\max_{\{\|\delta\| \leq \epsilon\}} \text{loss}(f(x+\delta), y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开古诗文本与版权已明确的材料中，当输入噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“鲁棒性、对抗样本与分布外检测”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开道路网络与虚构出行任务建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把校园能源使用的匿名汇总数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“输入噪声”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\max_{\{\|\delta\| \leq \epsilon\}} \text{loss}(f(x+\delta), y)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开道路网络与虚构出行任务中，当输入噪声变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存

`state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“输入噪声”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

231. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
232. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
233. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
234. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
235. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
236. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
237. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开古诗文本与版权已明确的材料作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择校园图书馆的公开借阅类别，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

238. **基础题：**用自己的话解释“鲁棒性、对抗样本与分布外检测”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

239. **应用题：**在公开科研论文的标题与摘要元数据中设计一个只改变“输入噪声”的实验，列出基线、三档设置、指标和失败样例。

240. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/robustness-adversarial-ood.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。

- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S022] Scott Cunningham, 《Causal Inference: The Mixtape》, [访问 mixtape.scunning.com](https://mixtape.scunning.com))
- [S023] Fairlearn contributors, 《Fairlearn User Guide》, [访问 fairlearn.org](https://fairlearn.org))
- [S024] NIST, 《Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations》, [访问 csrc.nist.gov](https://csrc.nist.gov))
- [S071] Margaret Mitchell et al., 《Model Cards for Model Reporting》, [访问 arxiv.org](https://arxiv.org))
- [S072] Timnit Gebru et al., 《Datasheets for Datasets》, [访问 arxiv.org](https://arxiv.org))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 draft → technical_review → fact_checked → stable 前进。

月 11 阶段项目：建立一个模型审计小组报告

项目结论：对同一模型做指标、校准、群体切片、数据漂移、因果假设和对抗测试，记录不确定性。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

241. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
242. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
243. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
244. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

第 45 章 感知机与多层网络

Perceptrons and MLPs

项目	内容
学习单元	月 12 · 神经网络、自动微分与 PyTorch
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 44 章 鲁棒性、对抗样本与分布外检测](#)

结论先行：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“感知机与多层网络”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 robustness-adversarial-ood。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。类比的帮助是先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 45-1 感知机与多层网络：单变量实验（教学模拟）

项目	内容
只改变量	隐藏层宽度，范围 2–256，步长 2
实验目标	只改变“隐藏层宽度”，观察结果、代价和风险如何一起变化，理解理解线性层、非线性激活、隐藏表示和通用近似的实际边界。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把隐藏层宽度分别设为 2、32 和 256，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“隐藏层宽度”，观察结果、代价和风险如何一起变化，理解理解线性层、非线性激活、隐藏表示和通用近似的实际边界。

静态替代说明：静态替代：把隐藏层宽度分别设为 2、32 和 256，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $h=\phi(Wx+b)$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 45-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“感知机与多层网络”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐藏层宽度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h=\phi(Wx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、

重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当隐藏层宽度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“感知机与多层网络”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐藏层宽度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h=\phi(Wx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当隐藏层宽度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“感知机与多层网络”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐藏层宽度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h=\phi(Wx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当隐藏层宽度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“感知机与多层网络”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构机器人在二维网格中的轨迹建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固

定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开道路网络与虚构出行任务作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐藏层宽度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h=\phi(Wx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构机器人在二维网格中的轨迹中，当隐藏层宽度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“感知机与多层网络”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解线性层、非线性激活、隐藏表示和通用近似的实际边界。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。

多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成的植物叶片特征建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开古诗文本与版权已明确的材料作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“隐藏层宽度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $h=\phi(Wx+b)$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成的植物叶片特征中，当隐藏层宽度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“隐藏层宽度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

245. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
246. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
247. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

248. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。

249. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。

250. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。

251. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择虚构机器人在二维网格中的轨迹作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择不含个人信息的教学模拟天气记录，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

252. **基础题：** 用自己的话解释“感知机与多层网络”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

253. **应用题：** 在合成电商目录但不含真实用户行为中设计一个只改变“隐藏层宽度”的实验，列出基线、三档设置、指标和失败样例。

254. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/perceptron-mlp.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S015] PyTorch, 《Automatic Differentiation with torch.autograd》，[访问 pytorch.org](https://pytorch.org))
- [S025] PyTorch, 《PyTorch Tutorials》，[访问 pytorch.org](https://pytorch.org))

核实日期：2026-08-06。评审状态：**draft**。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 46 章 反向传播与自动微分实现

Backpropagation and autodiff

项目	内容
学习单元	月 12 · 神经网络、自动微分与 PyTorch
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 45 章 感知机与多层网络](#)

结论先行：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“反向传播与自动微分实现”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 perceptron-mlp。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 46-1 反向传播与自动微分实现：单变量实验（教学模拟）

项目	内容
只改变量	网络深度，范围 1-100 层，步长 1
实验目标	只改变“网络深度”，观察结果、代价和风险如何一起变化，理解沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把网络深度分别设为 1 层、12 层和 100 层，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“网络深度”，观察结果、代价和风险如何一起变化，理解沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。

静态替代说明：静态替代：把网络深度分别设为 1 层、12 层和 100 层，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\partial L / \partial x = \partial L / \partial y \cdot \partial y / \partial x$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：

如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；

如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 46-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})
```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“反向传播与自动微分实现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“网络深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\partial L / \partial x = \partial L / \partial y \cdot \partial y / \partial x$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、

极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当网络深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“反向传播与自动微分实现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“网络深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\partial L / \partial x = \partial L / \partial y \cdot \partial y / \partial x$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当网络深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“反向传播与自动微分实现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“网络深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\partial L / \partial x = \partial L / \partial y \cdot \partial y / \partial x$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当网络深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“反向传播与自动微分实现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开科研论文的标题与摘要元数据建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如

固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成的植物叶片特征作为迁移案例，检查相同表示和假设是否仍成立。

若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“网络深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\partial L / \partial x = \partial L / \partial y \cdot \partial y / \partial x$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开科研论文的标题与摘要元数据中，当网络深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“反向传播与自动微分实现”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当

成稳定能力。总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用虚构游戏角色的状态日志建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构机器人在二维网格中的轨迹作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“网络深度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\partial L / \partial x = \partial L / \partial y \cdot \partial y / \partial x$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在虚构游戏角色的状态日志中，当网络深度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“网络深度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

255. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
256. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
257. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。

- 258. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 259. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 260. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 261. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择公开科研论文的标题与摘要元数据作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择公开道路网络与虚构出行任务，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 262. **基础题：** 用自己的话解释“反向传播与自动微分实现”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
- 263. **应用题：** 在校园图书馆的公开借阅类别中设计一个只改变“网络深度”的实验，列出基线、三档设置、指标和失败样例。
- 264. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/backprop-autodiff.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S015] PyTorch, 《Automatic Differentiation with torch.autograd》，[访问 pytorch.org](https://pytorch.org))
- [S025] PyTorch, 《PyTorch Tutorials》，[访问 pytorch.org](https://pytorch.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 47 章 初始化、归一化、正则化与训练诊断

Training and regularization

项目	内容
学习单元	月 12 · 神经网络、自动微分与 PyTorch
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 46 章 反向传播与自动微分实现](#)

结论先行：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。**版本边界：**本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“初始化、归一化、正则化与训练诊断”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 `backprop-autodiff`。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 47-1 初始化、归一化、正则化与训练诊断：单变量实验（教学模拟）

项目	内容
只改变量	正则强度，范围 0–1，步长 0.01
实验目标	只改变“正则强度”，观察结果、代价和风险如何一起变化，理解理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把正则强度分别设为 0、0.1 和 1，手工比较三组教学模拟输出。
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性

项目	内容
	能或现实世界因果效果。

实验目标：只改变“正则强度”，观察结果、代价和风险如何一起变化，理解理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。

静态替代说明：静态替代：把正则强度分别设为 0、0.1 和 1，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系： $\text{objective} = \text{data_loss} + \lambda \cdot \text{complexity}$ 。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 47-1 python 最小实现

```
import torch

torch.manual_seed(20260806)
x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
```

```

    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“初始化、归一化、正则化与训练诊断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{objective} = \text{data_loss} + \lambda \cdot \text{complexity}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“初始化、归一化、正则化与训练诊断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接

受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{objective} = \text{data_loss} + \lambda \cdot \text{complexity}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它们替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“初始化、归一化、正则化与训练诊断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。

这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{objective} = \text{data_loss} + \lambda \cdot \text{complexity}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“初始化、归一化、正则化与训练诊断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成电商目录但不含真实用户行为建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把虚构游戏角色的状态日志作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{objective} = \text{data_loss} + \lambda \cdot \text{complexity}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么训练像练球：难度太低会只记动作，

强度太高会受伤；需观察长期曲线而非单次成绩。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成电商目录但不含真实用户行为中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“初始化、归一化、正则化与训练诊断”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用公开天文观测的缩小样例建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开科研论文的标题与摘要元数据作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“正则强度”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{objective} = \text{data_loss} + \lambda \cdot \text{complexity}$ ，逐个说明符号的对象、形状、单位和取值范

围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在公开天文观测的缩小样例中，当正则强度变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“正则强度”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 265. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 266. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 267. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 268. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 269. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 270. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 271. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择合成电商目录但不含真实用户行为作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择合成的植物叶片特征，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

272. **基础题：**用自己的话解释“初始化、归一化、正则化与训练诊断”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。
273. **应用题：**在不含个人信息的教学模拟天气记录中设计一个只改变“正则强度”的实验，列出基线、三档设置、指标和失败样例。
274. **挑战题：**设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/training-regularization.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S015] PyTorch, 《Automatic Differentiation with torch.autograd》，[访问 pytorch.org](https://pytorch.org))

- [S025] PyTorch, 《PyTorch Tutorials》, [访问 pytorch.org](https://pytorch.org))

核实日期：2026-08-06。评审状态：`draft`。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

第 48 章 PyTorch 工程：数据、检查点与推理

PyTorch engineering

项目	内容
学习单元	月 12 · 神经网络、自动微分与 PyTorch
建议时间	8 小时
评审状态	draft · 核实 2026-08-06
练习与实验	3 道练习 · 1 个互动实验

先修关系：[第 47 章 初始化、归一化、正则化与训练诊断](#)

结论先行：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。**版本边界：**

本章属于 2026.08 Beta，评审状态为 draft。它是可运行、可核查的教学稿，不等于专家签字后的正式版。

1. 为什么重要，现在能做什么

学完本章，你应该能把“PyTorch 工程：数据、检查点与推理”从一个名词变成可操作的分析框架。你能解释它在完整 AI 系统中的位置，写出最小数学关系，运行一个 CPU 可承受的示例，设计只改变一个变量的实验，识别至少三种失败原因，并把方法迁移到新的公开或虚构场景。真正的能力不是复述答案，而是面对结果不理想时知道先检查哪里、需要补什么证据、何时应该拒绝继续自动化。

本章聚焦：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。这件事重要，因为 AI 系统的错误通常不是“模型突然不聪明”，而是问题定义、数据表示、目标函数、评估或部署条

件之间不一致。把这些环节分开，才能定位故障，也才能对他人解释为什么一个结果值得或不值得相信。

2. 先修检查与生活类比

先用一分钟自测：你能否说明输入和输出分别是什么？能否区分规则、算法、模型与系统？能否读懂变量、函数和简单图表？如果其中一项还不稳，先回到前置章节 training-regularization。不要跳过地基；后续章节默认你会保存代码、记录环境、保护隐私，并区分教学模拟与真实测量。

生活类比：模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。类比的价值是帮助你先获得方向感，但它会省略重要细节。本章后面会用数学、代码和反例修正直觉。请在笔记里写下类比对应关系，并额外写一条“类比失效处”；这一步能防止把好记的故事误当成严谨证据。

3. 互动实验：只改变一个变量

实验 48-1 PyTorch 工程：数据、检查点与推理：单变量实验（教学模拟）

项目	内容
只改变量	批量大小，范围 1-256，步长 1
实验目标	只改变“批量大小”，观察结果、代价和风险如何一起变化，理解掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。
键盘操作	聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。
静态替代	静态替代：把批量大小分别设为 1、32 和 256，手工比较三组教学模拟输出。

项目	内容
结果边界	页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

实验目标：只改变“批量大小”，观察结果、代价和风险如何一起变化，理解掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。

静态替代说明：静态替代：把批量大小分别设为 1、32 和 256，手工比较三组教学模拟输出。

键盘操作：聚焦滑块后使用方向键微调，Page Up/Page Down 大步调整，Home/End 到达边界。

风险提示：页面数值由公开公式生成，只用于教学，不代表真实模型性能或现实世界因果效果。

4. 直觉到数学

核心关系：`artifact = weights + config + preprocessing + version`。先不要急着推公式。逐个给符号贴标签：它表示现实中的什么对象、是标量还是向量、单位是什么、允许取什么范围、由谁测量。然后用最小数字手算一次，再交给代码。若手算、代码和图形不一致，优先检查单位、索引、符号方向和边界，而不是马上换模型。

数学的作用不是让内容显得困难，而是把模糊说法变成可检查关系。公式必须和任务定义相互约束：如果目标变化，损失和指标也可能需要变化；如果数据尺度变化，数值稳定性和优化速度也可能变化；如果输出被用于真实决策，还必须把错误代价、拒答和人工复核写进系统。

5. 最小实现

下面代码保留固定种子和“教学模拟”标签。它的目的不是刷新榜单，而是让你看到输入、状态、计算和输出怎样连接。真实项目要补充数据许可、环境文件、独立测试、异常处理和测量记录。

代码 48-1 python 最小实现

```
import torch
torch.manual_seed(20260806)
```



```

x = torch.randn(16, 4)
y = (x.sum(dim=1, keepdim=True) > 0).float()
model = torch.nn.Sequential(
    torch.nn.Linear(4, 8),
    torch.nn.ReLU(),
    torch.nn.Linear(8, 1),
)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-2)
for _ in range(40):
    loss = torch.nn.functional.binary_cross_entropy_with_logits(model(x), y)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
print({"loss": round(loss.item(), 4), "seed": 20260806})

```

运行后先检查输出类型、范围和极端值，再改变一个参数。不要一次改很多行；每次改变都记录“我预期什么、实际发生什么、可能原因是什么”。如果缺少 PyTorch，可先完成浏览器互动实验和纯 Python 主线；GPU 只属于后续研究复现轨道。

6. 五轮系统化理解

1. 问题边界

先把要解决的任务写成输入、输出、约束、失败代价与不做什么。对“PyTorch 工程：数据、检查点与推理”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“批量大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边

界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 `artifact = weights + config + preprocessing + version`，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当批量大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

3. 机制与因果链

沿着数据进入、内部计算、输出形成和外部行动逐步追踪，而不是只记名词。对“PyTorch 工程：数据、检查点与推理”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低

可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“批量大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{artifact} = \text{weights} + \text{config} + \text{preprocessing} + \text{version}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当批量大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. 评估与不确定性

把平均指标拆成切片、置信区间、校准、错误代价和最坏情况。对“PyTorch 工程：数据、检查点与推理”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“批量大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{artifact} = \text{weights} + \text{config} + \text{preprocessing} + \text{version}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当批量大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

9. 工程与复现

固定版本、环境、种子和测量记录，使别人能从原始输入重建结论。对“PyTorch 工程：数据、检查点与推理”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。。接着逐项标记观察到的事实、由

公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用校园图书馆的公开借阅类别建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把公开天文观测的缩小样例作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“批量大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{artifact} = \text{weights} + \text{config} + \text{preprocessing} + \text{version}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在校园图书馆的公开借阅类别中，当批量大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

11. 迁移与研究

把本章方法搬到新情境，用对照和消融判断哪些部分可迁移、哪些需要重做。对“PyTorch 工程：数据、检查点与推理”而言，第一步不是背下定义，而是把核心问题写成一句能被反驳、能被测量的话：掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。。接着逐项标记观察到的事实、由公式推导的结果、来自文献的结论和仍然不确定的部分。这样做能防止把界面效果、单个案例或流畅文字误当成稳定能力。模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。这个类比只帮助建立直觉，不能代替形式定义；当类比与公式冲突时，以经过核实的定义、实验和边界为准。

先用合成音频波形与环境声标签建立最小案例。只保留一个输入、一个可解释的中间量和一个输出，写清数据是“教学模拟”还是“实际测量”。然后建立一个不使用本章复杂方法的基线，例如固定规则、均值、随机策略或人工清单。基线不是为了显得简单，而是为了回答：如果什么都不优化，最低可接受表现是什么？再把合成电商目录但不含真实用户行为作为迁移案例，检查相同表示和假设是否仍成立。若必须改变数据定义、损失或评价标准，就把改变记录为新的实验条件，而不是悄悄调整。

本层实验只调节“批量大小”。先预测它变小时结果、成本和风险会怎样，再移动滑块核对教学模拟曲线。至少记录低、中、高三个点，不能只保留最好看的点。若曲线与预测不符，依次排查单位、边界、符号方向、初始化、随机性和实现。若曲线符合预测，也不能直接声称现实有效：页面只验证公式在简化条件下的行为，真实结论仍需要代表性数据、独立测试和部署观察。

常见错误是一次同时改变数据、模型和阈值，最后无法知道改进来自哪里。正确做法是建立实验矩阵：行表示唯一改变的变量，列记录假设、设置、种子、运行环境、主要指标、切片指标、资源成本和失败样例。对于 $\text{artifact} = \text{weights} + \text{config} + \text{preprocessing} + \text{version}$ ，逐个说明符号的对象、形状、单位和取值范围；把零值、极大值、空输入、重复输入和相反方向代入，检查是否符合直觉。任何无法解释的异常都应保留在失败日志，而不是删除。

完成本层的标准不是“看懂了”，而是你能不看正文讲出三件事：本章方法解决什么问题；它依赖哪些假设；怎样设计一个能让它失败的测试。再向同学解释为什么模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。只是入口，并给出一个类比失效的地方。如果解释中出现“肯定”“完

全”“智能”“理解”等宽泛词，就把它替换成可观察行为和适用条件。最后写一句研究问题：在合成音频波形与环境声标签中，当批量大小变化时，哪个机制最可能先成为瓶颈，怎样用对照实验验证？

7. PyTorch 实现路线

当本章涉及学习模型时，统一采用 Python 与 PyTorch：先把输入变成带明确形状的张量，再定义 `torch.nn.Module`，选择与任务一致的损失，使用 `optimizer.zero_grad()`、`loss.backward()` 和 `optimizer.step()` 完成更新。评估阶段使用 `model.eval()` 与 `torch.inference_mode()`，并保存 `state_dict`、预处理配置、词表或标签表。若本章属于规则、搜索、数据库或治理主题，则 PyTorch 不是强行套用的答案；应保留可解释的 Python 基线，在研究线中再比较神经方法。

形状检查要写在代码里：批次轴、特征轴、序列轴、通道轴不能靠猜。训练曲线至少同时记录训练损失、验证指标和资源成本；只保存最好一次会产生选择偏差。出现 NaN 时先检查输入范围、除零、对数和指数，再检查学习率、梯度范数与混合精度。出现训练好而测试差时先排查泄漏、重复样本、切分方式和分布偏移。

8. 实验设计与结果记录

为本章建立一张实验表。固定问题定义、数据版本、基线、硬件、依赖和随机种子，只改变“批量大小”。每个设置至少运行三次教学模拟；真实研究应根据方差和成本决定更充分的重复次数。结果栏必须标记为“实际测量”“教学模拟”或“文献引用”之一，三者不能混写。

推荐记录字段：实验 ID、日期、假设、唯一改变、数据哈希、代码提交、Python/PyTorch 版本、设备、种子、主要指标、切片指标、时间、内存、失败样例、是否符合预期、下一步。若采用外部结果，记录作者、标题、DOI 或 URL、访问日期、实验环境和原文限制；不要把论文中的一个数字搬到不同数据与硬件上直接比较。

9. 常见失败与排查顺序

- 275. **问题定义错位。** 指标提高但现实任务没有改善。重新写输入、决策和失败代价。
- 276. **数据泄漏或重复。** 训练与测试共享样本、时间或主体信息。按真实部署边界重新切分。
- 277. **表示不一致。** 训练和推理使用不同预处理、词表、单位或类别顺序。把预处理随模型版本化。
- 278. **只看平均值。** 少数群体、极端场景或长尾失败被平均掩盖。增加切片、区间和最坏情况。
- 279. **一次改变太多。** 无法归因改进来源。回到单变量实验和消融。
- 280. **把流畅当正确。** 生成结果自然但证据、逻辑或计算错误。增加引用核对、约束和拒答。
- 281. **部署条件变化。** 延迟、内存、网络、传感器或用户行为不同。建立监控、降级与回滚。

排查时按“数据→目标→实现→优化→评估→部署→治理”顺序逐层缩小范围。先构造最小可复现失败，再看日志和中间量；不要通过删除失败样本让图表变好。涉及人身、医疗、法律、财务、教育机会或其他高风险决策时，本章练习不能替代专业判断，也不能直接上线。

10. 迁移任务

工程线：把本章最小实现包装成一个可重复运行的命令，加入输入校验、测试、环境文件、日志、模型或规则版本、CPU 资源上限和 README。选择校园图书馆的公开借阅类别作为公开或合成案例，提供一个正常输入、一个边界输入和一个应拒绝的输入。

研究线：选择虚构游戏角色的状态日志，提出一个可证伪问题。复现简单基线，再改变一个机制；至少做一次消融、一次分布变化测试和一次失败类别分析。报告效应大小与不确定性，不只报告最好分数。若结果没有改善，也要解释它排除了什么假设。

11. 三级练习

- 282. **基础题：** 用自己的话解释“PyTorch 工程：数据、检查点与推理”，说明它的输入、输出和一个关键假设；再指出生活类比的一个局限。

283. **应用题：** 在公开道路网络与虚构出行任务中设计一个只改变“批量大小”的实验，列出基线、三档设置、指标和失败样例。

284. **挑战题：** 设计一个会让本章方法失效的分布变化或对抗条件，给出检测、拒答、人工复核和回滚方案。

独立答案位于 `content/ai-book/answers/pytorch-engineering.md`，网页中默认折叠，建议先完成再核对。

12. 本章能力检查

- 我能在两分钟内说清本章解决的问题与不解决的问题。
- 我能解释核心公式中每个符号的对象、形状和单位。
- 我能运行最小代码，并预测改变一个变量的方向性影响。
- 我能区分教学模拟、实际测量和文献引用。
- 我能给出至少三种失败原因，并按顺序排查。
- 我能说明何时必须停止自动化并交给给人。
- 我能提出一个可证伪研究问题和最低成本基线。

13. 来源与核实

本章来源于下列一手框架、官方文档、开放教材或研究论文的结构化核对；正文为原创解释，不复制原图、习题答案或大段原文。时效内容在每个版本重新核实。

- [S010] Aston Zhang et al., 《Dive into Deep Learning》，[访问 d2l.ai](https://d2l.ai))
- [S015] PyTorch, 《Automatic Differentiation with torch.autograd》，[访问 pytorch.org](https://pytorch.org))
- [S025] PyTorch, 《PyTorch Tutorials》，[访问 pytorch.org](https://pytorch.org))

核实日期：2026-08-06。评审状态：draft。只有技术评审、事实核查和专家签字完成后，状态才能按 `draft` → `technical_review` → `fact_checked` → `stable` 前进。

月 12 阶段项目：训练并发布一个 CPU 可运行的小型神经网络

项目结论： 从张量、数据加载、模型、优化、检查点到推理脚本形成可复现实验包。当前状态为 draft。

交付物

- 问题与边界说明
- 可运行最小实现
- 实验记录与失败分析
- 来源与许可清单
- 风险、隐私与人工闸门
- 复现 README

验收方法

285. 在 CPU 或浏览器主线运行；若使用 GPU，另提供缩小数据冒烟测试。
286. 保存环境、依赖、随机种子、数据许可、测量记录和失败样例。
287. 逐条标记实际测量、教学模拟和文献引用，写清适用边界。
288. 不收集个人信息；高风险输出必须有人审阅、可拒答、可回滚。

独立练习答案

先完成题目、保留自己的推理和错误，再查看参考答案。答案给出最低证据结构，不是唯一正确表述。

第 25 章答案 · 状态空间与 BFS、DFS、UCS

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，明确状态、动作、转移、目标和代价，比较宽搜、深搜与一致代价搜索。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“在迷宫里每个路口是状态，每次移动是动作；搜索策略决定先查看哪条岔路。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成的植物叶片特征，先用固定规则或简单统计量作为基线。把“分支因子”设置为 2、3 和 10，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 26 章答案 · 启发式与 A*：用估计减少盲目探索

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解可采纳、一致启发式和搜索效率之间的关系。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“导航会同时考虑已经走过的距离和到终点的估计距离。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构游戏角色的状态日志，先用固定规则或简单统计量作为基线。把“启发式权重”设置为 0、1 和 2，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 27 章答案·约束满足、回溯与 SAT

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，把变量、值域、约束和传播分开，利用顺序与剪枝缩小组合爆炸。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“排课不是生成所有课表再看冲突，而是每安排一门就传播教师、教室和时间限制。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开天文观测的缩小样例，先用固定规则或简单统计量作为基线。把“约束密度”设置为 0%、40%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 28 章答案·规划、博弈树与对手

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解动作前提、效果、效用、极小极大和剪枝，区分确定与随机对手。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“下棋不仅要自己想自己下一步，还要模拟对手最合理的回应。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成音频波形与环境声标签，先用固定规则或简单统计量作为基线。把“搜索深度”设置为 1 层、4 层和 12 层，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 29 章答案 · 命题逻辑与一阶逻辑

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解语法、语义、满足、量词和推理规则，区分真值与现实证据。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“逻辑像精确的积木接口：每块如何连接很清楚，但材料是否来自现实仍需证据。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园能源使用的匿名汇总数据，先用固定规则或简单统计量作为基线。把“规则数量”设置为 1 条、12 条和 100 条，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 30 章答案 · 知识图谱、本体与语义关系

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，建模实体、关系、类别、约束和溯源，理解开放世界假设。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“图书馆目录把作者、作品、主题连成网络；没有记录不等于事实为假。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开古诗文本与版权已明确的材料，先用固定规则或简单统计量作为基线。把“图连接率”设置为 0%、30%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 31 章答案 · 贝叶斯网、马尔可夫网与 HMM

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，用图表达条件独立性，在序列中区分隐藏状态和观测。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“隔墙听到天气声音是观测，真正天气是隐藏状态；前后时刻相互关联。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构机器人在二维网格中的轨迹，先用固定规则或简单统计量作为基线。把“转移噪声”设置为 0、0.15 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 32 章答案 · 决策论与概率程序

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，把不确定推断与偏好、代价、行动连接，理解抽样近似。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“医生式决策不只问最可能是什么，还问不同错误的后果；本书只用虚构低风险案例。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开科研论文的标题与摘要元数据，先用固定规则或简单统计量作为基线。把“误判代价”设置为 0、40 和 100，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 33 章答案 · 从问题定义到数据切分

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，把业务问题变成输入、目标、损失和决策，并防止训练测试泄漏。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“期末考试题不能提前出现在练习册答案里；测试集必须模拟真正未见情况。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成电商目录但不含真实用户行为，先用固定规则或简单统计量作为基线。把“测试集比例”设置为 5%、20%和 50%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 34 章答案 · 线性回归与逻辑回归

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解线性组合、最小二乘、对数几率、正则化和系数解释边界。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“调音台把多个输入按权重混合；逻辑函数再把分数压到零到一。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园图书馆的公开借阅类别，先用固定规则或简单统计量作为基线。把“决策阈值”设置为 0、0.5 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 35 章答案 · 决策树、随机森林与梯度提升

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解递归切分、袋装、随机特征和逐步纠错，比较偏差与方差。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“多位评委独立打分再汇总，往往比只问一位更稳，但共同偏见仍会保留。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择不含个人信息的教学模拟天气记录，先用固定规则或简单统计量作为基线。把“树深度”设置为 1 层、6 层和 30 层，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 36 章答案 · 间隔、核方法与泛化

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，从最大间隔理解 SVM，用核技巧表达非线性，并把复杂度与数据量联系。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“在两群人之间画分界线时，离双方都远的线通常更稳。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开道路网络与虚构出行任务，先用固定规则或简单统计量作为基线。把“正则强度”设置为 0、0.1 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 37 章答案·聚类：发现结构但不发明真相

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，比较 K 均值、层次聚类和密度方法，理解尺度、距离和簇数选择。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“整理一箱乐高可按颜色、形状或用途分组；不同距离定义会得到不同答案。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成的植物叶片特征，先用固定规则或简单统计量作为基线。把“簇数量”设置为 2 个、4 个和 20 个，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 38 章答案 · 密度估计与异常检测

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解概率密度、邻域、孤立和阈值，区分罕见与有害。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“图书借阅异常可能只是节假日或新书活动，罕见记录需要情境核查。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构游戏角色的状态日志，先用固定规则或简单统计量作为基线。把“异常阈值”设置为 0%、80%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 39 章答案 · 降维、流形与可视化陷阱

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，比较 PCA 与非线性嵌入，避免把二维图上距离直接当现实因果。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“把地球压到平面地图必然变形；不同投影保留面积、角度或距离中的一部分。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开天文观测的缩小样例，先用固定规则或简单统计量作为基线。把“嵌入维度”设置为 2 维、8 维和 64 维，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 40 章答案 · 自监督任务与表示学习

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解遮盖、对比、预测和数据增强如何构造无需人工标签的监督信号。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“把句子遮住一个词让模型猜，答案来自数据自身，而不是人工逐条标注。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成音频波形与环境声标签，先用固定规则或简单统计量作为基线。把“遮盖比例”设置为 0%、15%和 90%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 41 章答案 · 指标、阈值、校准与不确定性

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，根据错误代价选择指标，区分排序、分类、概率质量和置信表达。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“火警器调得太敏感会频繁误报，太迟钝会漏报；阈值必须结合后果。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园能源使用的匿名汇总数据，先用固定规则或简单统计量作为基线。把“决策阈值”设置为 0、0.5 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 42 章答案 · 实验设计与因果推断

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，区分相关、预测和干预，理解随机实验、混杂、反事实和识别假设。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“带伞和下雨相关，但带伞不会导致下雨；共同原因是天气预报。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开古诗文本与版权已明确的材料，先用固定规则或简单统计量作为基线。把“混杂强度”设置为 0%、30%和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 43 章答案 · 公平、偏差与数据分布变化

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，识别测量偏差、代表性偏差、历史偏差和不同公平指标间冲突。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“统一尺码不一定公平；不同使用情境需要明确目标、受影响者和补救方式。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择虚构机器人在二维网格中的轨迹，先用固定规则或简单统计量作为基线。把“少数组占比”设置为 0%、50% 和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 44 章答案 · 鲁棒性、对抗样本与分布外检测

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，测试噪声、最坏扰动、未知类别和环境变化，建立拒答与降级策略；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“自动门在晴天好用不代表雨雪、遮挡和断电时仍可靠。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开科研论文的标题与摘要元数据，先用固定规则或简单统计量作为基线。把“输入噪声”设置为 0%、10% 和 100%，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 45 章答案 · 感知机与多层网络

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解线性层、非线性激活、隐藏表示和通用近似的实际边界。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“多道门把原始特征逐步重组；没有非线性，多层线性门仍等价于一层。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择合成电商目录但不含真实用户行为，先用固定规则或简单统计量作为基线。把“隐藏层宽度”设置为 2、32 和 256，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 46 章答案 · 反向传播与自动微分实现

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，沿计算图应用链式法则，理解梯度累积、detach 和原地操作风险。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“总成绩由作业、测验和考试组成，反向传播追踪每一项对最终误差的贡献。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择校园图书馆的公开借阅类别，先用固定规则或简单统计量作为基线。把“网络深度”设置为 1 层、12 层和 100 层，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 47 章答案 · 初始化、归一化、正则化与训练诊断

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，理解梯度尺度、过拟合、权重衰减、dropout、归一化和学习曲线。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“训练像练球：难度太低会只记动作，强度太高会受伤；需观察长期曲线而非单次成绩。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择不含个人信息的教学模拟天气记录，先用固定规则或简单统计量作为基线。把“正则强度”设置为 0、0.1 和 1，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上

一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

第 48 章答案 · PyTorch 工程：数据、检查点与推理

答案用于自检，不是唯一表述。先保留自己的推理，再比较遗漏。

基础题参考

合格答案应包含三部分：第一，掌握 Dataset、DataLoader、Module、state_dict、设备和推理模式。；第二，输入、内部机制和输出之间的关系；第三，至少一个适用假设。生活类比“模型文件像乐谱，不含演奏环境和乐器说明就难以复现声音。”能提供直觉，但它省略了数学约束、数据偏差和部署风险，因此不能作为实验证据。若只写概念定义、没有边界或反例，答案仍不完整。

应用题参考

选择公开道路网络与虚构出行任务，先用固定规则或简单统计量作为基线。把“批量大小”设置为 1、32 和 256，其余条件保持一致。记录主要指标、资源成本、至少一个切片和失败样例。页面结果必须写“教学模拟”；如果自己运行真实代码，则保存环境、数据版本、种子和测量记录后才能写“实际测量”。

挑战题参考

可以主动改变输入分布、加入噪声、制造缺失、交换标签含义或改变部署资源。检测层使用输入校验、漂移或分布外分数；决策层设置置信阈值和拒答；流程层对高风险输出要求人工复核；运维层保留上

一版本、变更记录和回滚触发条件。关键不是列工具名，而是说明每道控制对应哪个失败路径、由谁负责、怎样验证控制有效。

自评量规

- 0 分：只复述名词。
- 1 分：说明了机制，但没有实验或边界。
- 2 分：有基线、单变量、指标和失败分析。
- 3 分：进一步包含不确定性、迁移、人工闸门和可复现记录。

本卷来源台账

本卷共引用 15 条来源。正文为原创解释；不复制受版权保护的图表、习题答案或大段原文。访问日期均为当前 Beta 版核实记录。

[S010] Dive into Deep Learning

项目	内容
作者/机构	Aston Zhang et al.
来源等级	开放教材
发布日期	持续更新
访问日期	2026-08-06
许可	CC BY-SA 4.0
DOI/URL	访问 d2l.ai

[S015] Automatic Differentiation with torch.autograd

项目	内容
作者/机构	PyTorch
来源等级	官方教程
发布日期	持续更新

项目	内容
访问日期	2026-08-06
许可	BSD-3-Clause
DOI/URL	访问 pytorch.org

[S016] Artificial Intelligence: A Modern Approach

项目	内容
作者/机构	Stuart Russell, Peter Norvig
来源等级	经典教材官方站
发布日期	持续更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 aima.cs.berkeley.edu

[S017] CS221: Artificial Intelligence

项目	内容
作者/机构	Stanford University

项目	内容
来源等级	大学课程一手资料
发布日期	持续更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 stanford-cs221.github.io

[S018] OWL 2 Web Ontology Language Document Overview

项目	内容
作者/机构	W3C
来源等级	Web 标准一手资料
发布日期	2012
访问日期	2026-08-06
许可	W3C Document License
DOI/URL	访问 w3.org

[S019] Probabilistic Graphical Models

项目	内容
作者/机构	Daphne Koller, Nir Friedman
来源等级	经典教材出版信息
发布日期	2009
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 mitpress.mit.edu

[S020] scikit-learn User Guide

项目	内容
作者/机构	scikit-learn developers
来源等级	官方文档
发布日期	持续更新
访问日期	2026-08-06
许可	BSD-3-Clause
DOI/URL	访问 scikit-learn.org

[S021] Machine Learning Crash Course

项目	内容
作者/机构	Google
来源等级	官方公开课程
发布日期	持续更新
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 developers.google.com

[S022] Causal Inference: The Mixtape

项目	内容
作者/机构	Scott Cunningham
来源等级	开放教材
发布日期	2021
访问日期	2026-08-06
许可	CC BY-NC-SA 4.0
DOI/URL	访问 mixtape.scunning.com

[S023] Fairlearn User Guide

项目	内容
作者/机构	Fairlearn contributors
来源等级	开源项目官方文档
发布日期	持续更新
访问日期	2026-08-06
许可	MIT
DOI/URL	访问 fairlearn.org

[S024] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

项目	内容
作者/机构	NIST
来源等级	标准机构一手资料
发布日期	2025
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名

项目	内容
DOI/URL	访问 csrc.nist.gov

[S025] PyTorch Tutorials

项目	内容
作者/机构	PyTorch
来源等级	官方教程
发布日期	持续更新
访问日期	2026-08-06
许可	BSD-3-Clause
DOI/URL	访问 pytorch.org

[S030] Attention Is All You Need

项目	内容
作者/机构	Ashish Vaswani et al.
来源等级	同行评审论文预印本
发布日期	2017

项目	内容
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S071] Model Cards for Model Reporting

项目	内容
作者/机构	Margaret Mitchell et al.
来源等级	同行评审论文预印本
发布日期	2018
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

[S072] Datasheets for Datasets

项目	内容
----	----

项目	内容
作者/机构	Timnit Gebru et al.
来源等级	同行评审论文预印本
发布日期	2018
访问日期	2026-08-06
许可	以来源页许可为准；本书仅作原创解释、短引用与明确署名
DOI/URL	访问 arxiv.org

全书术语表

共 81 个核心术语。定义用于建立共同语言，不能替代章节中的数学、实验和边界分析。

术语	英文	小白解释
人工智能	Artificial Intelligence	让机器执行通常需要感知、推理、学习、规划或生成等能力的研究与工程领域。
算法	Algorithm	把输入转换为输出的一组明确步骤；算法本身不等于训练后的模型。
模型	Model	从数据或规则中得到、用于预测或决策的数学结构及其参数。
参数	Parameter	训练过程中被优化的数值，例如神经网络权重。
超参数	Hyperparameter	由开发者或搜索过程设定、不是直接从单次梯度更新学得配置。
数据集	Dataset	带有明确来源、字段、许可和用途的一组样本。

术语	英文	小白解释
特征	Feature	模型使用的输入表示。
标签	Label	监督学习中希望模型预测的目标值。
训练	Training	用目标函数和优化方法调整模型参数的过程。
推理	Inference	用已经得到的模型对新输入计算输出。
损失函数	Loss function	把预测与目标之间的差异变成可优化数值的函数。
目标函数	Objective	训练或决策希望最大化或最小化的量，可能包含多个损失和约束。
梯度	Gradient	函数对多个变量的一阶变化率组成的向量。
泛化	Generalization	模型在未参与训练、但与目标场景相关的数据上保持能力。
过拟合	Overfitting	模型过度适应训练数据细节，导致新数据表现下降。

术语	英文	小白解释
欠拟合	Underfitting	模型或训练不足以捕捉任务中的重要结构。
正则化	Regularization	通过惩罚、噪声或结构限制降低过拟合风险的方法。
基线	Baseline	用于判断新方法是否真正改进的简单或既有方法。
消融实验	Ablation	移除或替换系统一部分，观察它对结果的贡献。
随机种子	Random seed	初始化伪随机数生成器的值；固定种子不是完整可复现性的替代品。
张量	Tensor	带有一个或多个轴的数值数组。
嵌入	Embedding	把离散对象映射到连续向量空间的表示。
注意力	Attention	根据查询与键的匹配程度加权汇总值的机制。
Transformer	—	以注意力、前馈层、残差和归一化为核心的序列模型架构。

术语	英文	小白解释
词元	Token	文本经过词元器切分后的基本单位，可是字符、子词或其他片段。
上下文窗口	Context window	模型一次计算中能够直接处理的词元范围。
基础模型	Foundation model	在广泛数据上训练、可适配多种下游任务的大规模模型。
语言模型	Language model	为词元序列分配概率或预测后续词元的模型。
检索增强生成	RAG	在生成前检索外部资料，并把证据提供给生成模型的系统模式。
Agent	—	围绕目标执行观察、计划、工具调用、验证和停止循环的系统。
提示注入	Prompt injection	不可信输入试图改变模型或 Agent 原定指令和权限的攻击。
强化学习	Reinforcement learning	智能体通过与环境交互、基于

术语	英文	小白解释
		奖励学习策略的方法。
策略	Policy	在给定状态下选择动作的规则或概率分布。
价值函数	Value function	从状态或状态动作对出发的预期累计回报。
马尔可夫决策过程	MDP	用状态、动作、转移、奖励和折扣描述序列决策的形式框架。
卷积	Convolution	用共享局部核在输入位置上滑动计算的运算。
计算机视觉	Computer vision	让机器从图像或视频中提取、生成和使用信息的领域。
自然语言处理	NLP	让计算机处理、理解、检索或生成人类语言的领域。
自动语音识别	ASR	把语音波形转换为文字或其他符号序列。
扩散模型	Diffusion model	通过学习逐步去噪过程生成样本的一类模型。

术语	英文	小白解释
生成对抗网络	GAN	通过生成器与判别器对抗训练学习生成分布的架构。
图神经网络	GNN	在图结构上通过邻域聚合或消息传递学习表示的模型。
校准	Calibration	预测概率与事件长期发生频率之间的一致程度。
分布偏移	Distribution shift	训练、评估与部署数据分布不一致的现象。
分布外	OOD	与训练或预期部署分布明显不同的输入。
因果推断	Causal inference	在明确假设下估计干预对结果影响的方法。
公平性	Fairness	围绕不同群体或个体的机会、错误、待遇和影响制定并审查的规范目标。
鲁棒性	Robustness	系统面对扰动、噪声、故障或环境变化仍保持可接受行为的能力。
可解释性	Explainability	为模型行为提供可理解说明的

术语	英文	小白解释
		方法和性质；解释清楚不等于预测正确。
差分隐私	Differential privacy	通过限制单条记录对输出分布的影响提供可量化隐私保证的框架。
联邦学习	Federated learning	在数据留在多个设备或组织的情况下协同训练的范式。
模型卡	Model card	记录模型用途、评估、限制和 risk 的结构化文档。
数据表	Datasheet	记录数据来源、构建、用途、限制和维护方式的结构化文档。
MLOps	—	管理机器学习数据、训练、部署、监控、版本和事故响应的工程实践。
延迟	Latency	一次请求从进入到完成所需的时间。
吞吐	Throughput	单位时间内系统能处理的请求或词元数量。

术语	英文	小白解释
量化	Quantization	用较低精度数值表示权重或激活以降低资源成本的方法。
蒸馏	Distillation	让较小模型学习较大模型输出或内部结构的方法。
并行训练	Distributed training	把训练计算、数据或参数分布到多个设备或进程。
检查点	Checkpoint	保存训练状态以便恢复、评估或部署的版本化工件。
可复现	Reproducible	他人依据充分工件和说明能够得到相容结果的性质。
同行评审	Peer review	由相关领域同行对方法、证据和表述进行独立审查。
置信区间	Confidence interval	按指定重复抽样程序构造、具有长期覆盖率解释的区间。
p 值	p-value	在零假设及检验假设成立时，观察到至少同样极端统计量的概率。
效应量	Effect size	差异或关联实际大小的度量。

术语	英文	小白解释
熵	Entropy	概率分布不确定性的度量。
交叉熵	Cross entropy	用一个分布编码另一个分布时的平均信息量，也是常见分类损失。
KL 散度	KL divergence	衡量一个概率分布相对另一个分布差异的非对称量。
贝叶斯更新	Bayesian update	用似然把先验信念更新为后验信念。
知识图谱	Knowledge graph	用实体、关系和属性构成的结构化知识网络。
本体	Ontology	对领域概念、关系和约束的显式规范。
搜索启发式	Heuristic	估计距离或代价、用于决定优先探索方向的函数。
约束满足	CSP	为变量选择值，使所有约束同时满足的问题。
SLAM	—	机器人同时定位并建立环境地图的问题。

术语	英文	小白解释
控制	Control	根据目标与反馈选择输入，使动态系统按期望演化。
仿真到现实	Sim-to-real	把在仿真中得到的方法迁移到真实系统的过程及差距问题。
世界模型	World model	学习环境状态和动作如何随时间演化的内部模型。
多模态	Multimodal	联合处理文本、图像、音频、视频或动作等多种信息形式。
AI for Science	—	使用 AI 辅助科学建模、模拟、实验设计、分析和发现的方向。
人机协同	Human-in-the-loop	在系统关键环节保留明确人工检查、判断或控制权的设计。
风险	Risk	事件发生可能性与后果的组合，必须结合具体情境评估。

版权、版本与勘误

2026.08 Beta · 第二卷：经典 AI 与机器学习 · 评审状态 draft。

Copyright © 2026 NEWVAR LAB. All rights reserved.

全部原创正文、实验定义和代码保留全部版权。第三方名称、论文与链接仅作明确署名的教学引用，其权利归原权利人。

本书仍处于 Beta：可能存在表达、事实、代码或排版问题。请通过 NEWVAR 版本页查看核实日期、评审状态与更正记录。

任何时效性结论都应在使用前回到一手来源重新核实；医疗、法律、财务、教育机会和安全关键场景必须由合格专业人员判断。

学习网站：[访问 newvarlab.com](https://newvarlab.com)）。