论文领读与课堂测试

机器学习导引 • 第二章模型评估与选择（配套材料）

Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning[J/OL]. arXiv:1811.12808, 2018.

一、读懂论文所需的基础概念（对照第二章）

概念	论文中的体现	对应第二章
经验误差 vs 泛化误差	用训练集评估是「乐观偏差」，评估应针对未见数据	二、经验误差与过拟合
错误率与准确度	$ACC = 1 - ERR$，用 0-1 损失定义 ERR	四 • 4.1 错误率与准确度
过拟合 / 欠拟合	训练集上高准确 = 记忆而非泛化；高方差 → 过拟合、高偏差 → 欠拟合	二 • 2.2
留出法 hold-out	2/3 训练、1/3 测试；小数据集不推荐	三 • 3.1 留出法
分层 stratification	划分时保持各类别比例，避免破坏统计独立性	三、评估方法的讲究
k 折交叉验证	k 次迭代、每样本恰验证一次、取均值	三 • 3.2 交叉验证
留一法 LOO	$k = n$：$n - 1$ 训练、1 验证；适合小数据集	三 • 3.2 的特例
重复留出法 Monte Carlo CV	重复 k 次随机划分取平均，反映模型稳定性	三、评估方法（进阶）
引导法 bootstrap	有放回重采样估计采样分布，量化不确定性	三、评估方法（进阶）
袋外样本 OOB	每轮未被抽中的约 36.8% 样本，作测试集	三、评估方法（进阶）
.632 / .632+ 引导	固定 / 自适应权重修正偏差	五、偏差与方差
经验置信区间	百分位法取 2.5% / 97.5% 分位数	（不确定性估计）
参数 vs 超参数	权重/截距 = 模型参数；正则强度/最大深度 = 超参数	一 • 1.2 参数与超参数
三路划分 train/val/test	three-way holdout 的六步流程	一 • 1.5 / 三阶段
评估 vs 选择 vs 算法选择	性能估计的三个子任务	一 • 1.4 评估与选择
偏差与方差	统计 bias/variance；折数 k 的权衡	五、偏差与方差
嵌套交叉验证	内层选超参、外层估泛化，减小选择偏差	三 • 3.3 调参（防数据泄露）
统计检验	McNemar、5x2cv F 检验等	（进阶：算法选择）

二、课堂测试题

（一）选择题：模型评估、选择与算法选择

Q1 论文中，预测准确度 ACC 与预测错误率 ERR 的关系是？

A. $ACC = 1 - ERR$　B. $ACC = ERR$　C. $ACC + ERR = \frac{1}{2}$　D. 二者无关

Q2 论文用「0-1 损失」定义错误率，0-1 损失的含义是？

A. 预测错误记 1、正确记 0　B. 错误率取值在 0 到 1 之间

C. 损失函数是线性的　D. 只能用于二分类

Q3 论文 1.3 节指出，用「训练模型的那份数据」来评估模型，这种做法叫？

A. 留出法　B. 再代入验证（resubstitution validation）　C. 交叉验证　D. 引导法 bootstrap

Q4　线性回归里的「斜率（权重系数）」和「截距（bias / y 轴截距项）」，论文把它们归为？
A. 超参数　B. 模型参数（学出来的）　C. 既非参数也非超参数　D. 统计量

Q5　下列哪一项是论文明确列举的「超参数」例子？
A. 线性回归的斜率　B. 决策树的最大深度　C. 线性回归的截距项　D. 0-1 损失

Q6　three-way holdout（三路留出法）中，训练集 / 验证集 / 测试集的分工是？
A. 训练 = 拟合、验证 = 选超参、测试 = 最终评估
B. 训练 = 评估、验证 = 拟合、测试 = 选超参
C. 训练 = 选超参、验证 = 评估、测试 = 拟合

Q7　k 折交叉验证中，每个样本会被用作「验证集」恰好几次？
A. 0 次　B. 1 次　C. $k-1$ 次　D. k 次

Q8　当折数 $k=n$（折数等于样本数）时，k 折交叉验证退化为？
A. 留出法　B. 留一法（LOO）　C. 引导法 bootstrap　D. 嵌套交叉验证

Q9　论文 3.5 节指出，k 太小时会带来什么？
A. 悲观偏差与方差都增大　B. 偏差与方差都减小　C. 只增加计算量　D. 没有任何影响

Q10　嵌套交叉验证（nested CV）中，内层循环与外层循环分别负责？
A. 内层估泛化、外层选超参　B. 内层选超参（模型选择）、外层估泛化
C. 内层和外层都选超参　D. 内层和外层都估泛化

（二）选择题：引导法与不确定性

Q11　论文第 2 章里，bootstrap（引导法）生成「新样本」的方式是？
A. 有放回抽样　B. 无放回抽样　C. 按类别分层抽样　D. 按时间顺序抽样

Q12　从 n 个样本中有放回地抽 n 次，某个样本在本次 bootstrap 样本中「没有被抽中」的概率约为？
A. $1/e \approx 36.8\%$　B. $1/2$　C. 63.2%　D. $1/n$

Q13　在 Leave-One-Out Bootstrap（LOOB）里，模型是在哪份数据上评估性能的？
A. 训练集本身　B. 袋外样本（out-of-bag，OOB）　C. 全数据集　D. 一个额外独立收集的新数据集

（三）简答题

Q17　用第二章的「训练集学、验证集选、测试集考」和「三阶段」，简述论文 3.3 节三路留出法的核心思想；并说明其中「用独立测试集只测一次」对应我们强调的哪条红线。

Q18　论文 1.1 节把性能估计分成三个子任务。请把它们与第二章的「评估（Evaluation）vs 选择（Selection）」对应起来，并指出「算法选择」对应论文的哪一章。

三、论文领读讨论题

1. **三任务与三阶段**：论文 1.1 节列出的性能估计三个子任务，如何对应你讲义里的「评估 vs 选择」以及「阶段一（Q1+Q2 循环）→ 阶段二（封存）→ 阶段三（Q3 启用）」？

2. **参数 vs 超参数**：论文 1.2 节如何区分「模型参数」与「超参数」？它把线性回归的截距项（bias/截距）归为哪一类？这与我们「b 的值是参数、要不要 bias 项是超参数」的结论是否一致？

3. **折数 k 的偏差-方差权衡**：论文 3.5 节为何说 k 太小会「悲观偏差与方差都增大」？为什么实验上 $k = 10$ 是个常用折中？联系第二章的偏差-方差。

4. **测试集只用一次**：三路留出法第 4–5 步为何「选完模型后合并训练 + 验证重训，再用独立测试集只测一次」？这对应第二章的哪条红线（数据泄露）？

5. **嵌套交叉验证**：为什么「用同一份验证集既选超参又报性能」会引入偏差？嵌套 CV 用内、外两层循环如何解决？（联系第二章「验证集选、测试集考」的分离原则。）

6. **留出法 vs 交叉验证**：论文结论（Figure 23）为何「小数据集不推荐留出法、推荐 k 折或留一法」？用第二章留出法的局限解释。

四、参考答案与解析

选择题

Q1 A——$ACC = 1 - ERR$，二者互补，和为 1。

Q2 A——0-1 损失：预测正确记 0、错误记 1，对所有样本取平均即错误率。

Q3 B——再代入验证：用训练集本身评估，会带来乐观偏差（过拟合），对应第二章「经验误差不是越小越好」。

Q4 B——论文明确把「权重系数（斜率）与其 bias 项（y 轴截距）」列为**模型参数**（由学习算法拟合到训练数据）。

Q5 B——决策树最大深度是超参数；斜率、截距是模型参数；0-1 损失既非参数也非超参数。

Q6 A——训练集拟合、验证集选超参、测试集最终评估（只测一次）。

Q7 B——每个样本在 k 轮里恰好充当验证集一次，这是 k 折交叉验证的核心性质。

Q8 B——$k = n$ 即留一法（LOOCV），每次用 $n - 1$ 个样本训练、剩下的 1 个验证。

Q9 A——k 太小则训练数据少，悲观偏差增大，且模型对划分方式更敏感、方差也增大。

Q10 B——内层循环做超参调优（模型选择），外层循环估计泛化性能。

Q11 A——引导法是有放回抽样；重复留出法才是无放回抽样。

Q12 A——$P(\text{未抽中}) = (1 - \frac{1}{n})^n \to 1/e \approx 36.8\%$。

Q13 B——LOOB 用袋外样本（OOB）作测试集，避免「训练集本身评估」带来的乐观偏差。

简答题

Q17　三路留出法把数据分成训练/验证/测试三份：在训练集上拟合（对应「训练集学」），在验证集上反复比较超参、选出最优配置（对应「验证集选」与「阶段一」循环），选好后再合并训练 + 验证重训（对应「阶段二 • 封存」），最后用从未见过的测试集**只测一次**（对应「阶段三 • Q3 启用」与「测试集考」）。「测试集只测一次」正是为了守住**数据泄露**这条红线——反复偷看测试集会得到过于乐观的估计。

Q18 论文的三个子任务：(1) 估计泛化性能、(2) 通过调超参选出最优模型、(3) 比较并选出最适合的算法。(1)(2) 对应第二章的「选择（Selection）」，(1) 的最终报告又对应「评估（Evaluation）」；(3)「算法选择」对应论文第 4 章（Algorithm Comparison，含 McNemar、5x2cv、嵌套交叉验证等）。

讨论题要点

- **1 三任务与三阶段**：子任务 (1)（估泛化）→ 阶段三 Q3；子任务 (2)（选最优模型）→ 阶段一 Q1+Q2 循环；阶段二「封存」对应论文三路留出法的第 4 步（选完不再动模型）。
- **2 参数 vs 超参数**：论文把「拟合到数据」的权重、截距归为模型参数，把「算法旋钮」正则强度、树最大深度归为超参数；与我们「b 的值是参数、要不要 bias 是超参数」完全一致。
- **3 折数 k**：k 小则训练数据少、估计悲观（偏差大），且对划分敏感（方差大）；$k = 10$ 在偏差-方差与计算量间折中，是经验上的甜点值。
- **4 测试集只用一次**：反复用测试集选模型会让「测试集泄露信息」，估计过于乐观；合并训练 + 验证重训能减少悲观偏差，测试集保持独立只测一次。
- **5 嵌套 CV**：同一份验证集既选超参又报性能，等于把模型选择当成训练的一部分，引入选择偏差；嵌套 CV 用内层选超参、外层估泛化，让外层评估基于「未参与选择」的数据，从而接近无偏。
- **6 留出法 vs 交叉验证**：小数据下留出法浪费数据、估计不稳定（方差大），且训练数据不足（悲观偏差）；k 折/留一法让每个样本既训练又验证，充分利用数据，估计更稳。