机器学习导引

第二章　模型评估与选择（讲义）

主讲：何新卫

一、从哪里说起：为什么要评估模型

1.1 一个导入问题：去掉 bias 会更好吗？

先从一个具体问题切入：对于一个线性分类任务，**去掉 bias（偏置 / intercept）项的模型与没有去掉 bias 的模型，哪一个会更好？**

- 模型 1（有 bias）：$y = w_1 x + b$；
- 模型 2（无 bias）：$y = w_1 x$。

同学们可能会有两种直觉：**「有 bias 的更好」**——因为模型更灵活，能拟合不通过原点的直线，理论上能力更强；**「无 bias 的更好」**——因为模型更简单，可能更不容易过拟合。那么，正确答案是什么？这个问题没法凭感觉回答，而且它其实牵出「参数」与「超参数」的区别——先把这对概念分清，再谈如何评估。

1.2 参数与超参数：学出来的 vs 设计出来的

回到导入问题的两个模型，它们的**唯一差别**就是「有没有 bias 项」。

- **参数（Parameter）**：模型在训练中**自动学出来**的量，由训练数据驱动、经优化算法（如梯度下降）不断更新。线性模型 $z = \mathbf{x}W^\top + b$ 里的权重 W 与偏置 b 都是参数——不是人写死的，而是「学」出来的；
- **超参数（Hyperparameter）**：训练开始前由人**手动设计/设定**的量，**不能**直接从训练数据中学出，需靠经验或验证集来调。例如学习率、网络层数 M、每层神经元数、**是否保留 bias 项**、正则化系数等。

	参数（Parameter）	超参数（Hyperparameter）
由谁决定	训练数据 + 优化算法「学」出来	人手动设计 / 设定
何时确定	训练过程中不断更新	训练前就固定
能否从数据学出	能（这就是「学习」本身）	不能，需用验证集比较挑选
典型例子	权重 W、偏置 b	学习率、层数 M、神经元数、是否加 bias、正则化系数

一句话口诀：**参数是「学」出来的，超参数是「设计」出来的。**于是导入问题可以拆成两层：b 本身是**参数**（若保留，会被学习更新）；而「要不要这个 bias 项」是**超参数**层面的结构选择。至于「去掉 bias 与保留 bias 到底哪个更好」，这个对比问题仍然留着——它正是本章要用评估方法去回答的核心问题。

1.3 回顾：Mitchell 定义与三个问题

第一章给出了 Mitchell（1997）的形式化定义：给定任务 T 与性能度量 P，利用经验 E 优化 θ，使 h_θ 在 T 上的性能（以 P 衡量）随 E 提升。围绕这个定义，一个完整的学习流程需要回答三个问题：

- **Q1 如何选择模型、如何优化参数？**——训练集、损失函数、优化算法；

- **Q2 学习率等超参数设置问题？**——用验证集调优；
- **Q3 如何评估训练好的模型好坏？**——用测试集最终评估（只用一次）。

本章聚焦 Q3，但它其实是 Q1、Q2 的「裁判」：没有一个可靠的评估方法，就无法判断哪种模型、哪组超参数更好。这个「裁判」作用体现在完整的评估流程里，可以分成三个阶段：

- **阶段一（Q1 + Q2 的循环）**：在训练集上调参数（Q1），在验证集上选超参数（Q2）。这个「训练-验证」的循环可以反复进行无数次，直到满意为止；
- **阶段二（终局）**：彻底停止所有对模型参数和超参数的调整，此时模型被「封存」、状态固定；
- **阶段三（Q3 的启用）**：只有在这个「冻结」时刻之后，才将从未见过的测试集输入模型，一次性算出最终性能指标。

1.4 评估与选择：以数字字符分类为例

以手写数字识别为例，把 Mitchell 定义中的三个要素落到具体对象上：

- **任务** T：手写数字识别（10 分类），输入 x 为 28×28 灰度图，输出 $y \in \{0, \ldots, 9\}$；
- **经验** E：训练集，大量带标签的样本对 $\{(x_i, y_i)\}$；
- **模型选择**：线性层，$\theta = (W, b)$；
- **性能度量** P：测试集上的分类准确率（accuracy）。

于是「评估」与「选择」有了清晰的区分：

- **评估（Evaluation）**：回答「这个训练好的模型到底有多好」，用 P 在 **测试集**上衡量 h_θ 的性能。测试集**只用一次**，保证评价的是泛化能力，而不是记忆能力；
- **选择（Selection）**：回答「多个候选模型/超参数里，挑哪一个最好」。比如换学习率、换层数、换模型结构，得到多个 h_θ；用**验证集**反复比较，挑 P 最优的那个。

1.5 三个集合：训练集、验证集、测试集

集合	英文	作用	对应三问
训练集	Training Set	优化模型参数 θ（在这里「学」）	Q1
验证集	Validation Set	调超参数、选模型结构（可复用）	Q2
测试集	Test Set	最终评估泛化性能（只用一次）	Q3

一句话：**训练集体现学习能力，测试集反映泛化能力**；测试集性能越高，模型越好。

二、经验误差与过拟合

2.1 泛化误差与经验误差

- **经验误差（Empirical Error）**：学习器在训练集上的误差；
- **泛化误差（Generalization Error）**：学习器在「未见」（Unseen）样本上的误差。

一个关键问题：**经验误差是不是越小越好？**答案是 **No！**经验误差过小，可能意味着模型**过拟合**——它把训练数据里的噪声和随机波动也当成了规律来学习。

2.2 欠拟合、过拟合与合适的拟合

- **欠拟合（Underfitting）**：相较于数据而言，模型参数过少或结构过于简单，无法捕捉到数据中的规律；
- **过拟合（Overfitting）**：模型过于紧密或精确地匹配特定数据集，以致于无法良好地拟合其他数据或预测未来的观察结果；
- **合适的拟合（Appropriate fitting）**：模型能够恰当地拟合和捕捉到数据中的规律。

直观表现：欠拟合时训练误差和测试误差都高；过拟合时训练误差很低、测试误差却很高；合适拟合时两者都低，且差距不大。

三、评估方法

3.1 留出法（hold-out）

留出法直接将数据集 D 划分为两个互斥的集合：训练集 S 和测试集 T。在 S 上训练出模型后，用 T 来评估其测试误差。

示例：以二分类为例，假设 D 包含 1000 个样本，将其划分为 S（700 个样本）与 T（300 个样本）。用 S 训练后，若模型在 T 上有 90 个样本分类错误，则错误率为

$$\frac{90}{300} = 30\%.$$

注意：单次留出的估计受划分方式影响较大（不同随机划分结果会有波动），这是它的主要局限。

3.2 交叉验证法（cross validation）

交叉验证将数据集划分为 k 个大小相近的互斥子集，每次用其中 $k-1$ 个子集训练、剩下的 1 个子集作验证，共进行 k 轮，最后取 k 次验证结果的**平均**作为估计。

示例：假设某个学习算法在四个验证集合上的识别率分别为 50%、100%、50%、50%，则平均识别率为

$$\frac{50\% + 100\% + 50\% + 50\%}{4} = 62.5\%.$$

交叉验证比单次留出更充分利用数据，估计也更稳定，代价是计算开销更大。

3.3 调参与最终模型

实际工作流是「**训练集学、验证集选、测试集考**」：在训练集上优化参数，在验证集上反复比较并挑选超参数与模型结构，最后**只用一次**测试集给出最终泛化性能。任何反复「偷看」测试集的做法都会引入**数据泄露**，使评估失真。

四、性能度量

选择合适的评估指标能够帮助我们比较模型好坏。通常指标取值范围在 $[0, 1]$ 之间的连续实数，一类指标是**越高越好**（如分类精度），一类指标是**越低越好**（如分类误差）。常见的评估指标有三类：错误率与准确度、精度/召回率/F1、ROC 与 AUC。

4.1 错误率与准确度（error rate & accuracy）

- **错误率（error rate）**：分类错误的样本数占样本总数的比例；
- **准确度（accuracy）**：分类正确的样本数占样本总数的比例。

设 m 个样本中有 a 个分类错误，则

$$\text{error rate} = \frac{a}{m}, \qquad \text{accuracy} = 1 - \frac{a}{m}.$$

错误率越低越好，准确度越高越好。二者在**类别平衡**时很直观，但在类别不平衡时可能产生误导（见 4.2 的讨论）。

4.2 精度、召回率与 F1（precision, recall & F1）

以二分类为例（多分类类似），用混淆矩阵四个量描述预测结果：

- **TP（真阳性）**：正例被正确判为正例；
- **FP（假阳性）**：负例被误判为正例；
- **FN（假阴性）**：正例被误判为负例；
- **TN（真阴性）**：负例被正确判为负例。

混淆矩阵可画成下面这张 2×2 表（行 = 实际类别，列 = 预测类别）：

	预测为正例	预测为负例
实际为正例	TP（真阳性）	FN（假阴性）
实际为负例	FP（假阳性）	TN（真阴性）

由此定义：

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

精度衡量「预测为正的样本里有多少是真的正例」，召回率衡量「真正的正例里有多少被找到」。

一个具体例子（识别狗）：测试集共 100 张图片，其中狗 40 张（正例）、猫 60 张（负例）。模型把 50 张判为狗、50 张判为猫，结果填入混淆矩阵：

	预测为狗	预测为猫	合计
实际为狗	$TP = 30$	$FN = 10$	40
实际为猫	$FP = 20$	$TN = 40$	60
合计	50	50	100

于是可逐一算出：

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{30}{30 + 20} = 0.6 = 60\%, \qquad \text{Recall} = \frac{TP}{TP + FN} = \frac{30}{30 + 10} = 0.75 = 75\%,$$

$$F1 = \frac{2PR}{P + R} = \frac{2 \times 0.6 \times 0.75}{0.6 + 0.75} = \frac{2}{3} \approx 66.7\%, \qquad \text{Accuracy} = \frac{TP + TN}{\text{总数}} = \frac{30 + 40}{100} = 70\%.$$

为什么单独用都不够：通过检索所有项目，可以简单获得完美的召回率；通过只选择极少数极有可能的项目，也可以实现完美的精度。因此通常把二者结合为单一指标，最常用的是 **F1**（精度与召回率的调和平均）：

$$F1 = \frac{2 \cdot P \cdot R}{P + R},$$

其取值在 $[0,1]$ 之间，越高越好。

4.3 PR 曲线（PR-curve）

PR 曲线（Precision-Recall 曲线）显示不同阈值下精度与召回率的权衡；曲线下面积越大，表示精度和召回率都越高。比较两条 PR 曲线时，通常是「包住」对方的那条所对应的算法更优。

4.4 ROC 与 AUC

ROC（Receiver Operating Characteristic），中文译为「受试者工作特征」，最早源于第二次世界大战中用于敌机检测的雷达信号分析技术，之后被引入机器学习领域。它用两个量刻画分类器在不同阈值下的表现：

- **真正例率**：$TPR = \dfrac{TP}{TP+FN}$（即召回率）；
- **假正例率**：$FPR = \dfrac{FP}{FP+TN}$。

以 FPR 为横轴、TPR 为纵轴画出的曲线即 ROC 曲线；**AUC（Area Under Curve）**是 ROC 曲线下的面积，取值在 $[0,1]$ 之间：$AUC = 1$ 表示完美分类，$AUC = 0.5$ 等价于随机猜测，因此 **AUC 越大越好**。

五、偏差与方差

5.1 模型复杂度

复杂的模型拟合能力更强。以多层线性层为例，把单层「线性层」扩展为 M 层：

$$\text{CustomLinear}(28\times 28, 512) \to \text{CustomLinear}(512, 512) \to \cdots \to \text{CustomLinear}(512, 10).$$

这里 M 是一个超参数（hyperparameter）：M 增大，模型参数增多，模型复杂度（Model Complexity）也随之增加。两个自然的问题：为什么每层 Linear 之后要插入非线性层（如 tanh、ReLU）？M 取 0、1 还是 100 更好？

5.2 偏差-方差权衡（bias-variance trade-off）

对于给定的机器学习任务，选择合适复杂度的模型非常必要：

- **复杂模型（Complex model）**：参数多、结构复杂，容易**过拟合**（easily overfit）——把训练数据中的噪声和随机波动也当作规律，导致在测试集上表现很差；
- **简单模型（Simple model）**：参数少，无法捕捉数据中的重要特征和规律，训练集与测试集上表现都不佳（**欠拟合**，underfit）。

总误差可分解为**偏差（bias）**与**方差（variance）**之和。随着模型复杂度升高，偏差下降、方差上升，总误差呈 U 形：左端是**欠拟合区（Underfitting zone）**，右端是**过拟合区（Overfitting zone）**，U 形底部即复杂度合适、总误差最小的位置。目标是找到一个「既能准确捕捉训练数据中的规律，又能良好泛化到未见数据」的模型。

注意（易混淆）：这里的「bias」和线性层里的「bias」**不是**同一个概念——前者是误差分解中的**偏差**，后者是模型里的**偏置项**（一个可学习参数）。二者同名但含义完全不同，注意区分。

六、本章小结

- **一对概念**：参数（学出来的，如 W、b）vs 超参数（设计出来的，如学习率、层数 M、是否加 bias）；
- **两个误差**：经验误差（训练集）vs 泛化误差（测试集，只用一次）；
- **一个红线**：过拟合 = 训练误差低、测试误差高；欠拟合 = 两者都高；
- **两类评估方法**：留出法（hold-out）、k 折交叉验证（cross validation）；
- **两类评估指标**：错误率与准确度、精度/召回率/F1 与 ROC/AUC；
- **一个权衡**：偏差-方差（bias-variance trade-off），决定模型复杂度选择；
- **一个延伸问题**：目标检测中的 mAP（mean Average Precision）如何计算？

七、课堂自测题

7.1 选择题

Q1 以下哪一项属于**超参数**？
A. 线性层的权重 W　B. 线性层的偏置 b　C. 学习率　D. 模型对某样本的预测输出

Q2 某模型训练精度为 99%，测试误差为 72%，它最可能处于什么状态？
A. 欠拟合　B. 过拟合　C. 合适的拟合　D. 无法判断

Q3 5 折交叉验证，每折在验证集上的错误率分别为 10%、20%、30%、20%、20%，则平均错误率为？
A. 18%　B. 20%　C. 22%　D. 25%

Q4 疾病筛查测试集 1000 人：990 人健康、10 人患病；模型把所有人都预测为「健康」。该模型的 accuracy（准确度）约为？
A. 99%　B. 90%　C. 50%　D. 0%

Q5 随着模型复杂度增加，偏差（bias）和方差（variance）分别如何变化？
A. 偏差下降、方差上升　B. 偏差上升、方差下降　C. 都下降　D. 都上升

7.2 参考答案与解析

Q1 C——W、b 是「训练数据 + 优化算法学出来」的参数；学习率是训练前由人设定、不能从数据学出的超参数；D 是输出，既非参数也非超参数。
Q2 B——训练精度 99%（训练误差约 1%）、测试误差 72%，训练集上几乎全对、测试集上很差，是典型过拟合；欠拟合是两者都差。
Q3 B——平均的是「每折的错误率」$= (10\% + 20\% + 30\% + 20\% + 20\%)/5 = 20\%$，不是「总错误样本数 ÷ 总样本数」。
Q4 A——accuracy $= 990/1000 = 99\%$；但 99% 毫无意义，它把 10 个病人都漏诊了，此时应看召回率（recall$= 0$）或 F1——类别不平衡下 accuracy 会「骗人」。
Q5 A——复杂度升高，偏差下降、方差上升，总误差 U 形；务必记住高偏差 = 欠拟合（左端）、高方差 = 过拟合（右端）。

八、推荐阅读

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

附：中英文术语对照

参数	Parameter	超参数	Hyperparameter
偏置项	Bias / Intercept Term	经验误差	Empirical Error
泛化误差	Generalization Error	欠拟合	Underfitting
过拟合	Overfitting	合适的拟合	Appropriate Fitting
模型复杂度	Model Complexity	留出法	Hold-out
交叉验证	Cross Validation	k 折交叉验证	k-fold Cross Validation
调参	Hyperparameter Tuning	训练集 / 验证集 / 测试集	Training / Validation / Test Set
数据泄露	Data Leakage	错误率	Error Rate
准确度	Accuracy	精度	Precision
召回率	Recall	F1	F1 Score
混淆矩阵	Confusion Matrix	真阳性 / 假阳性	True/False Positive
真阴性 / 假阴性	True/False Negative	PR 曲线	Precision-Recall Curve
ROC 曲线	ROC Curve	受试者工作特征	Receiver Operating Characteristic
AUC	Area Under Curve	真正例率	True Positive Rate (TPR)
假正例率	False Positive Rate (FPR)	偏差	Bias
方差	Variance	偏差-方差权衡	Bias-Variance Trade-off
平均精度	mean Average Precision (mAP)	性能度量	Performance Measure