机器学习 第二章

模型评估与选择

汇报人：何思成

Agenda

01

部门概览与核心职能

部门组织架构与职责分工

02

团队协作与沟通有效性

团队协作的现状和问题分析

03

目标完成情况洞察

绩效完成情况的数据分析

04

探讨工作氛围与流程

工作环境和流程的问题分析

05

发展规划与改进策略

未达标KPI的改善与优化

经验误差和过拟合

学习器的实际预测输出和样本的真实输出之间的差异称为“误差”，学习器在训练集上的误差称为“训练误差”或“经验误差”，新样本上的误差称为“泛化误差”

过拟合：学习器有时由于学习能力过强，同时学习了训练样本中不太一般的特征，把训练样本自身的特征当做了所有潜在样本都有的特征，对训练样本预测性能很好但泛化能力差，即称为过拟合。 与过拟合相对的称为欠拟合。

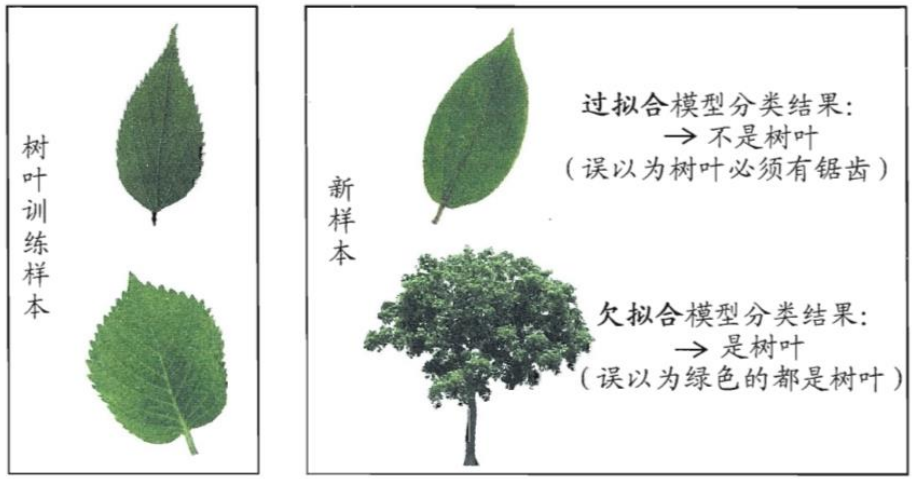


图 2.1 过拟合、欠拟合的直观类比

模型选择

针对不同的具体任务，我们可以选择不同的学习算法以及不同的参数配置，例如：决策树、支持向量机、随机森林、神经网络等等。

三个关键问题

1. 如何获得测试结果？ → 评估方法
2. 如何评估性能优劣？ → 性能度量
3. 如何判断算法实质差异？ → 比较检验

评估方法

在建模过程中，我们要用训练集来训练模型，用测试集检验其泛化能力

利用测试集得到的测试误差作为泛化误差的近似值

测试集应该与训练集“互斥”

如何从样本数据中产生训练样本和测试样本呢?

常见方法:

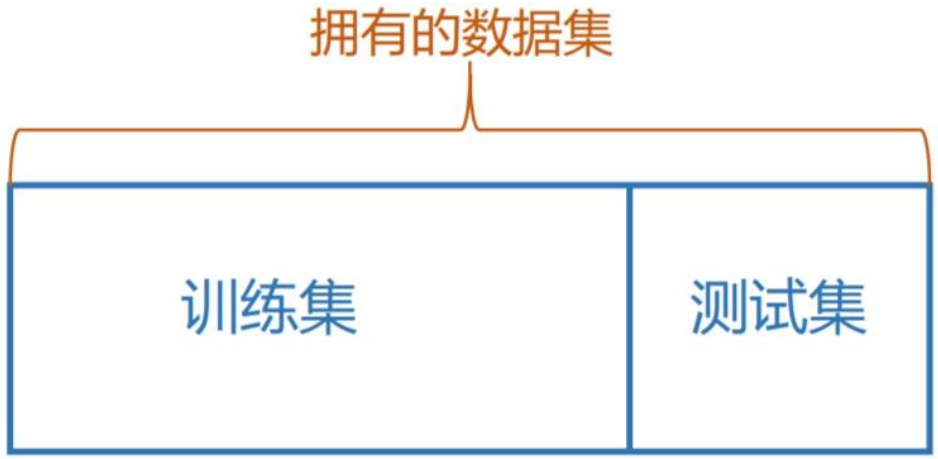
- 留出法 (hold-out)
- 交叉验证法 (cross validation)
- 自助法 (bootstrap)

评估方法

留出法

直接把数据集分成两个互斥的集合，一个作为训练集（S），一个作为测试集（T）

S的大小通常在样本总量的2/3~4/5，常用的是70%



注意：

- 保持数据分布一致性 （例如: 分层采样)
- 多次重复划分 (例如: 100次随机划分)
- 测试集不能太大、不能太小 (例如： 1/5~1/3)

评估方法

k-折交叉验证法

把数据集经分层抽样分成k个大小相似的互斥子集，每次用其中的k-1个子集的并集作为训练集，剩下那个用作测试集，最后返回k个测试结果的均值

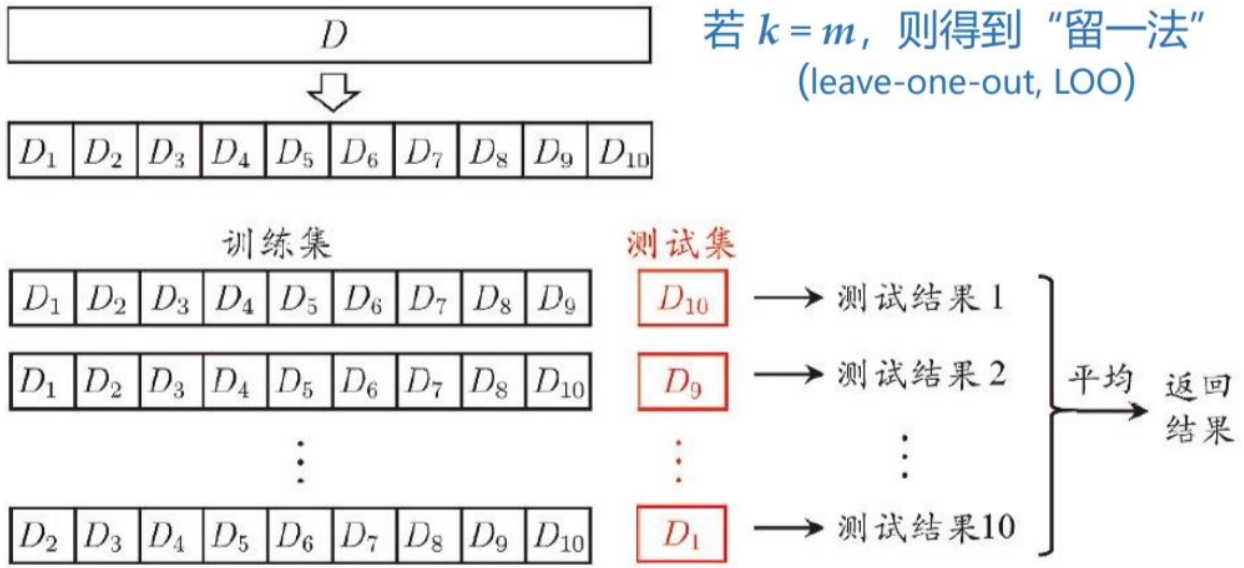


图 2.2　10 折交叉验证示意图

评估方法

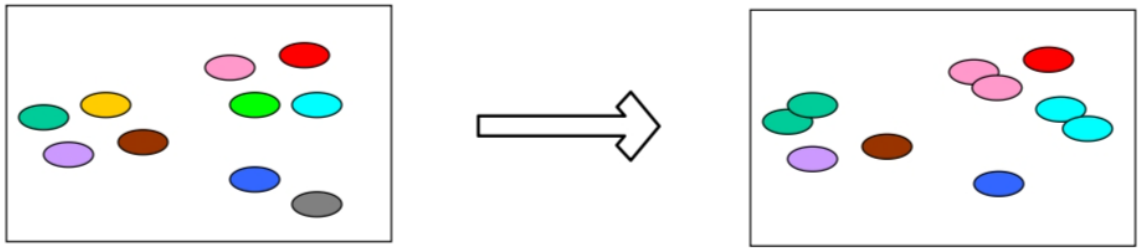
自助法

基于“自助采样” (bootstrap sampling)

亦称“有放回采样”、“可重复采样”

每次随机从样本总数据集中挑选一个样本放入数据集D’，进行有放回的随机抽样

数据集D’作为训练集，D\D’作为测试集



约有 36.8% 的样本不出现

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^{m}=\frac{1}{e}\approx 0.368$$

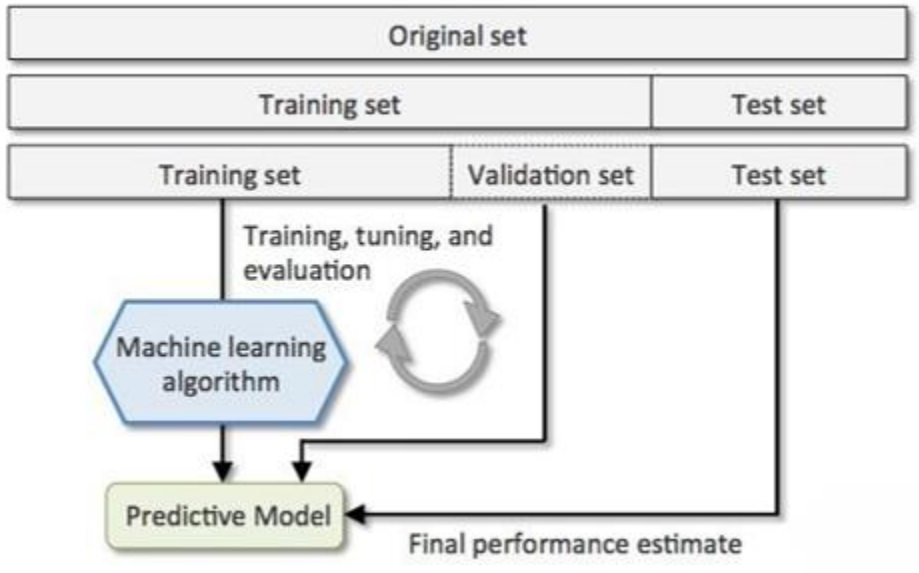
“包外估计” (out-of-bag estimation)

- 训练集与原样本集同规模
- 数据分布有所改变

自助法适用于数据集小、难以有效划分训练/测试样本时

评估方法

“调参”与最终模型



算法的参数：一般由人工设定，亦称“超参数”

模型的参数：一般由学习确定

调参过程相似：先产生若干模型，然后基于某种评估方法进行选择

参数调得好不好对性能往往对最终性能有关键影响

区别：训练集 vs. 测试集 vs. 验证集 (validation set)

算法参数选定后，要用“训练集+验证集”重新训练最终模型

模型评估时从训练集中划分出来用来调参的是验证集

评估模型实际使用的泛化能力的是测试集

性能度量

性能度量(performance measure)是衡量模型泛化能力的评价标准， 反映了任务需求
使用不同的性能度量往往会导致不同的评判结果

什么样的模型是“好”的，不仅取决于算法和数据，还取 决于任务需求

回归(regression) 任务常用均方误差

$$E(f;\mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}} (f(\boldsymbol{x}) - y)^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \ .$$

分类任务常用性能度量包括错误率、精度、查准率、查全率、F1

性能度量

错误率
vs.
精度

- 错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

错误率：分类错误的样本数(α)占样本总数(m)的比例，E=α/m

- 精度：

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

精度= 1- 错误率

性能度量

查准率
vs.
查全率

查准率：预测为正例中实际为正例的比例

查全率：实际为正例中被正确预测的比例

表 2.1　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- 查准率：$$P = \frac{TP}{TP + FP}$$
- 查全率：$$R = \frac{TP}{TP + FN}$$

PR曲线

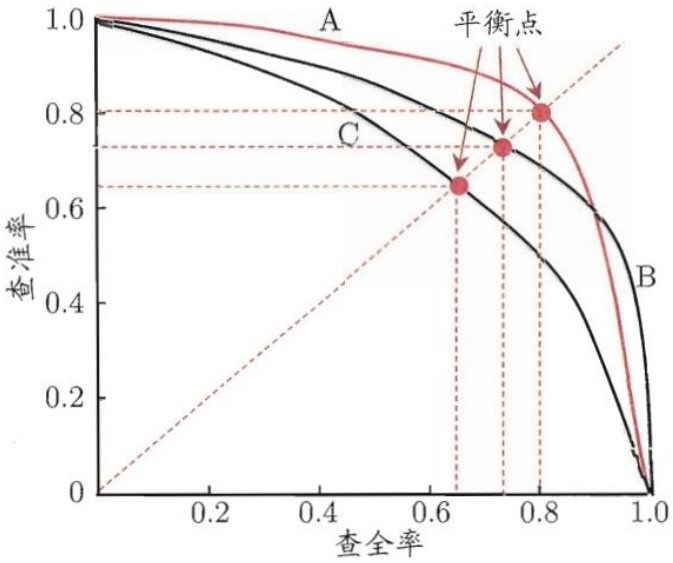


图 2.3　P-R曲线与平衡点示意图

性能度量

F1

$F1$ 是基于查准率与查全率的调和平均(harmonic mean)定义的:

$$\frac{1}{F1} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R} \right).$$

F_β 则是加权调和平均:

$$\frac{1}{F_\beta} = \frac{1}{1+\beta^2} \cdot \left(\frac{1}{P} + \frac{\beta^2}{R} \right).$$

与算术平均($\frac{P+R}{2}$)和几何平均($\sqrt{P \times R}$)相比, 调和平均更重视较小值.

F1 度量:

$$F1 = \frac{2 \times P \times R}{P + R} \qquad \frac{1}{F1} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R} \right)$$

$$= \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

若对查准率/查全率有不同偏好:

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R} \qquad \frac{1}{F_\beta} = \frac{1}{1+\beta^2} \cdot \left(\frac{1}{P} + \frac{\beta^2}{R} \right)$$

$\beta > 1$ 时查全率有更大影响; $\beta < 1$ 时查准率有更大影响

性能度量

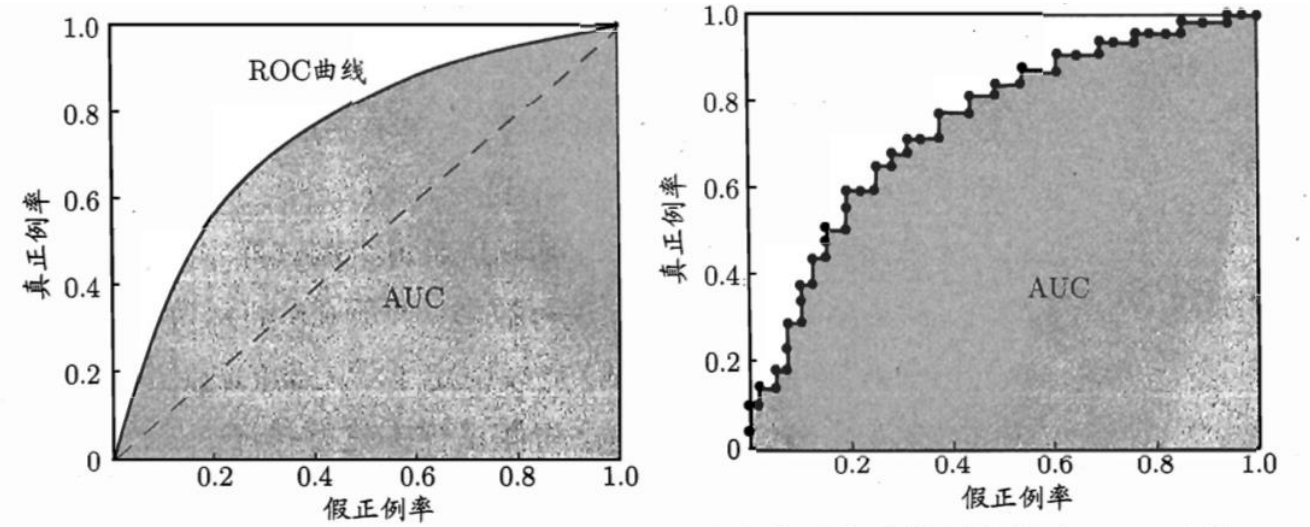
ROC（受试者工作特征）曲线是从选取分类阈值时样本排序本身的质量好坏的角度来研究学习器泛化性能的好坏

曲线纵坐标为真正例率（TPR），

横坐标为假正例率（FPR）

$$\mathrm{TPR} = \frac{TP}{TP + FN},$$

$$\mathrm{FPR} = \frac{FP}{TN + FP}.$$



(a) ROC 曲线与 AUC

(b) 基于有限样例绘制的 ROC 曲线与 AUC

图 2.4 ROC 曲线与 AUC 示意图

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1}).$$

性能度量

代价敏感错误率和代价曲线

为了衡量不同类型的错误造成的不同损失，将错误赋予“非均等代价”

表 2.2　二分类代价矩阵

真实类别	预测类别	
	第 0 类	第 1 类
第 0 类	0	$cost_{01}$
第 1 类	$cost_{10}$	0

代价敏感错误率计算公式：（D+是正例子集）

$$E(f;D;cost)=\frac{1}{m}\left(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{01}+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)\times cost_{10}\right).$$

在非均等代价下，ROC曲线不能直接反映出学习器的期望总体代价，需要用代价曲线来评估

代价曲线图的横轴是取值为[0,1]的正例概率代价

$$P(+)cost=\frac{p\times cost_{01}}{p\times cost_{01}+(1-p)\times cost_{10}}, \tag{2.24}$$

其中 p 是样例为正例的概率; 纵轴是取值为 $[0,1]$ 的归一化代价

$$cost_{norm}=\frac{\mathrm{FNR}\times p\times cost_{01}+\mathrm{FPR}\times(1-p)\times cost_{10}}{p\times cost_{01}+(1-p)\times cost_{10}}, \tag{2.25}$$

其中 FPR 是式(2.19)定义的假正例率, $\mathrm{FNR}=1-\mathrm{TPR}$ 是假反例率.

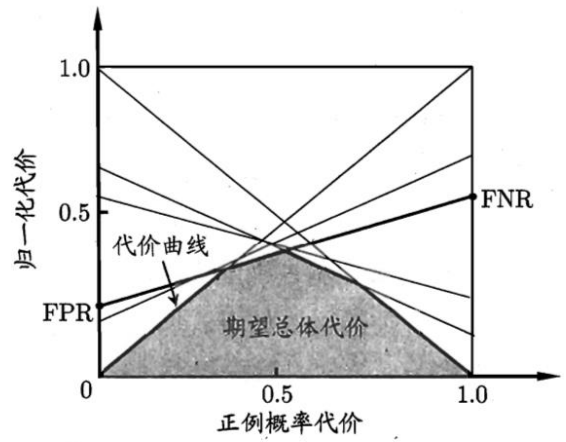


图 2.5　代价曲线与期望总体代价

比较检验

默认使用错误率（∈）为性能度量

在某种度量下取得评估结果后，是否可以直接比较以评判优劣？

NO！

因为：

- 测试性能不等于泛化性能
- 测试性能随着测试集的变化而变化
- 很多机器学习算法本身有一定的随机性

统计假设检验本质上是带有某种概率性质的反证法

- 首先假定原假设H_0（例如A算法优于B算法）和备择假设H1
- 然后确定显著水平α和样本容量
- 假定原假设成立，导出检验统计量的概率分布情况
- 确定拒绝域的临界值　　P= {拒绝H_0 | H_0为真}=α
- 根据置信度（1-α）做出接受或拒绝原假设的判断

比较检验

二项检验：已知数据样本服从二项分布，但是不知道这个分布的某个参数，提出假设判断样本二元变量是否服从某一假设

$$P(\hat{\epsilon};\epsilon)=\binom{m}{\hat{\epsilon}\times m}\epsilon^{\hat{\epsilon}\times m}(1-\epsilon)^{m-\hat{\epsilon}\times m}\,. \tag{2.26}$$

给定测试错误率, 则解 $\partial P(\hat{\epsilon};\epsilon)/\partial\epsilon=0$ 可知, $P(\hat{\epsilon};\epsilon)$ 在 $\epsilon=\hat{\epsilon}$ 时最大, $|\epsilon-\hat{\epsilon}|$ 增大时 $P(\hat{\epsilon};\epsilon)$ 减小. 这符合二项(binomial)分布, 如图 2.6 所示, 若 $\epsilon=0.3$, 则 10 个样本中测得 3 个被误分类的概率最大.

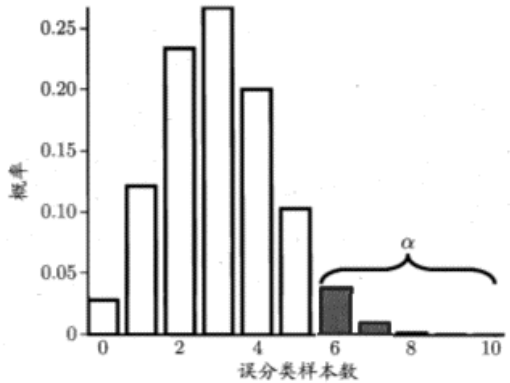


图 2.6 二项分布示意图($m=10,\epsilon=0.3$)

我们可使用“二项检验”(binomial test)来对“$\epsilon\leqslant 0.3$”(即“泛化错误率是否不大于 0.3”)这样的假设进行检验. 更一般的, 考虑假设“$\epsilon\leqslant\epsilon_0$”, 则在 $1-\alpha$ 的概率内所能观测到的最大错误率如下式计算. 这里 $1-\alpha$ 反映了结论的“置信度”(confidence), 直观地来看, 相应于图 2.6 中非阴影部分的范围.

$$\bar{\epsilon}=\max\epsilon\quad\text{s.t.}\quad\sum_{i=\epsilon_0\times m+1}^{m}\binom{m}{i}\epsilon^{i}(1-\epsilon)^{m-i}<\alpha\,. \tag{2.27}$$

t分布在应用中通常用于小样本或总体方差未知的情况下，而正态分布则更适用于大样本或总体方差已知的情况下。

“t 检验”(t-test). 假定我们得到了 k 个测试错误率, $\hat{\epsilon}_1,\hat{\epsilon}_2,\ldots,\hat{\epsilon}_k$, 则平均测试错误率 μ 和方差 σ^2 为

$$\mu=\frac{1}{k}\sum_{i=1}^{k}\hat{\epsilon}_i\,, \tag{2.28}$$

$$\sigma^2=\frac{1}{k-1}\sum_{i=1}^{k}(\hat{\epsilon}_i-\mu)^2\,. \tag{2.29}$$

考虑到这 k 个测试错误率可看作泛化错误率 ϵ_0 的独立采样, 则变量

$$\tau_t=\frac{\sqrt{k}(\mu-\epsilon_0)}{\sigma} \tag{2.30}$$

服从自由度为 $k-1$ 的 t 分布, 如图 2.7 所示.

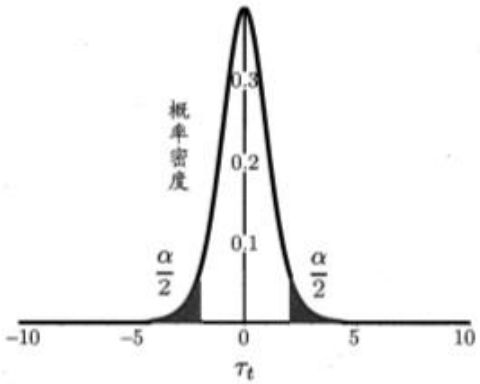


图 2.7 t 分布示意图($k=10$)

比较检验

在一个数据集上对不同学习器的性能进行比较

- 两学习器比较
  - 交叉验证 t 检验 (基于成对 t 检验)
    - k 折交叉验证； 5x2交叉验证
  - McNemar 检验 (基于列联表，卡方检验)

交叉验证t检验的基本思想是如果两个算法性能相同，则它们使用相同的训练/测试集得到的测试错误率相同。例如首先对算法A和B进行5次2折交叉验证，得到测试错误率后分别求差值（每次交叉验证将产生两对错误率），再进行t检验，计算出均值和方差

表 2.4 两学习器分类差别列联表

算法 B	算法 A	
	正确	错误
正确	e_{00}	e_{01}
错误	e_{10}	e_{11}

如果两个学习器性能相同，则A预测正确B预测错误数应该=B预测正确A预测错误数。

所以下面的变量服从自由度为1的卡方分布

若我们做的假设是两学习器性能相同, 则应有 $e_{01} = e_{10}$, 那么变量 $|e_{01} - e_{10}|$ 应当服从正态分布, 且均值为 1, 方差为 $e_{01} + e_{10}$. 因此变量

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \tag{2.33}$$

比较检验

在一组数据集上对多个算法进行比较

- 多学习器比较
  - Friedman + Nemenyi
    - Friedman检验 (基于序值，F检验; 判断”是否都相同”)
    - Nemenyi 后续检验 (基于序值，进一步判断两两差别)

如果假设“所有算法性能相同”被拒绝，则说明所有算法性能显著不同，这时需要通过Nemenyi检验进一步判断优劣

表 2.5 算法比较序值表

数据集	算法 A	算法 B	算法 C
D_1	1	2	3
D_2	1	2.5	2.5
D_3	1	2	3
D_4	1	2	3
平均序值	1	2.125	2.875

然后, 使用 Friedman 检验来判断这些算法是否性能都相同. 若相同, 则它们的平均序值应当相同. 假定我们在 N 个数据集上比较 k 个算法, 令 r_i 表示第 i 个算法的平均序值, 为简化讨论, 暂不考虑平分序值的情况, 则 r_i 服从正态分布, 其均值和方差分别为 $(k+1)/2$ 和 $(k^2-1)/12$. 变量

$$\begin{aligned}\tau_{\chi^2} &= \frac{k-1}{k}\cdot\frac{12N}{k^2-1}\sum_{i=1}^{k}\left(r_i-\frac{k+1}{2}\right)^2 \\ &= \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}r_i^2-\frac{k(k+1)^2}{4}\right)\end{aligned} \tag{2.34}$$

在 k 和 N 都较大时, 服从自由度为 $k-1$ 的 χ^2 分布.

然而, 上述这样的“原始 Friedman 检验”过于保守, 现在通常使用变量

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}, \tag{2.35}$$

比较检验

偏差（2.40）度量了学习算法的期望预测和真实结果的偏离程度，刻画的是学习算法本身的拟合能力

方差（2.38）度量了同样大小的训练集的变动所导致的学习性能的变化，刻画的是数据扰动所造成的影响

噪声（2.39）表达了在当前任务上任何学习算法能够达到的期望泛化误差下界，刻画的是学习问题本身的难度

使用样本数相同的不同训练集产生的方差为

$$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f\left(\boldsymbol{x}; D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] , \tag{2.38}$$

噪声为

$$\varepsilon^2 = \mathbb{E}_D\left[\left(y_D - y\right)^2\right] . \tag{2.39}$$

期望输出与真实标记的差别称为偏差(bias), 即

$$bias^2(\boldsymbol{x}) = \left(\bar{f}\left(\boldsymbol{x}\right) - y\right)^2 . \tag{2.40}$$

$$E(f; D) = bias^2\left(\boldsymbol{x}\right) + var\left(\boldsymbol{x}\right) + \varepsilon^2 , \tag{2.42}$$

也就是说, 泛化误差可分解为偏差、方差与噪声之和.

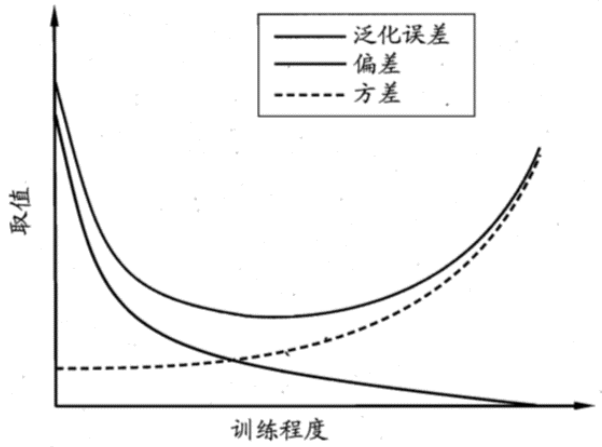
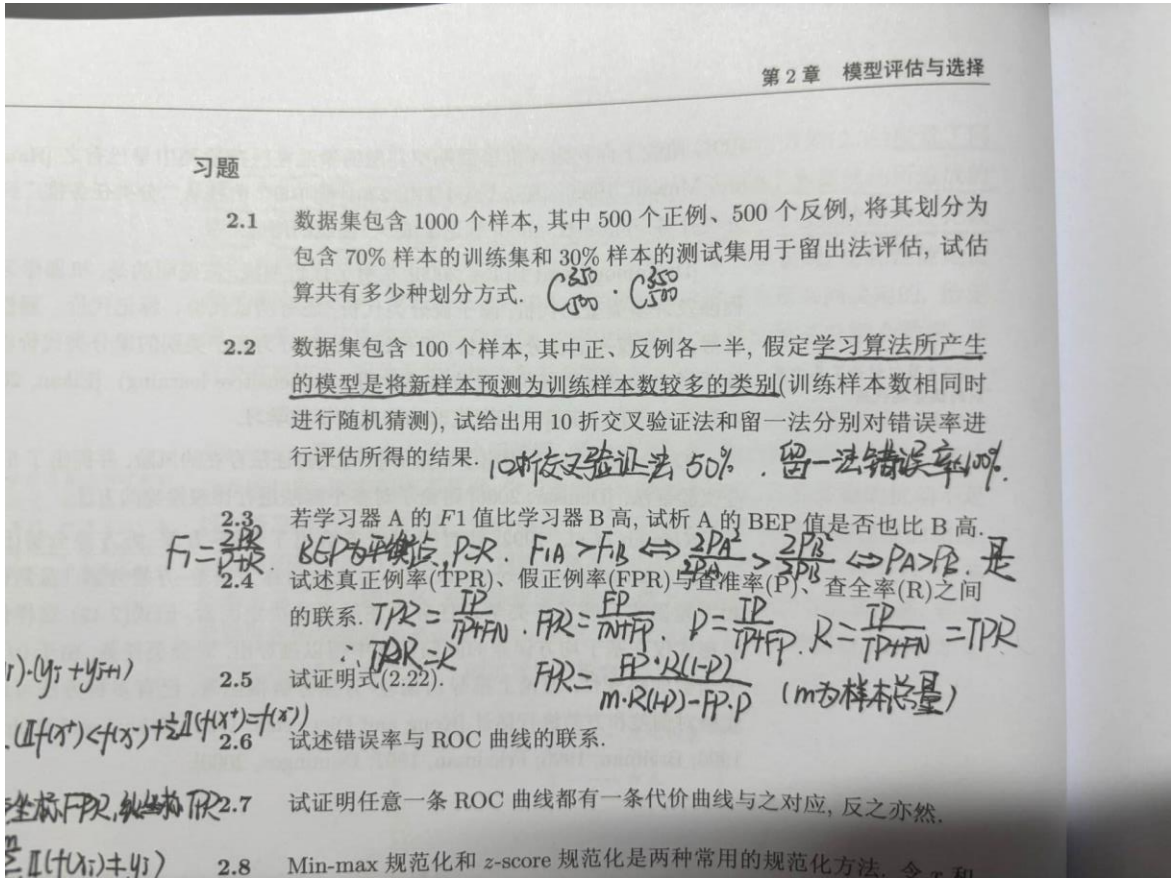


图 2.9 泛化误差与偏差、方差的关系示意图

习题



习题

2.1 数据集包含 1000 个样本, 其中 500 个正例、500 个反例, 将其划分为包含 70% 样本的训练集和 30% 样本的测试集用于留出法评估, 试估算共有多少种划分方式.

2.2 数据集包含 100 个样本, 其中正、反例各一半, 假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别(训练样本数相同时进行随机猜测), 试给出用 10 折交叉验证法和留一法分别对错误率进行评估所得的结果.

2.3 若学习器 A 的 $F1$ 值比学习器 B 高, 试析 A 的 BEP 值是否也比 B 高.

2.4 试述真正例率(TPR)、假正例率(FPR)与查准率(P)、查全率(R)之间的联系.

2.5 试证明式(2.22).

2.6 试述错误率与 ROC 曲线的联系.

2.7 试证明任意一条 ROC 曲线都有一条代价曲线与之对应, 反之亦然.

2.8 Min-max 规范化和 z-score 规范化是两种常用的规范化方法. 令 x 和

Thank You

www.officeplus.cn

Presenter name