机器学习理论及工程实践

第2章 模型评估与选择

徐文星

目录

第2章 模型评估与选择

基本概念

评估方法

性能度量

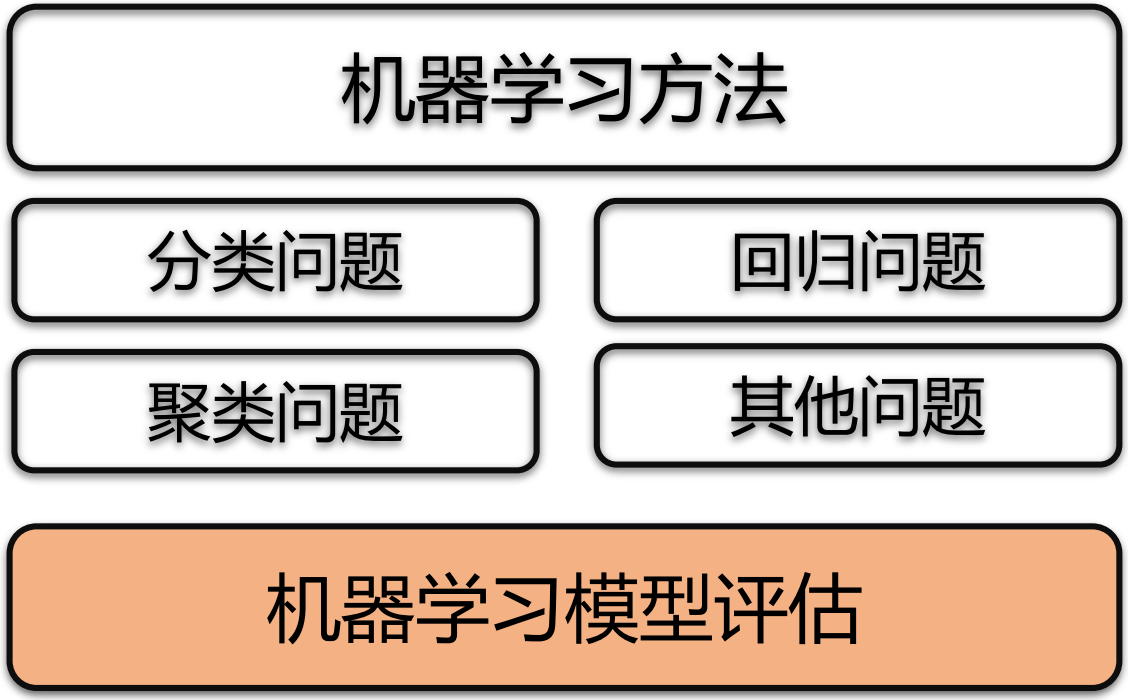
比较检验

2.1 基本术语

数据集	data set	样本/示例	sample/instance
属性/特征	attribute/feature	维数	dimentionality
学习/训练	learning/training	测试	testing
泛化	generalization	验证	validation
分类	classification	聚类	clustering
回归	regression	预测	prediction
监督学习	supervised learning	无监督学习	unsupervised learning
半监督学习	semi-supervised learning	强化学习	reinforcement learning
归纳	induction	演绎	deduction
精度	accuracy	错误率	error rate
欠拟合	underfitting	过拟合	overfitting

2.1 基本概念

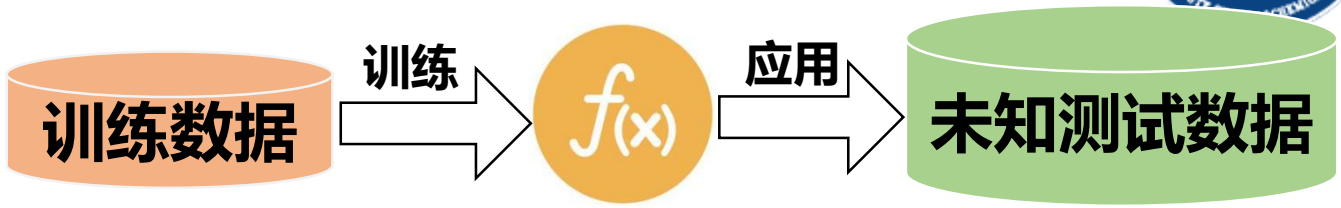
2.1 基本概念



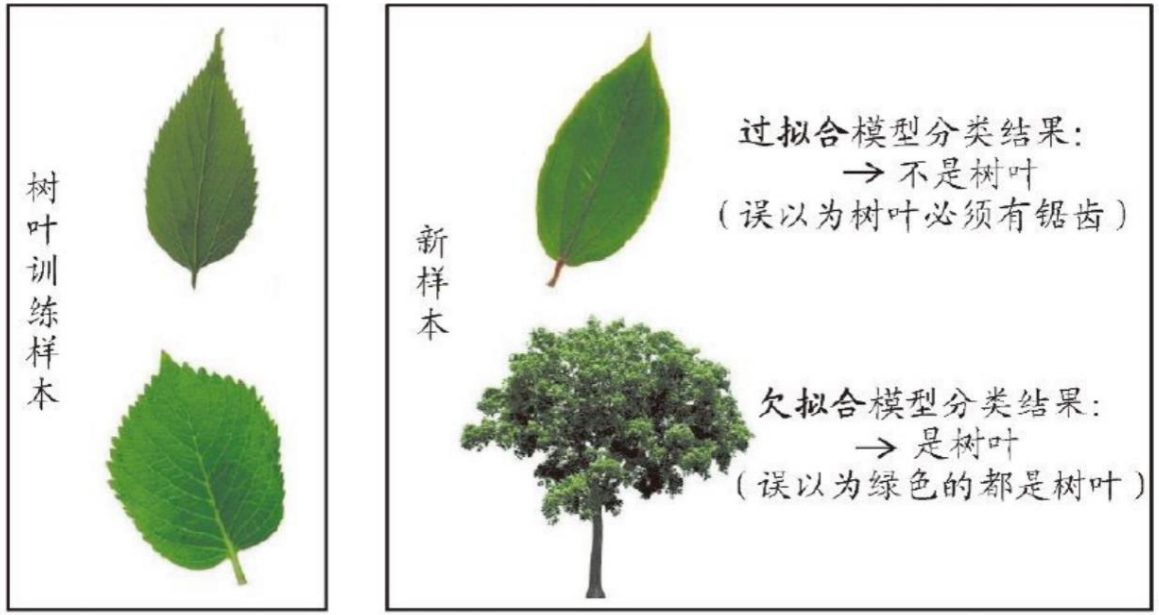
2.1 引入-为什么要进行模型评估与选择

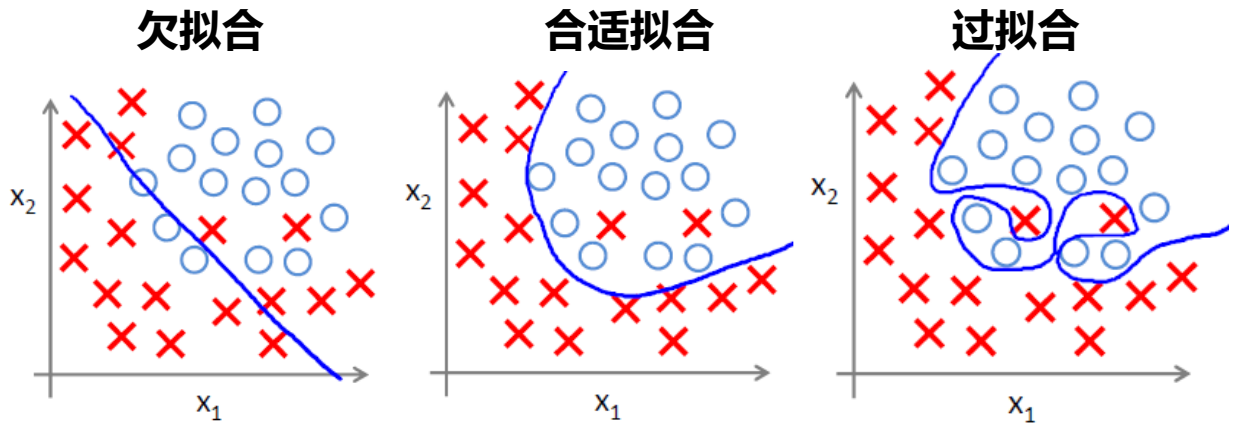
泛化误差：在“未来”样本上的误差

经验误差：在训练集上的误差



□ **经验误差越小越好?**





第2章 模型评估与选择

三个关键问题：

- 如何获得测试结果？ → 评估方法
- 如何评估性能优劣？ → 性能度量
- 如何判断实质差别？ → 比较检验

基本概念

2.2 评估方法

问题：怎样获得“测试集”？

关键：测试集应该与训练集“互斥”

常见方法

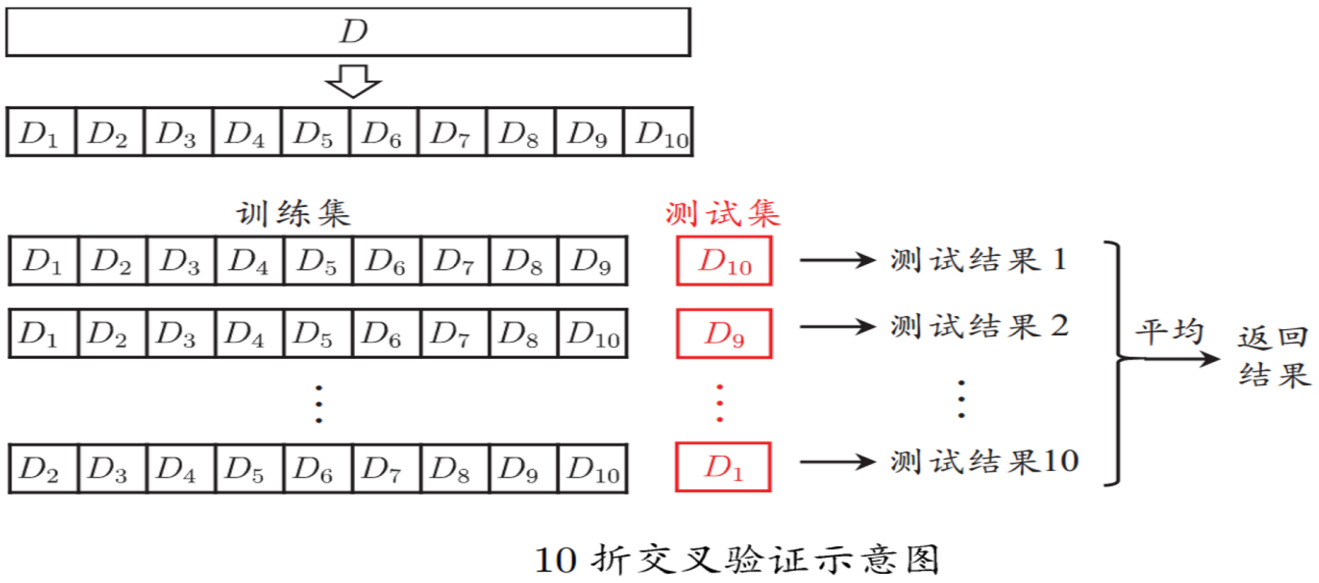
- 留出法（hold-out）
- 交叉验证法（cross validation）
- 自助法（bootstrapping）

留出法（Hold-out）

- ✓ 直接将数据集划分为两个互斥集合
- ✓ 训练/测试集划分要尽可能保持数据分布的一致性
- ✓ 一般若干次随机划分、重复实验取平均值
- ✓ 训练/测试样本比例通常为2:1~4:1

交叉验证法（Cross validation）

✓ 将数据集分层采样划分为k个大小相似的互斥子集，每 次用k-1个子集的并集作为训练集，余下的子集作为测 试集，最终返回k个测试结果的均值，k的常用取值是 10。

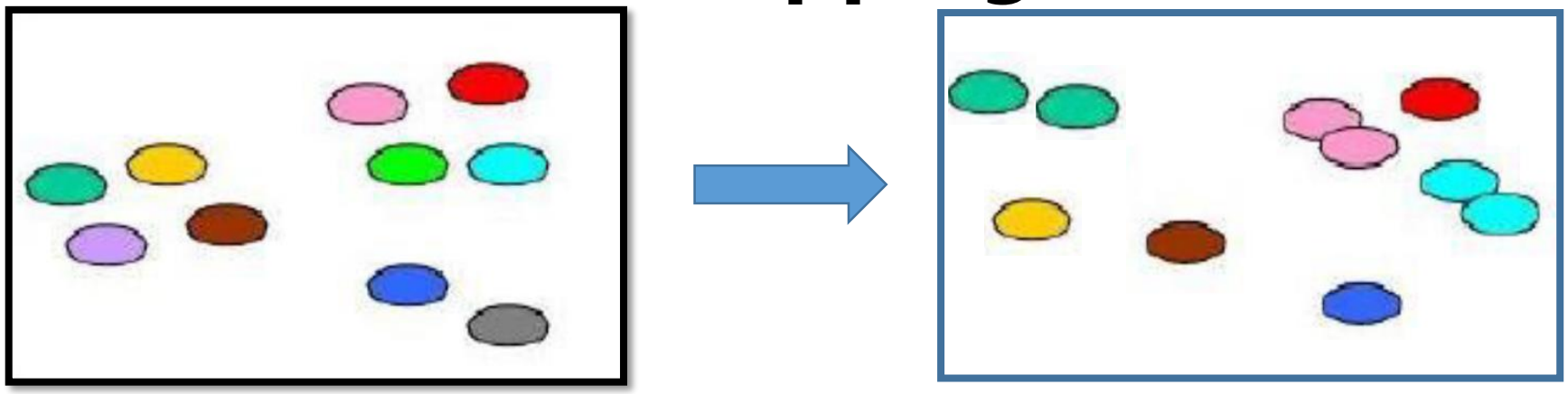


10 折交叉验证示意图

✓ 通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值

✓ 假设数据集D包含m个样本，若令k=m，则得到留一法。

自助法（Bootstrapping）



- ✓基于“自助采样”（bootstrap sampling），亦称“有放回采用”、“可重复采样”
- ✓训练集与原样本集同规模
- ✓数据分布改变，会引入估计偏差

- ✓约有36.8%的样本不出现
- ✓测试结果又称“包外估计”（out-of-bag estimate）

第2章 模型评估与选择

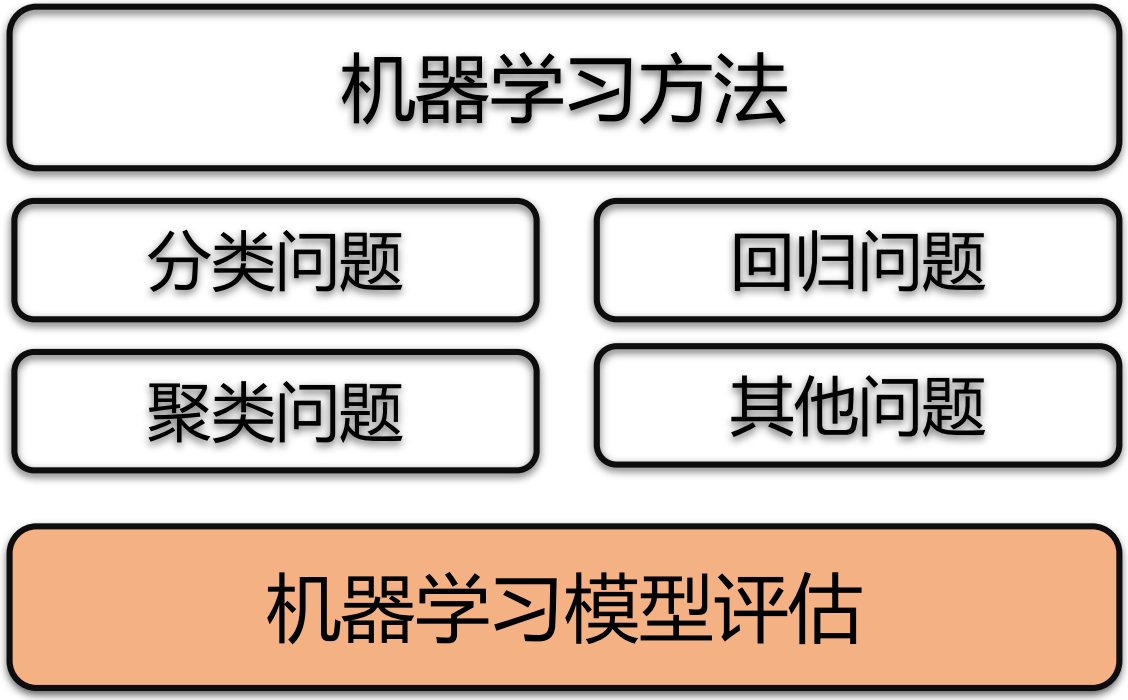
基本概念

评估方法

性能度量

比较检验

2.3 性能度量



性能评价指标-分类

准确率(Accuracy)：分类正确的样本个数占所有样本个数的比例

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

平均准确率(Average per-class accuracy)：每个类别下的准确率的算术平均

$$average\_accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$$

精确率/查准率(Precision)：分类正确的正样本个数占分类器所有的正样本个数的比例

$$Precision = \frac{TP}{TP + FP}$$

召回率/查全率(Recall)：分类正确的正样本个数占正样本个数的比例

$$Recall = \frac{TP}{TP + FN}$$

分类结果混淆矩阵

真实情况	预测结果	
	正例	负例
正例	TP（真正例）	FN（假负例）
负例	FP（假正例）	TN（真负例）

平衡点BEP：精确率与召回率相等时的取值

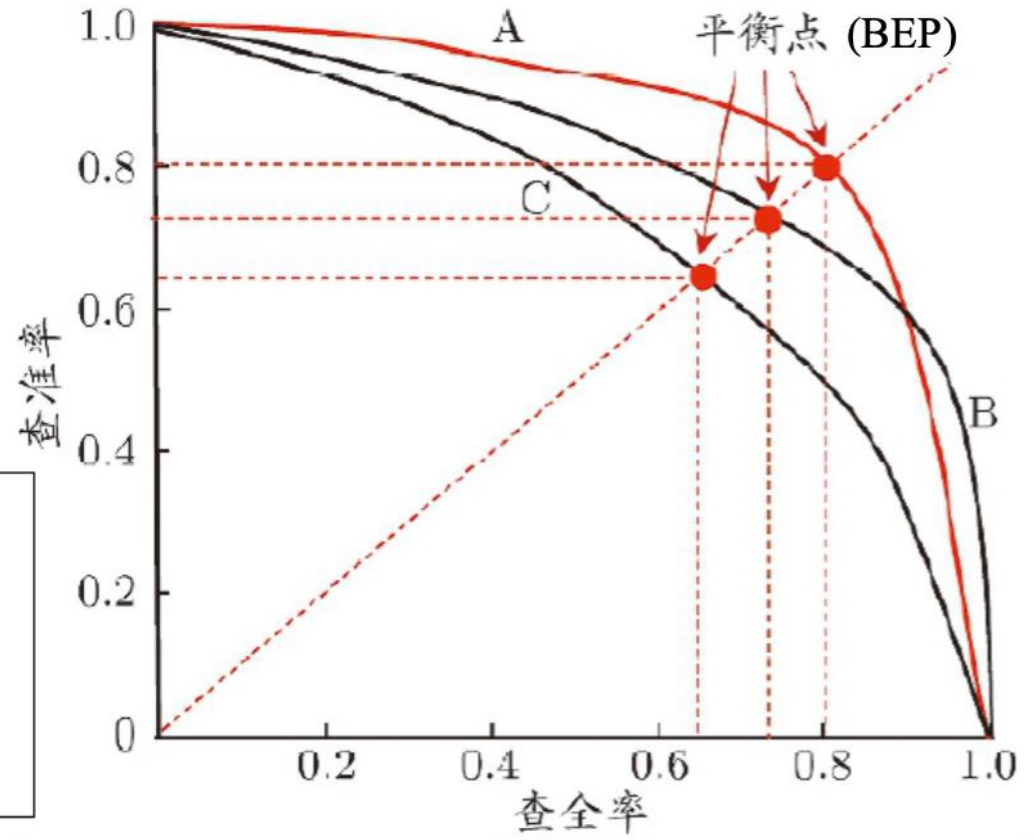
F1-Score：精确率与召回率的调和平均值，它的值更接近于Precision与Recall中较小的值

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

性能评价指标-分类

PR曲线：根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测。

- ✓ 如果一个学习器的**PR曲线**包住了另一个，则可以认为A的性能优于C
- ✓ 如果有交叉，如A、B，综合考虑PR性能，引入**平衡点(BEP)**，基于BEP比较，A优于B



PR图：
- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C
- 学习器 A ?? 学习器 B

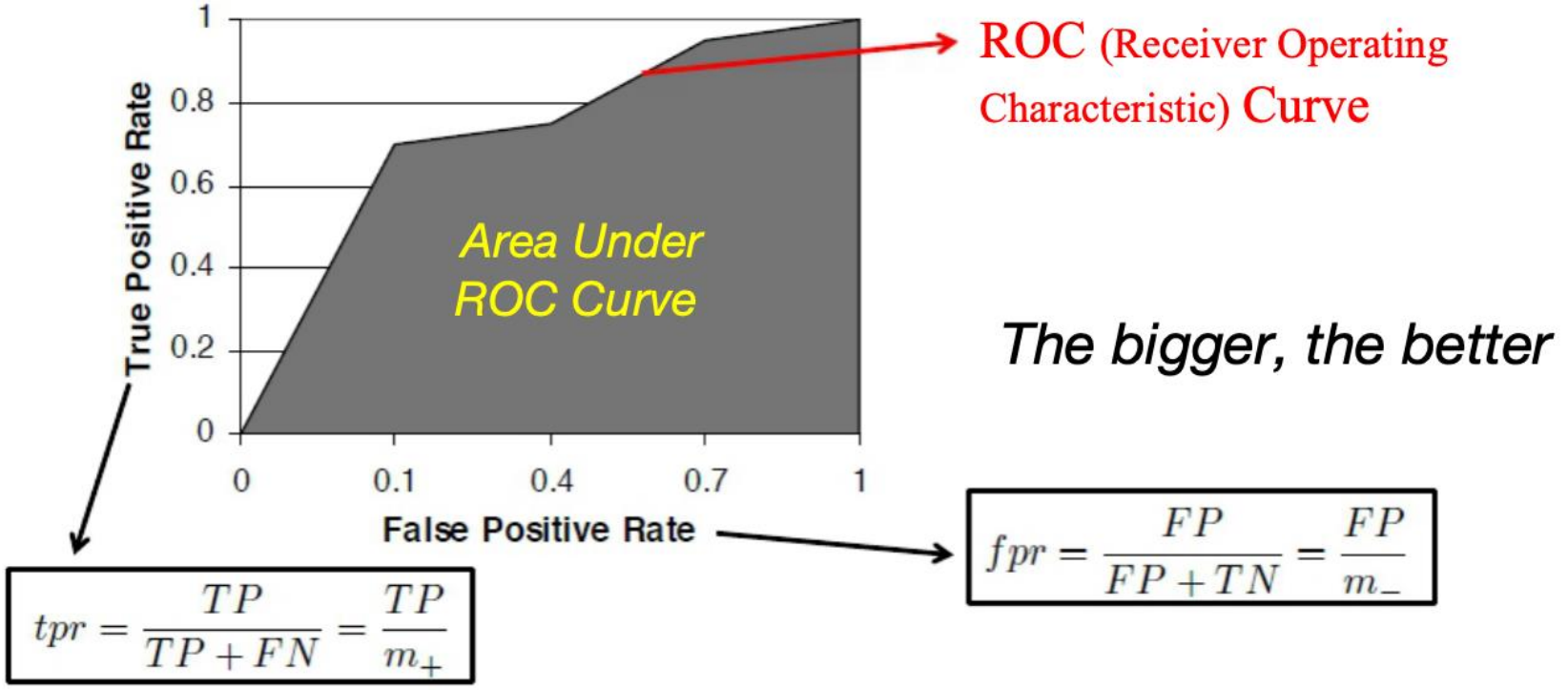
BEP：
- 学习器 A 优于 学习器 B
- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C

性能评价指标-分类

AUC(Area under the Curve(Receiver Operating Characteristic, ROC))

ROC：纵轴：真正例率TPR；横轴：假正例率FPR

AUC是ROC曲线下的面积。一般来说，如果ROC是光滑的，那么基本可以判断没有太大的overfitting，这个时候调模型可以只看AUC，面积越大一般认为模型越好。



性能评价指标-分类

宏平均&微平均

多分类问题中，若能得到**多个混淆矩阵**，例如多次训练/测试的结果，多分类的两两混淆矩阵：

宏(macro-)查准率、查全率、F1

$$\text{macro-}P = \frac{1}{n}\sum_{i=1}^{n} P_i \ ,$$

$$\text{macro-}R = \frac{1}{n}\sum_{i=1}^{n} R_i \ ,$$

$$\text{macro-}F1 = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R} \ .$$

微(micro-)查准率、查全率、F1

$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \ ,$$

$$\text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \ ,$$

$$\text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R} \ .$$

性能评价指标-回归

平均绝对误差：平均绝对误差MAE（Mean Absolute Error）又被称为l1范数损失（l1-norm loss）

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} |y_i - \hat{y}_i|$$

平均平方误差：平均平方误差MSE（Mean Squared Error）又被称为l2范数损失（l2-norm loss）：

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2$$

R Squared:是将预测值跟只使用均值的情况下相比，看能好多少。

$$R^2 = 1 - \frac{(\sum_i (\hat{y}_i - y_i)^2)/m}{(\sum_i (\hat{y}_i - y_i)^2)/m} = 1 - \frac{MSE(\hat{y}, y)}{Var(y)}$$

性能评价指标-聚类

外部指标对数据集D=$\{x_1, x_2, \ldots, x_m\}$，假定通过聚类给出的簇划分为 $C=\{C_1, C_2, \ldots, C_k\}$ 参考模型给出的簇划分为 $C^*=\{C_1^*, C_2^*, \ldots, C_k^*\}$，通过比对C和 C^* 来判定聚类结果的好坏。

Jaccard系数， FM 指数， Rand 指数，纯度purity，熵 entropy，互信息，Adjusted Rand Index (ARI)， F-measure， Probabilistic Rand Index (PRI)

内部指标对聚类数据结构上的描述，类内距离小，类间距离大较好。

DB 指数(Davies-Bouldin Index，简称DBI)：衡量同一簇中数据的紧密性，越小越好。

Dunn 指数(Dunn Index 简称DI)：衡量同一簇中数据的紧密性，越大越好。

Silouette：衡量同一簇中数据的紧密性，越大越好。

Modurity：衡量模块性，越大越好。

第2章 模型评估与选择

基本概念

评估方法

性能度量

比较检验

2.4 比较检验

- 有了实验评估方法和评估指标，看似可以对分类器的性能进行评估比较了：先使用某种实验评估方法测得分类器的某个评估指标结果，然后对这些结果进行比较。但怎么来做这个“比较”呢？是直接比较不同分类器的评估指标结果吗?
- 关于性能比较：
  - ✓ 测试性能并不等于泛化性能
  - ✓ 测试性能会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

  直接选取相应评估方法在相应度量下比大小的方法不可取！
- 假设检验为分类器的性能比较提供了重要依据，基于其结果我们可以推断出，若在测试集上观察到分类器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握有多大。

2.4 比较检验

核心：计算服从某种分布的统计量
比较统计量和临界值的大小--超过则认为有显著不同
如果是两两比较，则平均错误率小的更优

单个学习器

- 二项检验
  - 泛化错误率为ϵ的学习器，m个测试样本，测试错误率为$\hat{\epsilon}$
  - 假设“$\epsilon \leq \epsilon_0$”，置信度为$1-\alpha$，拒绝域为$\hat{\epsilon} \geq \bar{\epsilon}$，其中临界值

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

- t检验
  - k个测试错误率$\hat{\epsilon}_1, \cdots, \hat{\epsilon}_k$，可看做泛化错误率$\epsilon_0$的独立采样
  - 计算得平均测试错误率μ和样本方差σ^2，则$\tau_t = \frac{\sqrt{k}(\mu-\epsilon_0)}{\sigma}$服从自由度$k-1$的t分布
  - 假设“$\mu = \epsilon_0$”，显著度α，拒绝域为$|\tau_t| > t_{\alpha/2}$

2.4 比较检验

一个数据集多个学习器

- 成对t检验
  - 学习器A和B，k折交叉验证法得测试错误率ϵ_i^A和$\epsilon_i^B (i = 1, \dots, k)$
  - 计算得差值Δ_i及它们的均值μ和样本方差σ^2
  - 假设“$\epsilon_i^A = \epsilon_i^B$”，显著度$\alpha$，拒绝域为$\tau_t = \left|\frac{\sqrt{k}\mu}{\sigma}\right| > t_{\alpha/2}$
- McNemar检验
  - 学习器A和B，留出法得列联表（contingency table）
  - 假设“$e_{01} = e_{10}$”，显著度α，拒绝域为

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} > \chi_\alpha^2$$

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}

2.4 比较检验

多个数据集和多个学习器

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

- Friedman检验
  - 由数据集$D_1, \ldots, D_N$对算法$\mathrm{A}_i (i = 1, \ldots, k)$测试结果排序得算法平均序值$r_i$
  - 假设“各算法性能相同”，，显著度α，拒绝域为$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} > F_{k-1,(k-1)(N-1)}(\alpha/2)$，其中$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right)$

- Nemenyi后续检验
  - 若假设被拒绝，计算平均序值差别的临界值域$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$
  - 假设“两个算法性能相同”，显著度α，拒绝域为$|r_i - r_j| > CD$

2.5 偏差与方差

- 泛化误差可分解为偏差、方差和噪声之和。
  - $E(f;D) = \mathbb{E}_D[(f(\boldsymbol{x};D) - y_D)^2] = \cdots = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \varepsilon^2$

- 偏差度量了学习算法的偏离程度，$\mathbb{E}_D[(f(\boldsymbol{x};D) - \mathbb{E}_D[f(\boldsymbol{x};D)])^2]$
- 方差度量了数据扰动所造成的影响，$(\mathbb{E}_D[f(\boldsymbol{x};D)] - y)^2$
- 噪声刻画了学习问题本身的难度，$\mathbb{E}_D[(y_D - y)^2]$

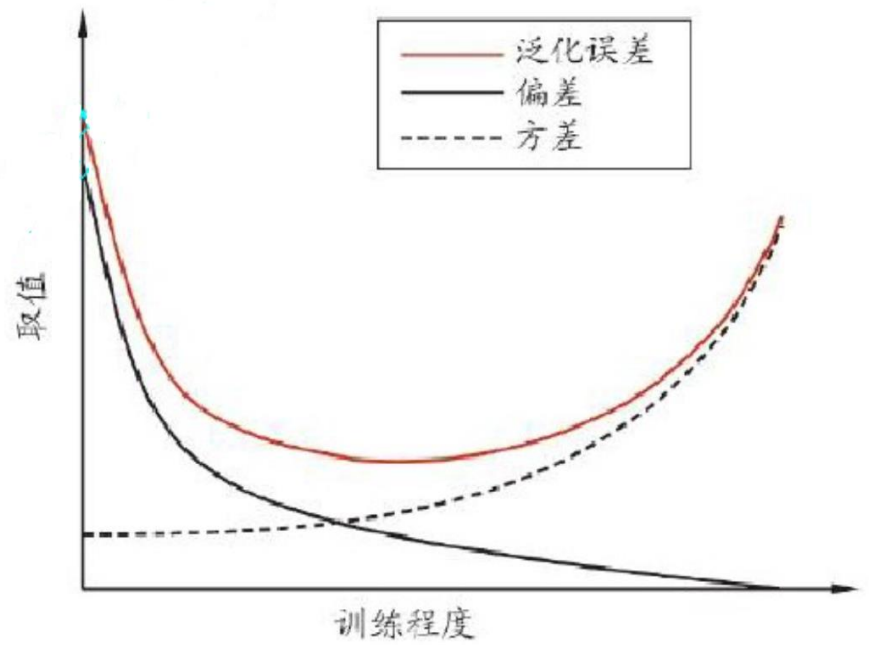


图 2.9 泛化误差与偏差、方差的关系示意图

感谢聆听

有什么问题吗?

留一法（Leave-one-out, LOO）优缺点

- ✓ 不受随机样本划分方式的影响
- ✓ 结果往往比较准确
- ✓ 当数据集比较大时，计算开销难以忍受