第2章 模型评估与选择

王博：自动化（人工智能）学院
wangbo@hdu. edu. cn

章节目录

经验误差与过拟合

错误率&误差

- 错误率: 错分样本的占比：$E = a/m$
- 误差：样本真实输出与预测输出之间的差异
  - 训练(经验)误差：训练集上
  - 测试误差：测试集
  - 泛化误差：除训练集外所有样本

由于事先并不知道新样本的特征，只能努力使经验误差最小化；

训练集上做到分类错误率为零？

经验误差与过拟合

过拟合

学习器把训练样本学习的“太好”，将训练样本本身的特点当做所有样本的一般性质，导致泛化性能下降

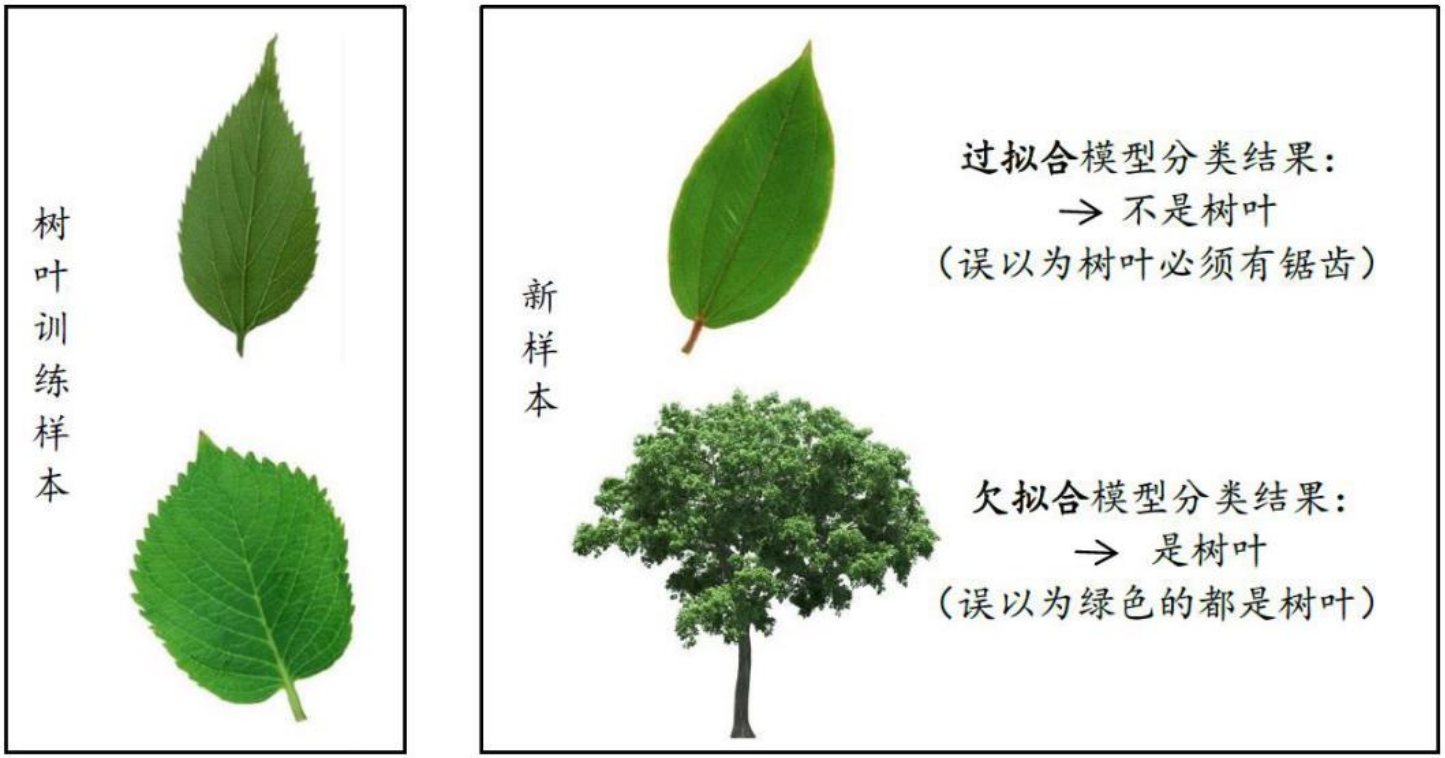
- 优化目标加正则项
- early stop

欠拟合

对训练样本的一般性质尚未学好

- 决策树:拓展分支
- 神经网络:增加训练轮数

经验误差与过拟合



过拟合、欠拟合的直观类比

章节目录

- 经验误差与过拟合
- 评估方法
- 性能度量
- 比较检验
- 偏差与方差

评估方法

对学习器的泛化性能、时间开销、存储开销、可解释性等方面的因素进行评估并做出选择

泛化误差

假设测试集是从样本真实分布中独立采样获得，将测试集上的“测试误差”作为泛化误差的近似。

测试集要和训练集中的样本尽量互斥。

问题：通常只有一个包含m个样本的数据集，既要训练又要测试，该如何划分？

评估方法-留出法

通常将包含个m 样本的数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
拆分成训练集 S 和测试集T

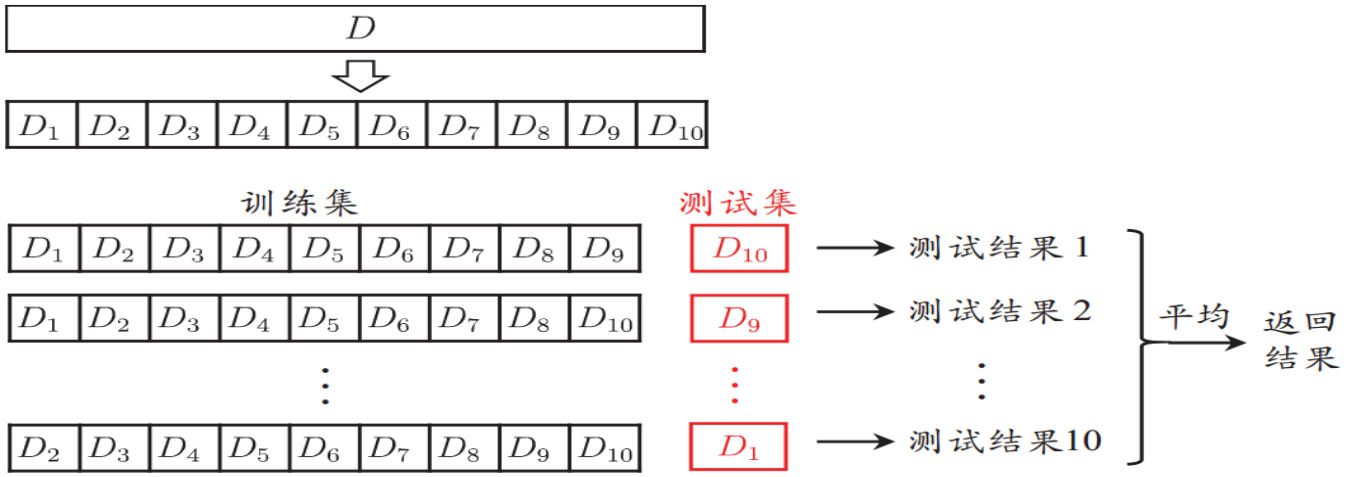
留出法：

- 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

评估方法-交叉验证法

交叉验证法：

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最常用的取值是10.



10 折交叉验证示意图

评估方法-留一法

与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证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，例如常见的“10次10折交叉验证”

假设数据集D包含m个样本，若令 $k = m$ ，则得到留一法：

- 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

评估方法-自助法

自助法

以自助采样法为基础，对数据集 D 有放回采样 m 次得到训练集 D'，$D \setminus D'$ 用做测试集。

- 实际模型与预期模型都使用 m 个训练样本
- 约有1/3的样本没在训练集中出现
- 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处
- 自助法在数据集较小、难以有效划分训练/测试集时很有用；由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用。

章节目录

- 经验误差与过拟合
- 评估方法
- **性能度量**
- 比较检验
- 偏差与方差

性能度量

性能度量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

在预测任务中，给定样例集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
评估学习器的性能 f 也即把预测结果 $f(\boldsymbol{x})$ 和真实标记比较.

回归任务最常用的性能度量是“均方误差”：

$$E(f; D) = \frac{1}{m} \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - y_i)^2$$

性能度量

对于分类任务，错误率和精度是最常用的两种性能度量：

- 错误率：分错样本占样本总数的比例
- 精度：分对样本占样本总数的比率

分类错误率

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

精度

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

性能度量

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比率，此时查准率和查全率比错误率和精度更适合。

准确率(Accuracy)、查准率(精确率Precision)、查全率(召回率Recall)

统计真实标记和预测结果的组合可以得到“混淆矩阵”

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
精确率

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

查全率
召回率

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

准确率：(TP + TN) / (ALL)

性能度量

举例

一个班有50人，在某场考试中有40人及格，10人不及格。现在需要根据一些特征预测出所有及格的学生。

某一模型执行下来，给出了39人，其中37人确实及格了，剩下2人实际上不及格。

求准确率、查准率(Precision)、召回率？

性能度量

举例

一个班有50人，在某场考试中有40人及格，10人不及格。
现在需要根据一些特征预测出所有及格的学生。

某一模型执行下来，给出了39人，其中37人确实及格了，剩下2人实际上不及格。

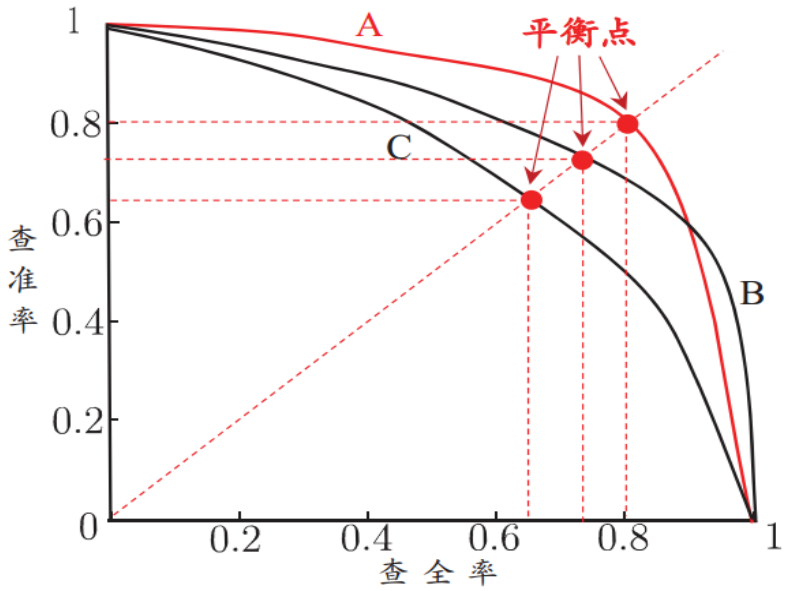
求准确率、查准率(Precision)、召回率？

Accuracy = 45 / 50
Precision = 37 / 39
Recall = 37 / 40

性能度量- P-R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，则可以得到查准率-查全率曲线，简称“P-R曲线”

平衡点是曲线上“查准率=查全率”时的取值，可用来用于度量P-R曲线
有交叉的分类器性能高低，越高可认为学习器更好



P-R曲线与平衡点示意图

p是positive样本（正例）,n当然就是negative（负例），score是分类器对于该样本属于正例的可能性的打分。因为一般模型输出的不是0,1的标注，而是小数，相当于置信度， 然后设置一个从高到低的阈值y，大于等于阈值y的被标注为正例，小于阈值y的被标注为负例。每一种阈值下都会求得查准率和查全率

性能度量- F1度量

比P-R曲线平衡点更用常用的是F1度量：

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

比F1更一般的形式 F_β ，

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

$\beta = 1$： 标准F1
$\beta > 1$：偏重查全率
$\beta < 1$：偏重查准率

$F1$ 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.

性能度量- F1度量

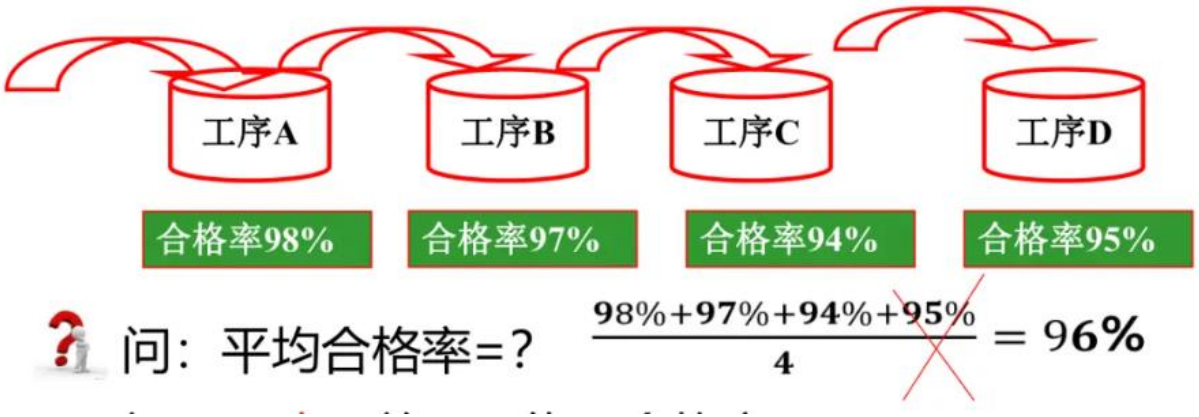
调和平均

$$H=\frac{1}{\frac{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}{n}}=\left[\frac{\left(x_1^{-1}+x_2^{-1}+\cdots+x_n^{-1}\right)}{n}\right]^{-1}$$

几何平均

$$G=\sqrt[n]{X_1\cdot X_2\cdot\cdots\cdot X_n}=\sqrt[n]{\prod_i X_i}$$

适用对象：计算平均比率或平均发展速度



性能度量- F1度量

	P	R	平均	F1
算法1	0.5	0.4	0.45	0.444
算法2	0.7	0.1	0.4	0.175
算法3	0.02	1	0.51	0.0392

如果两者的差距非常大?

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

性能度量-ROC曲线

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)

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

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

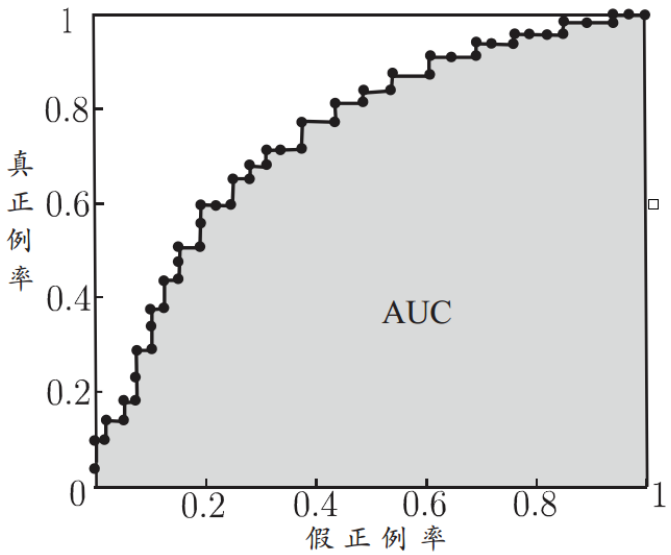
TPR真正例率为纵轴，FPR假正例率为横轴

ROC图的绘制：给定 m^+ 个正例和 m^- 个负例，根据学习器预测结果对样例进行排序，将分类阈值设为每个样例的预测值，当前标记点坐标为 (x, y)，当前若为真正例，则对应标记点的坐标为 $(x, y + \frac{1}{m^+})$；当前若为假正例，则对应标记点的坐标为 $(x + \frac{1}{m^-}, y)$，然后用线段连接相邻点.

二分类：刚开始阈值为1，坐标为(0,0)

性能度量-AUC

若某个学习器的ROC曲线被另一个学习器的曲线“包住”，则后者性能优于前者；否则如果曲线交叉，可以根据ROC曲线下面积大小进行比较，也即AUC值.



基于有限样例绘制的 ROC 曲线与 AUC

假设ROC曲线由$\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$的点按序连接而形成$(x_1 = 0, x_m = 1)$，则：AUC可估算为：

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

AUC衡量了样本预测的排序质量。

性能度量-ROC曲线

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

好瓜预测阈值就应该高，坏瓜预测阈值就应该低，才是好的学习器

情况1	
真实情况	预测
好瓜	0.9
好瓜	0.8
坏瓜	0.4
坏瓜	0.3

AUC排序能力解释：AUC等于随机挑选一个正样本 (P) 和负样本 (N) 时，分类器将正样本排在负样本前面的概率。

AUC=0.6 时，其含义可大致理解为：给定一个正样本和一个负样本，在 60% 的情况下，模型对正样本的打分高于对负样本的打分。可见在该解释下，关心的只有正负样本间的相对分数高低，而具体的分值则无关紧要

章节目录

- 经验误差与过拟合
- 评估方法
- 性能度量
- **比较检验**
- 偏差与方差

比较检验

关于性能比较：

- 测试性能并不等于泛化性能
- 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

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！

若在测试集上观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B？？？

假设检验

假设检验：检验统计假设的一种方法，假设检验是先对总体参数提出一个假设值，然后利用样本信息判断这一假设是否成立。

以错误率 ϵ 为性能度量，“假设”为对错误率分布的某种猜想或判断

记泛化错误率为 ϵ，测试错误率为 $\hat{\epsilon}$，假定测试样本从样本总体分布中独立采样而来，利用测试错误率估推出泛化错误率的分布。

泛化错误率为 ϵ 的学习器被测得测试错误率为 $\hat{\epsilon}$ 的概率为（假定测试样本从m个样本总体中独立采样而来）：

$$P(\hat{\varepsilon};\varepsilon)=\begin{pmatrix} m \\ \hat{\varepsilon}\times m \end{pmatrix}\varepsilon^{\hat{\varepsilon}\times m}(1-\varepsilon)^{m-\hat{\varepsilon}\times m}$$

给定测试错误率的情况下，什么时候概率取最大值？

二项检验

使用“二项检验”对 $\epsilon \leq \epsilon_0$ 进行假设检验。以下式子啥意思？？？

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

假设检验的概念是这样的：假设 $\epsilon \leq \epsilon_0$ 成立，若 $\hat{\epsilon} \leq \bar{\epsilon}$ 的概率不小于 $1-\alpha$，则接受假设 $\epsilon \leq \epsilon_0$，即若 $P(\hat{\epsilon} \leq \bar{\epsilon} \mid \epsilon \leq \epsilon_0) \geq 1-\alpha$ 成立，则认为假设 $\epsilon \leq \epsilon_0$ 猜对了！

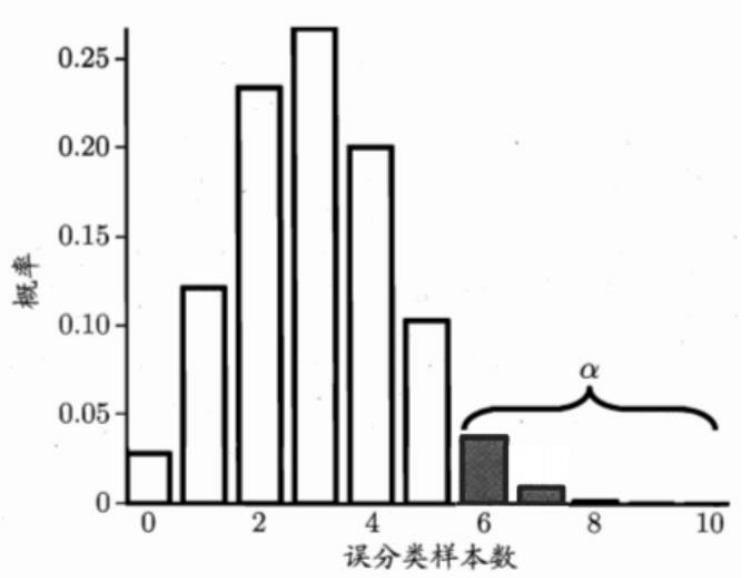


图 2.6 二项分布示意图($m = 10, \epsilon = 0.3$)

需要检验假设

H_0：$\epsilon \leq \epsilon_0$（即泛化错误率不大于ϵ_0）

H_1：$\epsilon > \epsilon_0$（即泛化错误率大于ϵ_0）

这是一个右边检验问题，其拒绝域为 $\hat{\epsilon} > \bar{\epsilon}$

$$P(\hat{\epsilon} > \bar{\epsilon} \mid \epsilon \leq \epsilon_0) < \alpha$$

正态分布、卡方分布、t分布

记总体均值为μ，总体方差为σ^2（未知），样本均值$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，样本标准差$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$，有：

$$X_i \sim N(\mu,\sigma^2) \rightarrow \bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \sim N(\mu,\frac{\sigma^2}{n})$$

$$\rightarrow \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$$

卡方变量的一个重要定理：$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

X服从N(0,1)，Y服从自由度为n－1的卡方分布，$X/\sqrt{Y/(n-1)}$服从自由度为n－1的t分布

$$\frac{\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{n-1}}}=\frac{\sqrt{n}(\bar{X}-\mu)}{s} \sim t(n-1)$$

t检验

1. 对于**要检验样本均值是否等于总体均值的双侧检验**，若根据样本数据算出来的t统计量的绝对值 $\left|\frac{\sqrt{n}(\bar{X}-\mu)}{s}\right| > t_{\frac{\alpha}{2},n-1}$，则拒绝原假设，认为样本均值与总体均值不等，否则不拒绝原假设。
2. 对于**要检验样本均值是否比总体均值大的单侧检验**，若根据样本数据算出来的t统计量 $\frac{\sqrt{n}(\bar{X}-\mu)}{s} < t_{\alpha,n-1}$，则拒绝原假设，认为样本均值不大于总体均值，否则不拒绝原假设。
3. 对于**要检验样本均值是否比总体均值小的单侧检验**，若根据样本数据算出来的t统计量 $\frac{\sqrt{n}(\bar{X}-\mu)}{s} > t_{\alpha,n-1}$，则拒绝原假设，认为样本均值不小于总体均值，否则不拒绝原假设。

举例

从某厂生产的零件中随机抽取若干件，检验其某种规格的均值是否与要求的规格相等（双侧检验）

检验某一线城市全体高三学生视力水平是否比全国全体高三学生视力水平低（单侧检验）

$$P(|Z| \geq z_{\alpha/2}) = \alpha,$$

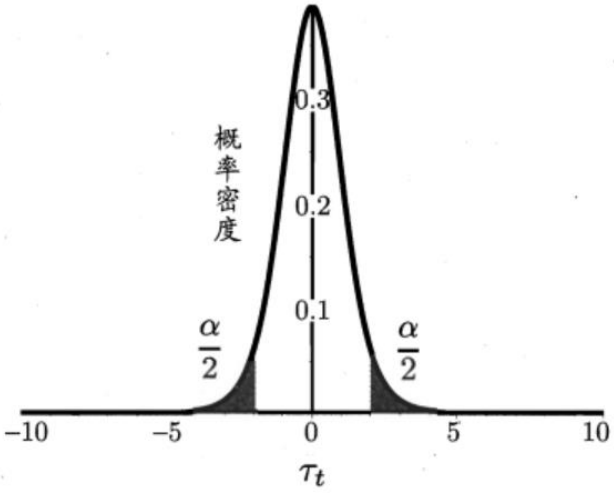


图 2.7 t 分布示意图($k=10$)

t检验

对应的，面对多次重复留出法或者交叉验证法进行多次训练/测试时可使用“t检验”。

假定得到了k个测试错误率，$\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_k$平均测试错误率及方差 μ, σ^2

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

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

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

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

服从自由度为 $k-1$ 的 t 分布

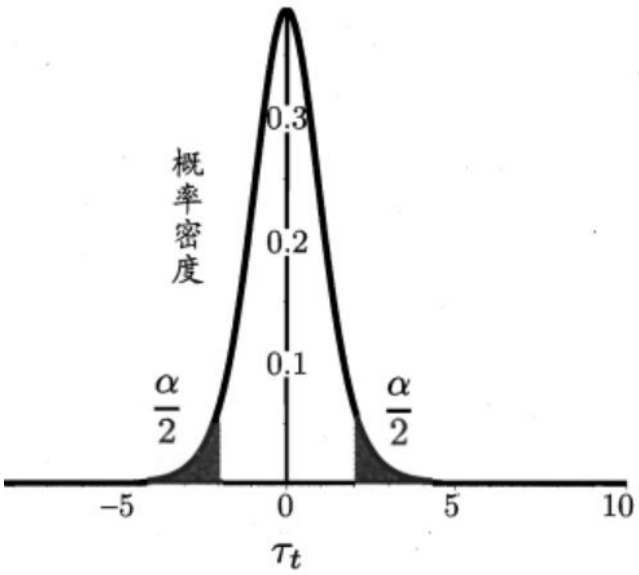


图 2.7 t 分布示意图($k=10$)

t检验

假设检验H0：样本均值与总体均值相等

$$\mu = \epsilon_0$$

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

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

服从自由度为 $k-1$ 的 t 分布

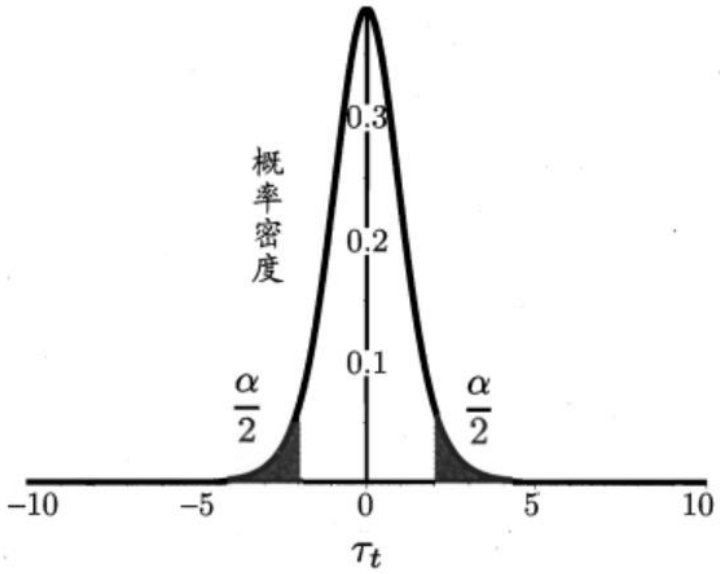


图 2.7 t 分布示意图($k=10$)

交叉验证t检验

现实任务中，更多时候需要对不同学习器的性能进行比较

对两个学习器A和B，若k折交叉验证得到的测试错误率分别为 $\epsilon_1^A, \ldots, \epsilon_k^A$ 和 $\epsilon_1^B, \ldots, \epsilon_k^B$ 可用k折交叉验证“成对t检验”进行比较检验。

章节目录

偏差与方差

通过实验可以估计学习算法的泛化性能，而“偏差-方差分解”可以用来帮助解释泛化性能。偏差-方差分解试图对学习算法期望的泛华错误率进行拆解。

对测试样本x,令y_D为x在数据集中的标记,y为x的真实标记,$f(\boldsymbol{x};D)$为训练集D上学得模型f在x上的预测输出。以回归任务为例：学习算法的期望预期为：

$$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x};D)]$$

使用样本数目相同的不同训练集产生的方差为

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

噪声为

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

偏差与方差

期望输出与真实标记的差别称为偏差，即 $bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$

为便与讨论，假定噪声期望为0，也即 $\mathbb{E}_D[y_D - y] = 0$ 对泛化误差分解

$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&\quad + \mathbb{E}_D\left[2\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)\left(\bar{f}(\boldsymbol{x}) - y_D\right)\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right] + \mathbb{E}_D\left[\left(\bar{f}(\boldsymbol{x}) - y_D\right)^2\right]
\end{aligned}
$$

偏差与方差

$$
\begin{aligned}
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right)-\bar{f}\left(\boldsymbol{x}\right)\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right)-y+y-y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right)-\bar{f}\left(\boldsymbol{x}\right)\right)^2\right]+\mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right)-y\right)^2\right]+\mathbb{E}_D\left[\left(y-y_D\right)^2\right] \\
&\quad +2\mathbb{E}_D\left[\left(\bar{f}\left(\boldsymbol{x}\right)-y\right)\left(y-y_D\right)\right]
\end{aligned}
$$

又由假设中噪声期望为0，可得

$$
E(f;D)=\mathbb{E}_D\left[\left(f\left(\boldsymbol{x};D\right)-\bar{f}\left(\boldsymbol{x}\right)\right)^2\right]+\left(\bar{f}\left(\boldsymbol{x}\right)-y\right)^2+\mathbb{E}_D\left[\left(y_D-y\right)^2\right]
$$

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

即泛化误差可分解为偏差、方差与噪声之和。

偏差与方差

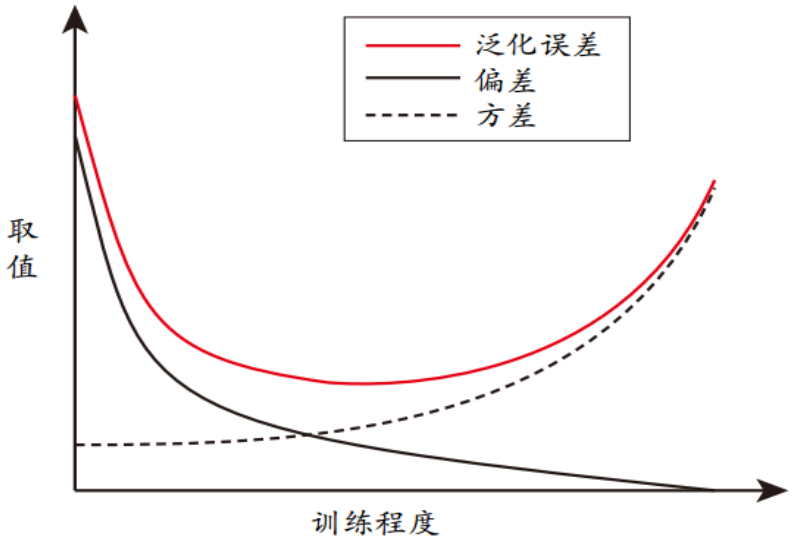
- 偏差度量了学习算法期望预测与真实结果的偏离程度；即刻画了学习算法本身的拟合能力；
- 方差度量了同样大小训练集的变动所导致的学习性能的变化；即刻画了数据集变化（数据扰动）所造成的影响；
- 噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界；即刻画了学习问题本身的难度。

泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度所共同决定的。给定学习任务为了取得好的泛化性能，需要使偏差小(充分拟合数据)而且方差较小(减少数据扰动产生的影响)。

偏差与方差

一般来说，偏差与方差是有冲突的，称为偏差-方差窘境。如右图所示，假如我们能控制算法的训练程度：

- 在训练不足时，学习器拟合能力不强，训练数据的扰动不足以使学习器的拟合能力产生显著变化，此时偏差主导泛化错误率；
- 随着训练程度加深，学习器拟合能力逐渐增强，方差逐渐主导泛化错误率；
- 训练充足后，学习器的拟合能力非常强，训练数据的轻微扰动都会导致学习器的显著变化，若训练数据自身非全局特性被学到则会发生过拟合。



泛化误差与偏差、方差的关系示意图