清華大學
Tsinghua University

# 大数据机器学习

## 第三讲：模型性能评估

袁春
清华大学深圳研究生院
2017/6

Tsinghua University

## 提纲

- 训练集和测试集的产生
  - 留出法
  - 交叉验证法
  - 自助法。
- 性能度量
  - PR曲线
  - ROC
  - 代价曲线
- 假设检验
  - 二项检验
  - T检验
  - 交叉t检验
- 偏差-方差分解

# 模型评估方法

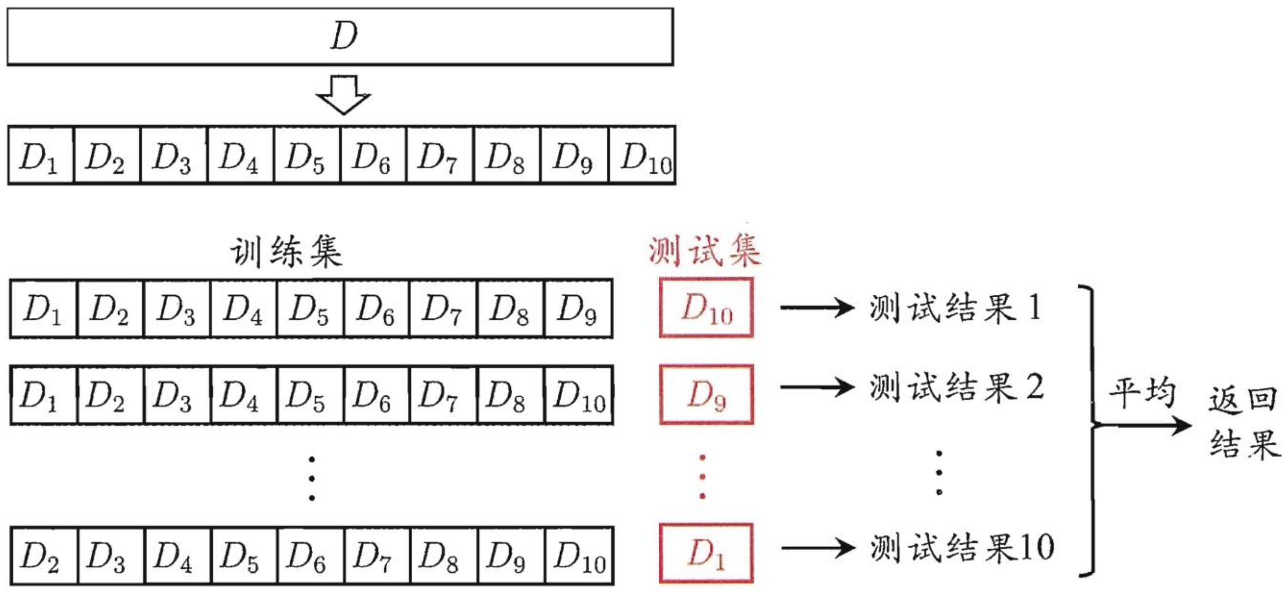
- 泛化误差评估：
  - 训练集 training set： 用于训练模型
  - 验证集 validation set： 用于模型选择
  - 测试集 test set： 用于模型泛化误差的近似

- 训练集和测试集的产生
  - 留出法
  - 交叉验证法
  - 自助法

## 留出法 Hold-out

- D=S∪T
- S∩T=∅
- 注意点:
  - 训练/测试集的划分尽可能保持数据分布的一致性，避免引入额外偏差;
  - 存在多种划分方式对初始数据集进行分割，采用若干次随机划分，重复实验;
- 存在问题:
  - S大，T小；S小，T大，都会带来负面影响;

## 交叉验证法 cross validation

- D →k个大小相等的互斥子集
- $D = D_1 \cup D_2 \cup \ldots \cup D_k,\ D_i \cap D_j = \varnothing\ (i \neq j)$
- K-1个子集并集为训练集，1个测试集
- 

## 自助法 boostrapping

- 自助采样法:

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

- 测试集：D \ D′
- 优点
  - 适用于数据集较小，难以划分；
  - 从数据集产生不同的训练集，适用于集成学习方法；
- 缺点
  - 产生的训练集改变了初始数据集的分布，会引入估计偏差。

## 性能度量

- 不同任务，性能度量不同；
- 回归任务-均方误差：

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

- 更一般：

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

## 性能度量

- 错误率和精度-分类任务
- 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

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

- 更一般：

$$E(f;\mathcal{D})=\int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}\left(f\left(\boldsymbol{x}\right)\neq y\right)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\begin{aligned}\operatorname{acc}(f;\mathcal{D}) &= \int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}\left(f\left(\boldsymbol{x}\right)=y\right)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= 1-E(f;\mathcal{D})\ .\end{aligned}$$

## 性能度量

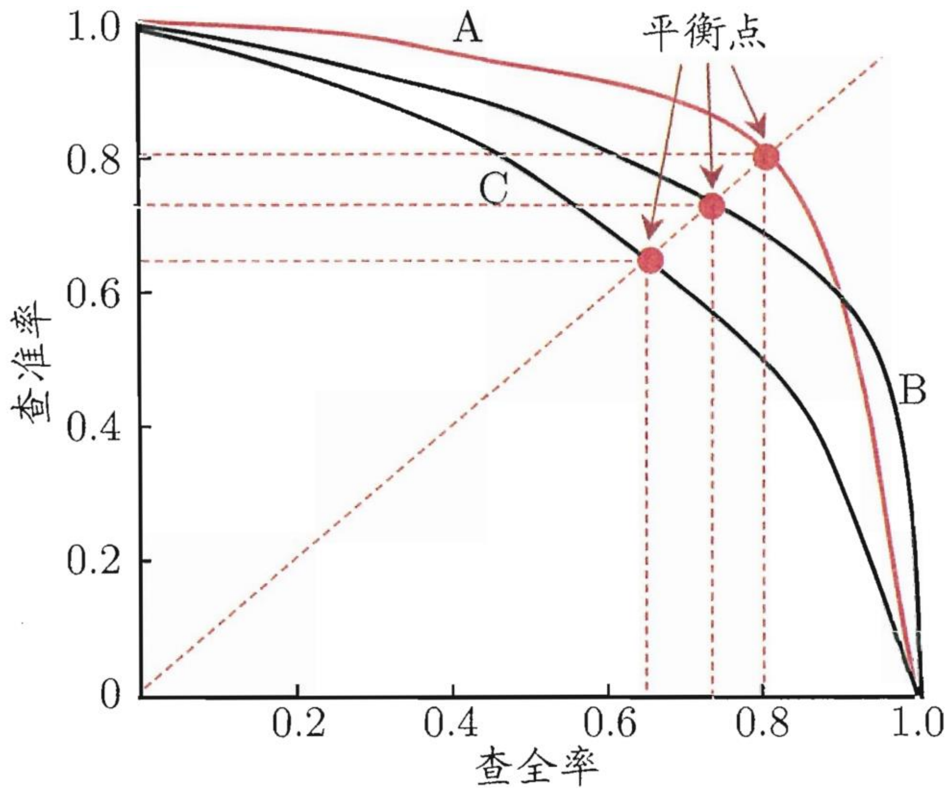
- 查准率precision 、 查全率recall与FI
- 二分类-混淆矩阵：

| 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： $$R = \frac{TP}{TP + FN}$$

## 性能度量

- 查准率precision 、 查全率recall与FI
- P-R曲线




- 平衡点BEP
  - 查准率=查全率

## 性能度量

- F1度量

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

- $F_\beta$ 度量

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

# 性能度量

- 多个二分类混淆矩阵：
  - 多次训练/测试
  - 多个数据集上训练/测试
  - 执行多分类任务
- 宏查准率(macro-P)/宏查全率(macro-R)/宏F1

$$\text{macro-}P = \frac{1}{n}\sum_{i=1}^{n} P_i \qquad \text{macro-}R = \frac{1}{n}\sum_{i=1}^{n} R_i \qquad \text{macro-}F1 = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R}$$

- 微查准率(micro-P)/微查全率”(micro-R)和“微FI

$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \qquad \text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \qquad \text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R}$$

## 性能度量

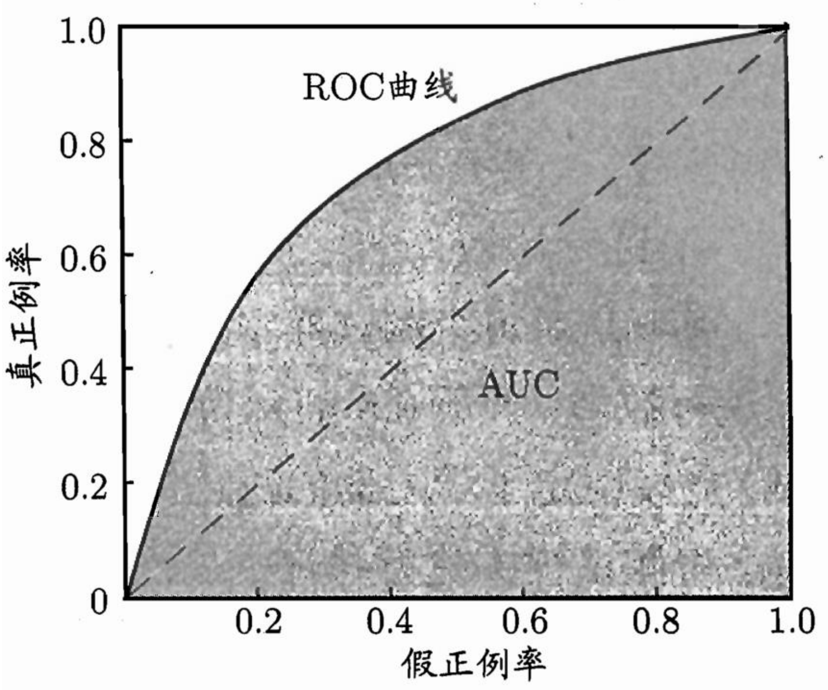
- ROC(Receiver Operating Characteristic)
- AUC(Area Under ROC Curve)

- 纵轴：“真正例率 ” (True Positive Rate, 简称 TPR)
- 横轴：“假正例率” (False Positive Rate, 简称 FPR),

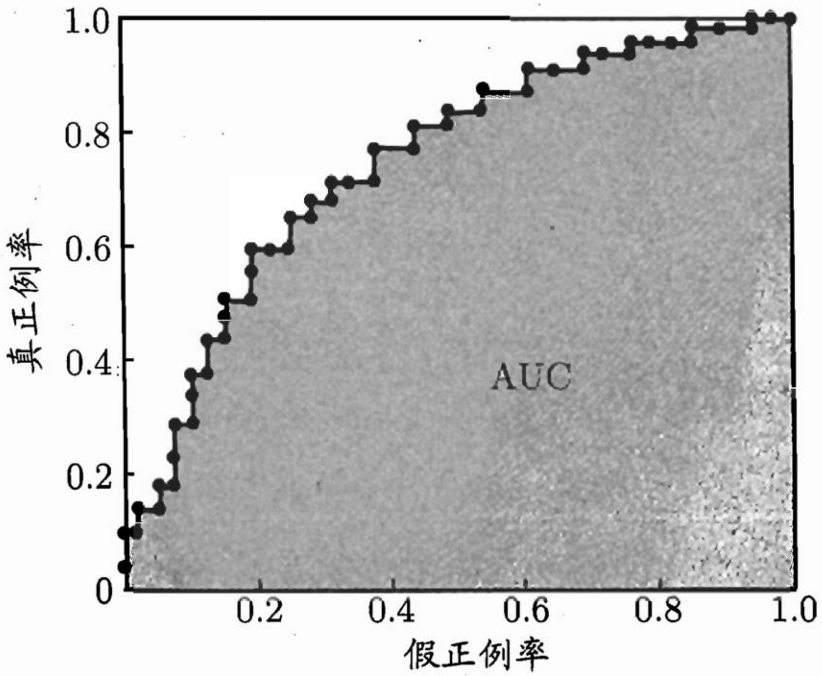
$$\mathrm{TPR} = \frac{TP}{TP + FN} \qquad \mathrm{FPR} = \frac{FP}{TN + FP}$$

# 性能度量

- ROC(Receiver Operating Characteristic)
- AUC(Area Under ROC Curve)




(a) ROC 曲线与 AUC




(b) 基于有限样例绘制的 ROC 曲线与 AUC

## 性能度量

- 代价敏感错误率与代价曲线
  - 应用背景：不同类型的错误所造成的后果不同；
- 二分类任务：代价矩阵（cost matrix）

| 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 |

对应代价敏感错误率

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

## 性能度量

- 代价曲线cost curve：非均等代价下ROC曲线不适用；
- 横轴：正例概率代价： P为样例为正例的概率。

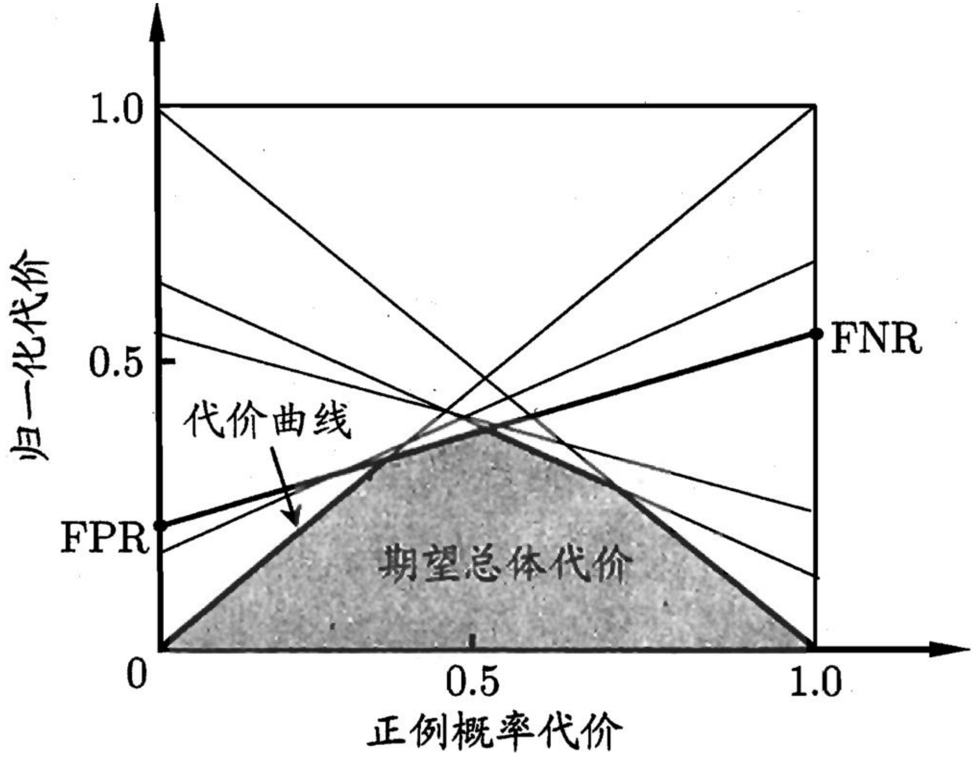
$$P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}}$$

- 纵轴：纵轴是取值为 [0,1] 的归一化代价

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

# 性能度量

- 代价曲线cost curve：非均等代价下ROC曲线不适用；
- 



# 性能度量

- **比较检验**
- 问题提出： 能否直接用上述评估方法获得的性能度量“比大小”？
  - 答案： 不能，
- 原因：
  - 希望比较泛化性能， 实验评估的是测试集性能；
  - 测试集性能和测试集的选择有关， 测试样例不同， 结果不同；
  - 机器学习算法本身有一定的随机性， 相同的参数， 相同的数据集， 结果也会不同。
- 方案： 统计假设检验(hypothesis test)
  - 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， 以及这个结论的把握有多大.

# 性能度量

- **假设检验**
- 对单个学习器泛化性能的假设进行检验
  - “二项检验” (binomial test)
  - “*t* 检验”(t-test)

- 对不同学习器的性能进行比较，
  - “成对 t 检验” (paired t-tests)

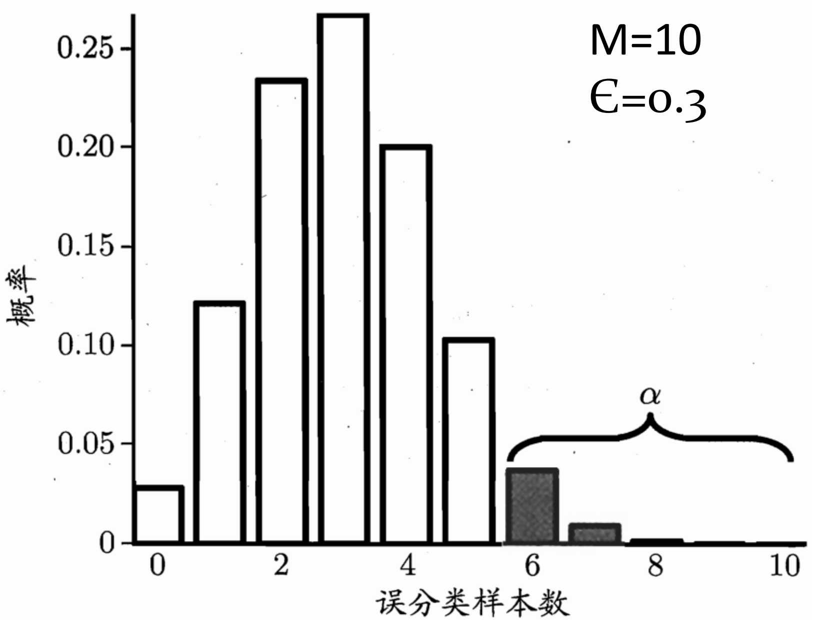
# 性能度量

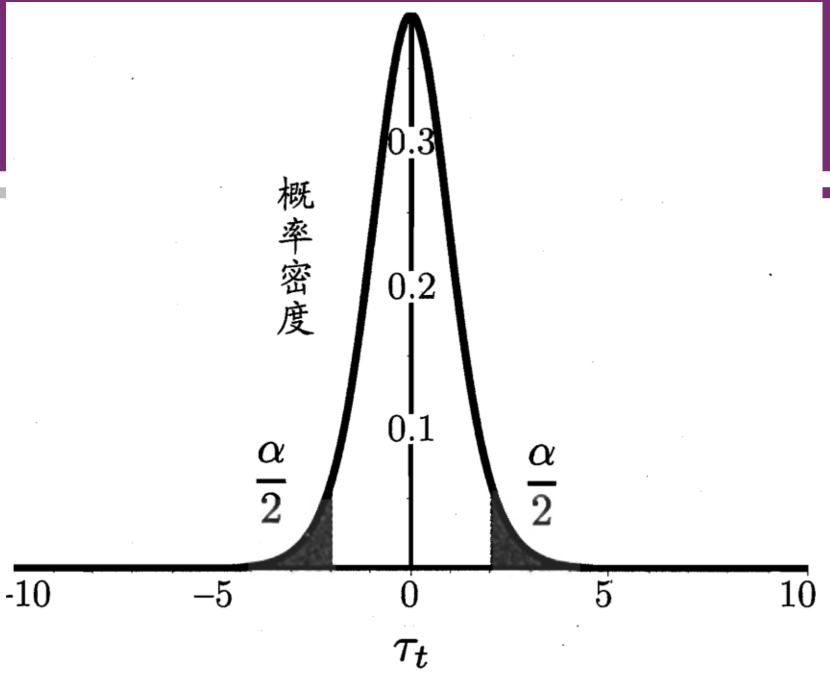
- **二项检验**
  - 假设检验：“假设”是对学习器泛化错误率分布的某种判断或猜想，如ϵ
  - 现实任务中我们只能获知测试错误率
  - 那么：泛化错误率为ϵ 的学习器将其中$\hat{\epsilon}$m'个样本误分类的概率：

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

  - 使用二项检验对泛化误差ϵ ≤0.3的假设进行检验；
  - 1-a 的概率内所能观测到的最大错误率：

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






## 性能度量

- t检验
  - 多次重复训练/测试，得到多个测试错误率；
  - K个测试错误率， $\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_k$;
  - 均值 $\mu = \frac{1}{k}\sum_{i=1}^{k}\hat{\epsilon}_i$ 方差 $\sigma^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{\epsilon}_i - \mu)^2$
  - 考虑到k个测试错误率可看作泛化错误率 $\epsilon_0$ 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
  - 对假设$\mu = \epsilon_o$ 和显著度a，可计算当测试错误率均值为$\epsilon_o$，在1-a 概率内能观测到的最大错误率，即临界值，如果$|\mu - \epsilon_o|$位于临界值内，则假设成立

## 性能度量

- **交叉验证t检验**
- 学习器A,B，得到：

$$\epsilon_1^A, \epsilon_2^A, \ldots, \epsilon_k^A \text{ 和 } \epsilon_1^B, \epsilon_2^B, \ldots, \epsilon_k^B$$

- 成对t检验：假设 $\epsilon_i^A = \epsilon_i^B$
- 计算：$\Delta_i = \epsilon_i^A - \epsilon_i^B$
- 计算均值和方差
- 在显著度a下，若 $\tau_t = \left| \frac{\sqrt{k}\mu}{\sigma} \right|$ 小于临界值，则假设不能被拒绝。

清华大学 Tsinghua University

# 偏差与方差

- **偏差–方差分解**
- 对测试样本x； 令 ***yD*** 为 ***x*** 在数据集中的标记； ***y*** 为 ***x*** 的真实标记，
- ***f***(***X*** ;***D***) 为训练集 D 上学得模型f在x上的预测输出
- 回归方法的期望预测： $\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$

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

- 噪声为： $\varepsilon^2 = \mathbb{E}_D\left[(y_D - y)^2\right]$
- 期望输出与真实标记的差别称为偏差

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

# 偏差与方差

- **偏差-方差分解**
- 假定噪声期望为0，对算法的期望泛化误差进行分解：

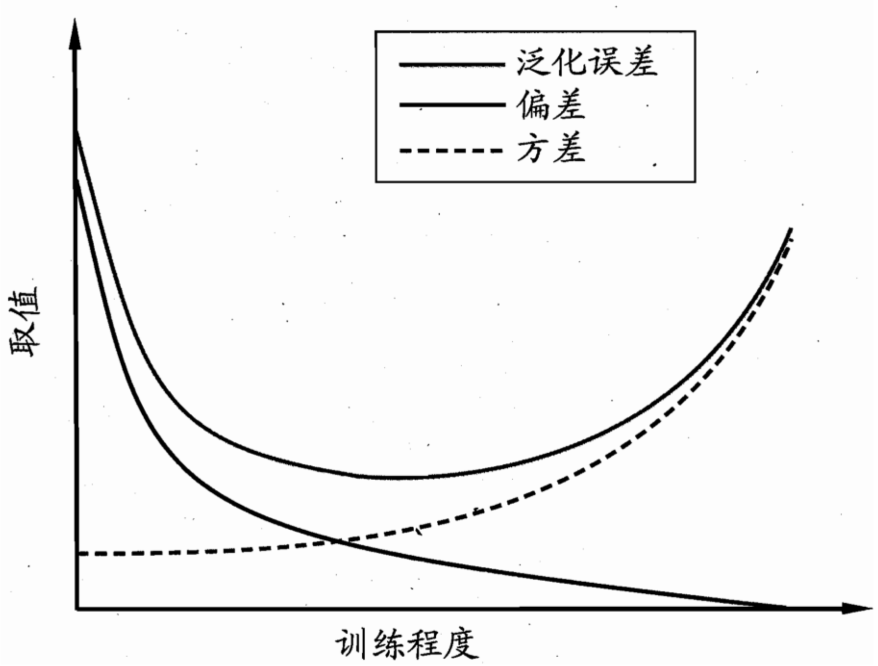
$$
\begin{aligned}
E(f;D) &= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - y_D\right)^2\right] \\
&= \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right]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

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

- 泛化误差可分解为偏差、 方差与噪声之和.

## 偏差与方差

- **偏差-方差窘境**(bias-variance dilemma)



# Q&A?