西安交通大學
XI'AN JIAOTONG UNIVERSITY

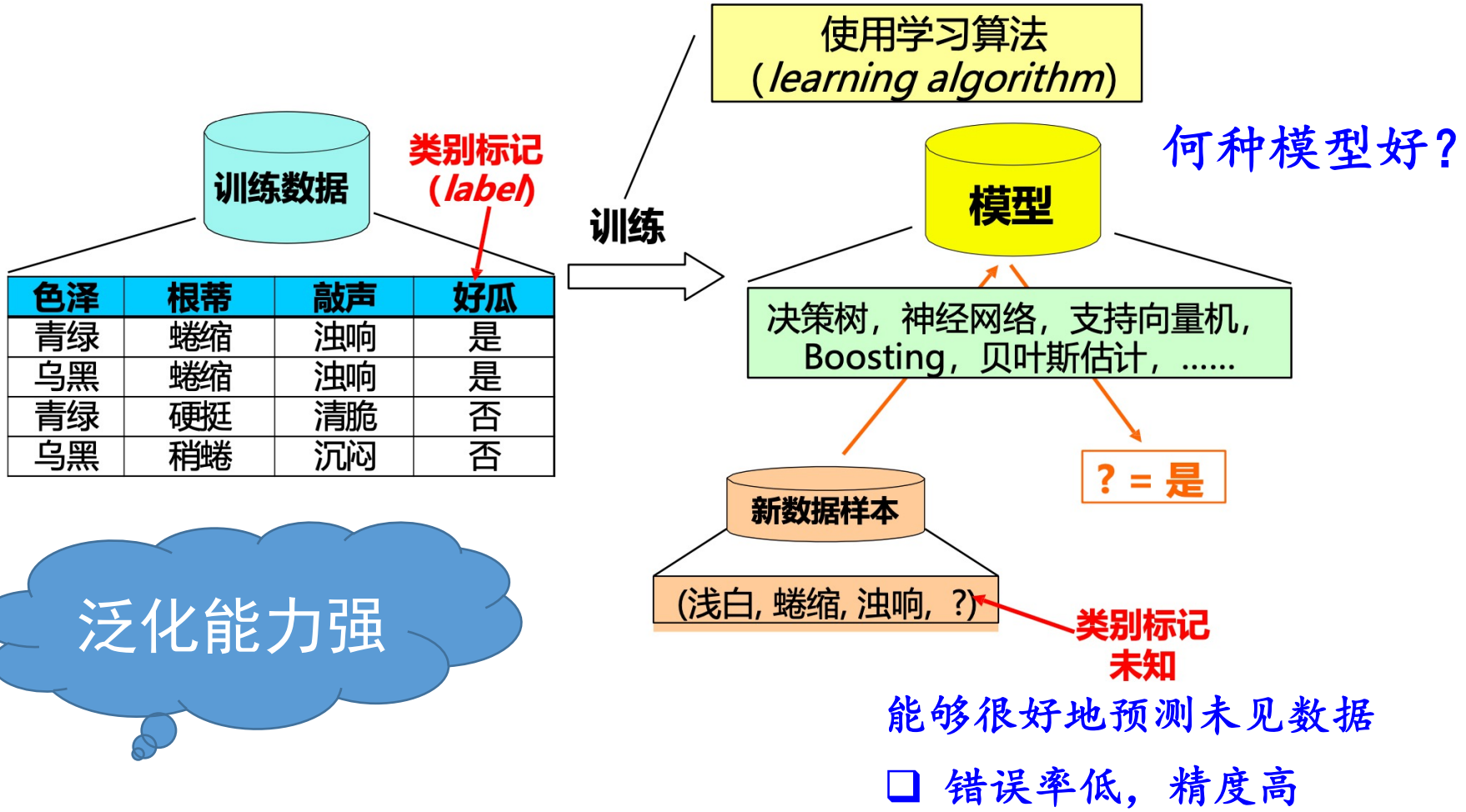
机器学习1: 模型选择与评估

Tieliang Gong 龚铁梁

School of Computer Science & Technology,

Xi'an Jiaotong University

典型的机器学习过程



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

能够很好地预测未见数据

- 错误率低，精度高

然而，我们没有“未见”数据

泛化误差 vs. 经验误差

泛化误差：在“未来”样本上的误差

经验误差：在训练样本上的误差，亦称为“训练误差”

- 泛化误差越小越好
- 经验误差是否越小越好？

不是，因为会出现过拟合(Overfitting)

过拟合 vs. 欠拟合

训练样本

新样本

过拟合模型分类结果：

不是树叶

（误以为树叶必须有锯齿）

欠拟合模型分类结果：

是树叶

（误以为绿色的都是树叶）

模型选择

三个关键问题：

- 如何获得测试结果？
- 如何评估性能优劣？
- 如何判断实质差别？

⇒

- 评估方法
- 性能度量
- 比较检验

评估方法

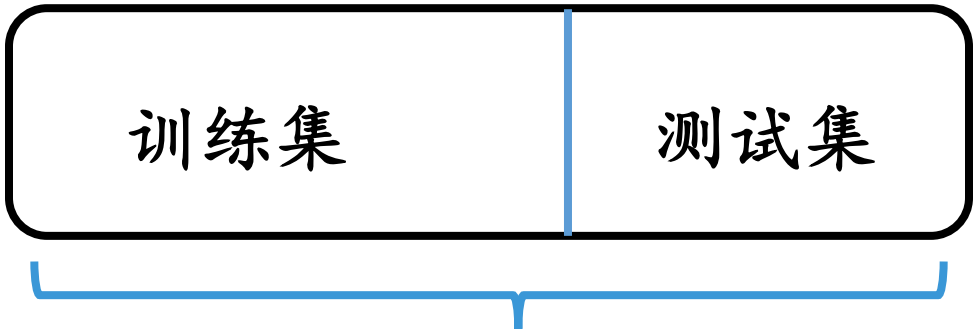
关键：如何获取“测试集”

- 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 (Bootstrap)

留出法

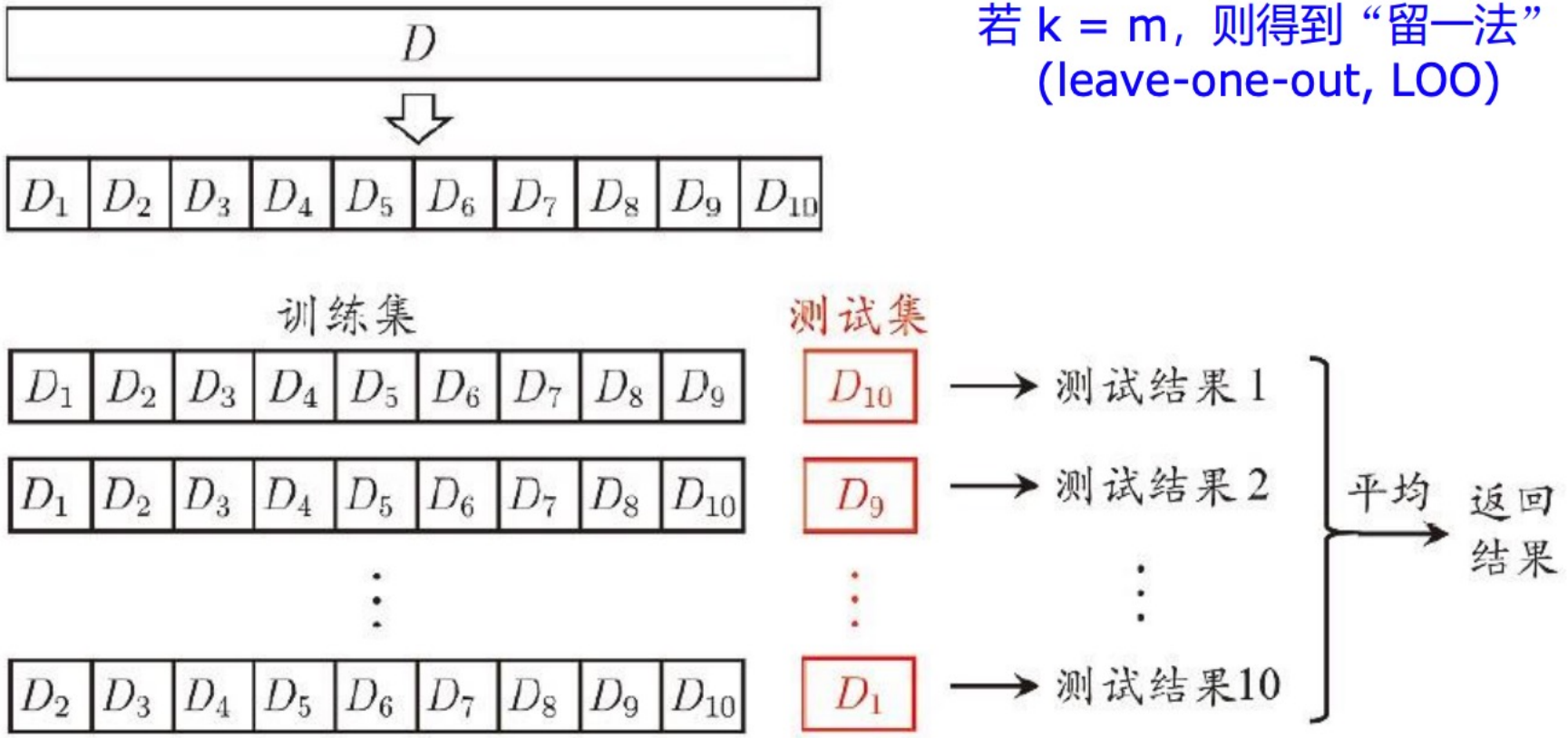


注意

- 保持数据分布一致性（分层抽样）
- 多次重复划分（例如：50次随机划分）
- 测试集不能太大、不能太小（1/5 – 1/3）

交叉验证法

若 k = m，则得到“留一法”
(leave-one-out, LOO)

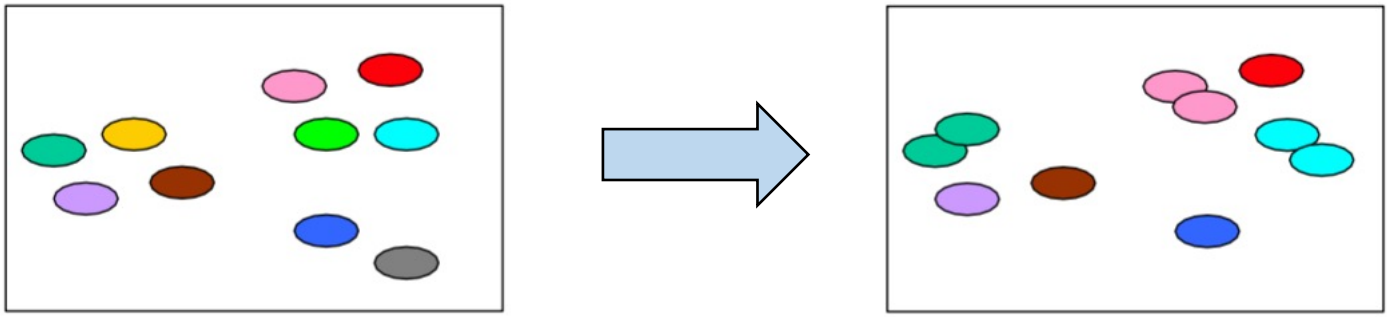


10折交叉验证示意图

自助法

基于“自助采样”(Bootstrap sampling)

亦称“有放回采样”、“可重复采样”



约有36.8%的样本不会出现

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

- 训练集与原样本集同规模
- 数据分布有所改变

“包外估计”(out-of-bag estimation)

“调参”与最终模型

- 算法的参数：一般人工设定，亦称“超参数”
- 模型的参数：一般由算法在学习过程中确定

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

调参对算法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，需用“训练集+验证集”重新训练最终模型

模型选择

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

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

使用不同性能度量往往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

❑ 回归任务常用均方误差：

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

错误率 vs. 精度

- 错误率：

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

- 精度：

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

查准率 vs. 查全 (召回)率

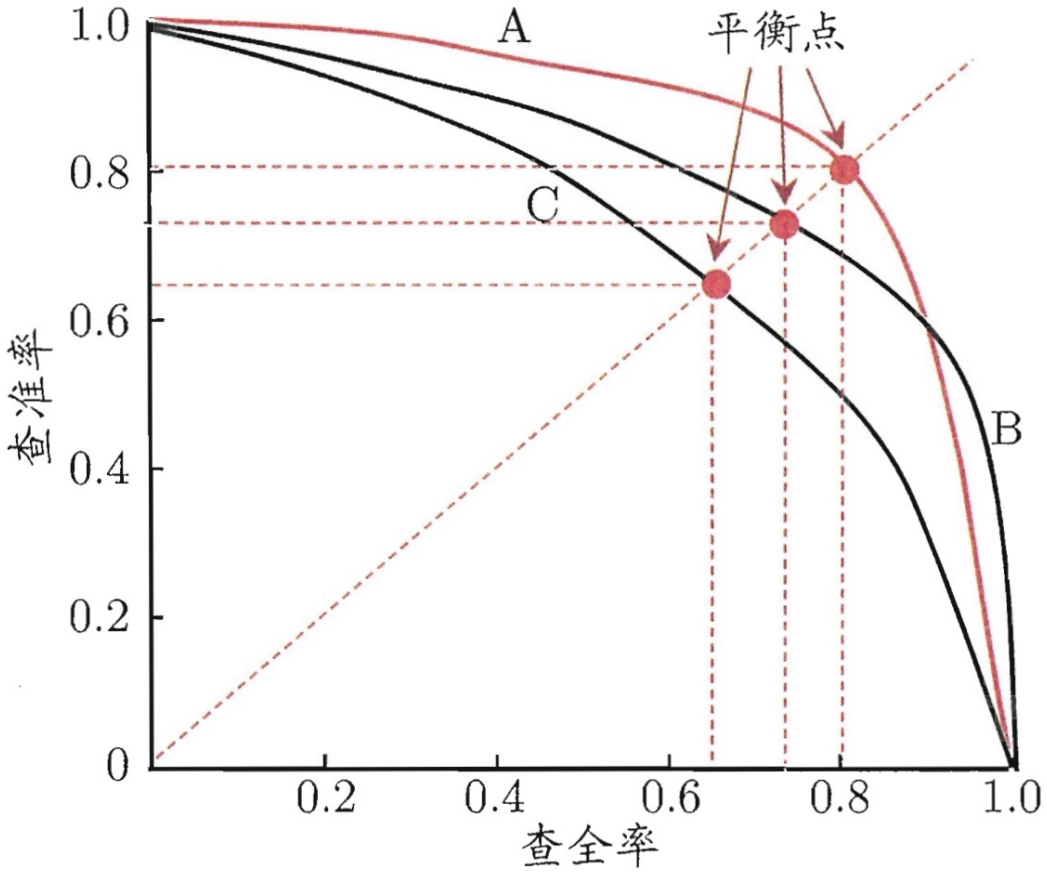
分类结果混淆矩阵

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

P-R图，BEP（Break- Event Point）

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。



P-R图：

- A 优于 C
- B 优于 C
- A ？ B

BEP

- A 优于 B
- A 优于 C
- B ？ C

F1-score

- 比 BEP 更常用的 F1 度量：

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

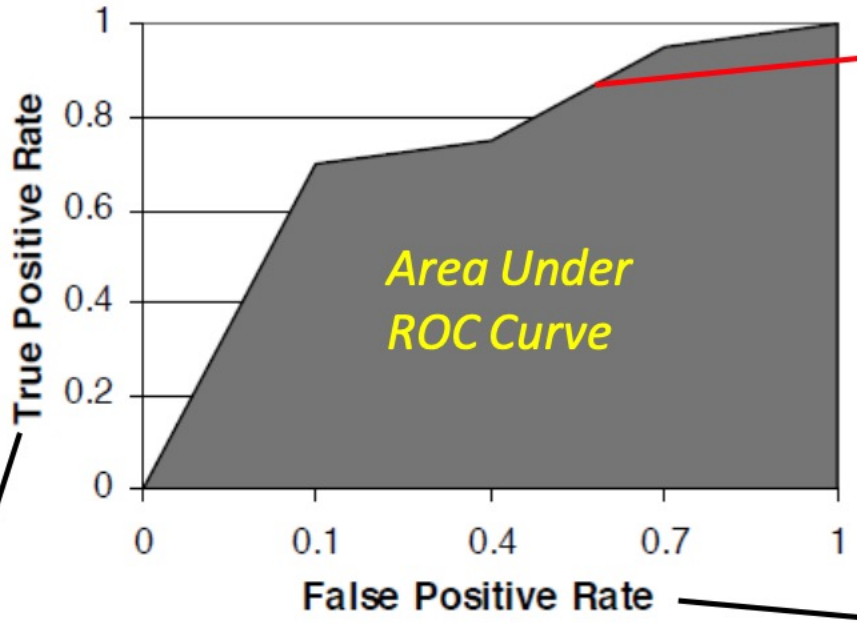
- 若对查准率/查全率有不同偏好：

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

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

ROC & AUC

AUC: **A**rea **U**nder the ROC **C**urve



ROC (Receiver Operating Characteristic) Curve [Green & Swets, Book 66; Spackman, IWML'89]

The bigger, the better

$$tpr = \frac{TP}{TP + FN} = \frac{TP}{m_+}$$

$$fpr = \frac{FP}{FP + TN} = \frac{FP}{m_-}$$

$$\mathrm{AUC} = 1 - \frac{1}{m^+ m^-} \sum_{\boldsymbol{x}^+ \in D^+} \sum_{\boldsymbol{x}^- \in D^-} \left(\mathbb{I}\left(f(\boldsymbol{x}^+) < f(\boldsymbol{x}^-)\right) + \frac{1}{2}\mathbb{I}\left(f(\boldsymbol{x}^+) = f(\boldsymbol{x}^-)\right) \right)$$

非均等代价

不同错误往往会导致不同程度的损失，需考虑“非均等代价”

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(Cost-sensitive)错误率

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

模型选择

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

比较检验

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？

不可以！原因如下：

- 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
- 机器学习算法具有一定的随机性

机器学习 ⇒ 概率近似正确

机器学习的理论基础

计算学习理论

Computational learning theory

Leslie Valiant
(莱斯利 维利昂特)
(1949-)
2010年图灵奖

PAC (Probably Approximately Correct) learning model [Valiant, 1984]

$$P(|f(\boldsymbol{x}) - y| \leq \epsilon) \geq 1 - \delta$$

假设检验

假设检验(Hypothesis test)中的假设是对学习器泛化错误率分布的某种判断或者猜想

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
- McNemar检验 (基于联列表，卡方检验)

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)

误差

“误差”包含了哪些因素？

换言之，机器学习的“误差”从何而来？

偏差-方差分解

对于回归任务，泛化误差可通过“偏差-方差分解”拆解为：

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

期望输出与真实输出的差别

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

同样大小的训练集的变动，所导致的性能变化

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

训练样本的标记与真实标记有区别

表达了当前任务上任何学习算法所能达到的期望泛化误差下界

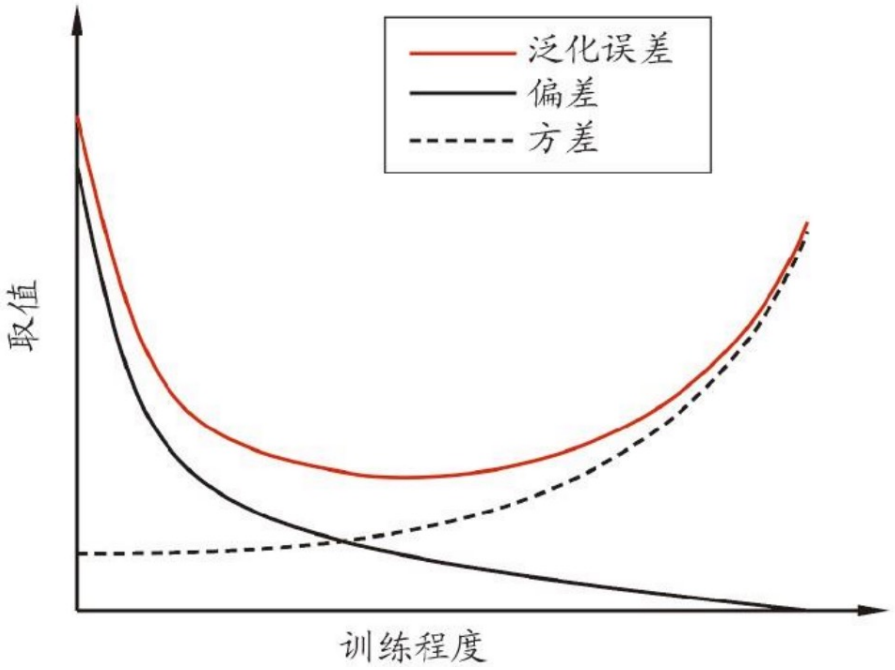
$$\varepsilon^2 = \mathbb{E}_D\left[(y_D - y)^2\right]$$

泛化性能由学习算法的能力，数据分布以及学习任务本身的难度共同决定

偏差-方差折中

一般而言，偏差与方差存在冲突：

- 训练不足时，学习器拟合能力不强，偏差主导
- 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
- 训练充足后，学习器的拟合能力很强，方差主导



泛化误差与偏差、方差的关系

谢谢!