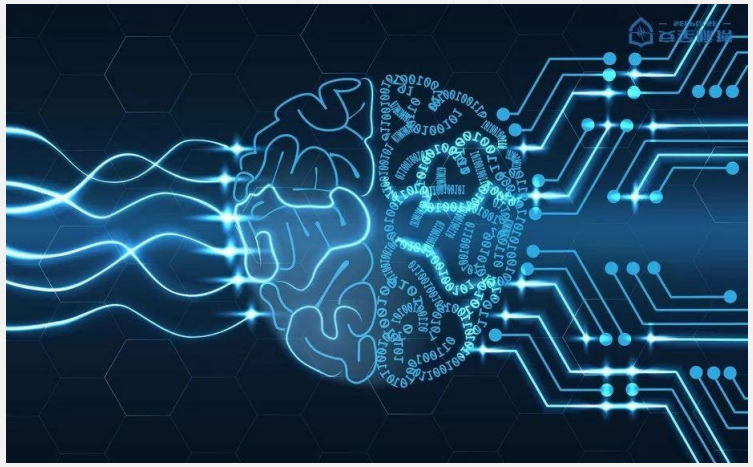


机器学习

Machine Learning

授课老师：刘雪明
电　　话：87543563
办 公 室：南一楼西303
邮　　箱：xm_liu@hust.edu.cn

第三章、模型优化与验证方法

目录
CONTENTS

本章学习的目的：掌握机器学习的基础知识

1、模型选择

➢回顾机器学习的工作流程：

❖ 选用哪一个学习算法?

❖ 使用哪一种参数配置?

模型选择问题

❖ 如何评估模型的好坏?

未知的目标函数
$f: \mathcal{X} \to \mathcal{Y}$

(理想的信用卡批准公式)

训练数据集
$\mathcal{D} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_m, y_m)\}$

(银行历史批准好坏的记录)

学习算法
$\mathcal{A}$

最终的学习函数
$g \approx f$

假设集合
$\mathcal{H}$

(包含好的和不好的假设，算法$\mathcal{A}$从中选择“最好”的作为g)

学习模型=$\mathcal{A} + \mathcal{H}$

1、模型选择

- 模型评估方法

  机器学习的目标是学到一个具有泛化能力的函数g，使得模型在新样本上表现得很好，即“泛化误差”小。

- 误差：模型的实际预测输出与样本的真实输出之间的差异
  - 训练误差（training error）：模型在训练集上的误差，也称“经验误差”，即E_{in}
  - 泛化误差（generalization error）：模型在新样本上的误差，即E_{out}，*无法直接获得*
    - ✓ 采用“测试误差”（testing error）作为泛化误差的近似

1、模型选择

- 测试误差度量的是模型在“测试集T”上的判别能力

  测试集T是从给定的数据集$\mathcal{D}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$中划分出来的，测试集$T$尽可能与训练集$S$互斥。

- 下面介绍三种常见的从$\mathcal{D}$中分离出S和T的方法：
  - 留出法（hold-out）
  - 交叉验证法（cross validation）
  - 自助法（bootstrapping）

1、模型选择

- 留出法（hold-out）

直接将数据集$\mathcal{D}$划分为两个互斥的集合，即$\mathcal{D} = S \cup T$，且$S \cap T = \emptyset$
在S上训练出模型后，用T来评估其测试误差，作为对泛化误差的估计

✓ 尽可能保证S和T分布的一致性，避免引入额外的偏差

如分类任务中保持样本的类别比例类似，类似于统计中的分层采样

✓ 单次使用留出法得到的估计结果不够稳定可靠，采用若干次随机划分，重复进行试验评估，再取平均值

✓ 存在的矛盾： S和T大小的分配问题

- *S大T小，评估结果不够稳定准确；*
- *S小T大，训练出的模型与用$\mathcal{D}$训练出的模型可能有较大差别*

1、模型选择

- 交叉验证法（cross validation），也称为“k折交叉验证”

1. 将数据集$\mathcal{D}$划分为分布尽可能一致k个互斥子集，即$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \ldots \cup \mathcal{D}_k$，且$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset\ (i \neq j)$；
2. 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；
3. 返回这k个测试结果的均值

✓ 随机使用不同的划分重复p次取均值

✓ 令$k = m$，得到**留一法**(leave-one-out)

评估结果比较准确，$\mathcal{D}$大时计算开销大

5折交叉验证示意图

1、模型选择

留出法和交叉验证法中训练集S比数据集$\mathcal{D}$小，会引入因训练样本规模不同而导致的偏差；留一法受规模不同的影响小但计算复杂度高，怎么办？

- 自助法（bootstrapping）

1. 给定m个样本的数据集$\mathcal{D}$，有放回地采样得到包含m个样本的数据集$\mathcal{D}'$
2. 用$\mathcal{D}'$作为训练集，$\mathcal{D}\backslash\mathcal{D}'$用作测试集
   - ✓ 有部分样本在$\mathcal{D}'$中重复出现，而另一部分样本不出现

     始终不出现的概率为$(1-\frac{1}{m})^m$，*其极限*$\lim_{m\to\infty}(1-\frac{1}{m})^m=\frac{1}{e}\approx 0.368$
   - ✓ 在数据集小，难以划分训练集和测试集有用
   - ✓ 改变了初始数据集的分布，引入估计偏差，数据量足够时，留出法和交叉验证法更常用

1、模型选择

➢ 调参与最终模型

- 学习算法的参数配置不同，会导致模型的性能有显著差别。

- 进行模型选择时，除了选择学习算法，还需进行参数调节（调参）
  - ✓ 对每个参数选定一个范围和变化步长

- 模型评估与选择过程只用了部分数据进行训练，在学习算法和参数配置完成后，应使用数据集D对模型进行重新训练，得到最终模型

1、模型选择

- 测验：给定包含1000个样本的数据集，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，试计算有多少种划分方式。

$$C_{500}^{300} \times C_{500}^{300}$$

1、模型选择

➢ 模型性能度量

模型泛化能力的评价标准，可使用E_{out}来度量，E_{out}可通过E_{in}来估计。

给定m个样本的数据集$\mathcal{D}$，样本的真实标记为y_i，模型预测的结果为$g(\boldsymbol{x}_i)$

- 对于**回归任务**，其性能可用**均方误差**$\mathrm{E}(g;\mathcal{D})$来度量

$$\mathrm{E}(g;\mathcal{D}) = E_{\text{in}} = \frac{1}{m}\sum_{i=1}^{m}(g(\boldsymbol{x}_i) - y_i)^2$$

- 对于**分类任务**，其性能可用**错误率**$\mathrm{E}(g;\mathcal{D})$来度量

$$\mathrm{E}(g;\mathcal{D}) = E_{\text{in}} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(g(\boldsymbol{x}_i) \neq y_i)$$

或用精度$\mathrm{acc}(g;\mathcal{D})$来度量　　$\mathrm{acc}(g;\mathcal{D}) = 1 - \mathrm{E}(g;\mathcal{D})$

1、模型选择

➢ 模型性能度量

对**分类任务**，除常用的错误率与精度外，还有如下性能度量指标

1. 查准率（Precision）、查全率（Recall）与F1

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)

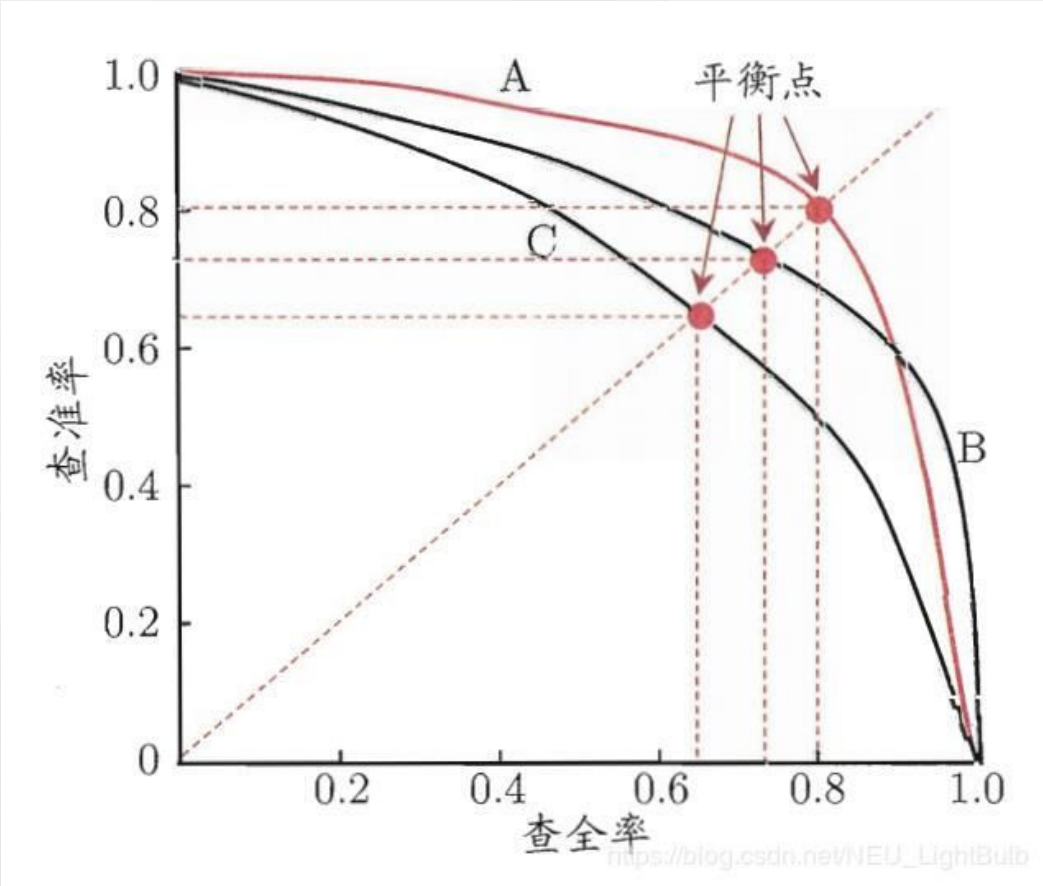
分类结果混淆矩阵

查准率P度量被预测为正样本的结果中有多少真正的正样本 $P = \frac{TP}{TP + FP}$

查全率R度量所有真实的正样本中有多少被预测正确 $R = \frac{TP}{TP + FN}$

1、模型选择

1. 查准率（Precision）、查全率（Recall）与$F1$



$P-R$曲线与平衡点示意图

- ✓ 查准率P与查全率R间存在矛盾
- ✓ 可采用$P-R$曲线下面积度量模型性能，缺点：不太容易估算
- ✓ 可采用“平衡点”作为模型性能度量，缺点：过于简化
- ➢ 常用两者的调和平均$F1$度量

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

1、模型选择

2. ROC与AUC

- ✓ 很多模型为样本产生一个**预测值**，若该预测值大于一个选定的**分类阈值**，则判为正例，否则为反例
- ✓ 根据预测值的大小可对样本进行**排序**，针对不同任务可采用不同阈值来截断，而该排序体现了模型在一般情况下的泛化性能的好坏
- ✓ **ROC**（Receiver Rperating Characteristic，受试者工作特征）**曲线**可用于度量该模型的泛化性能

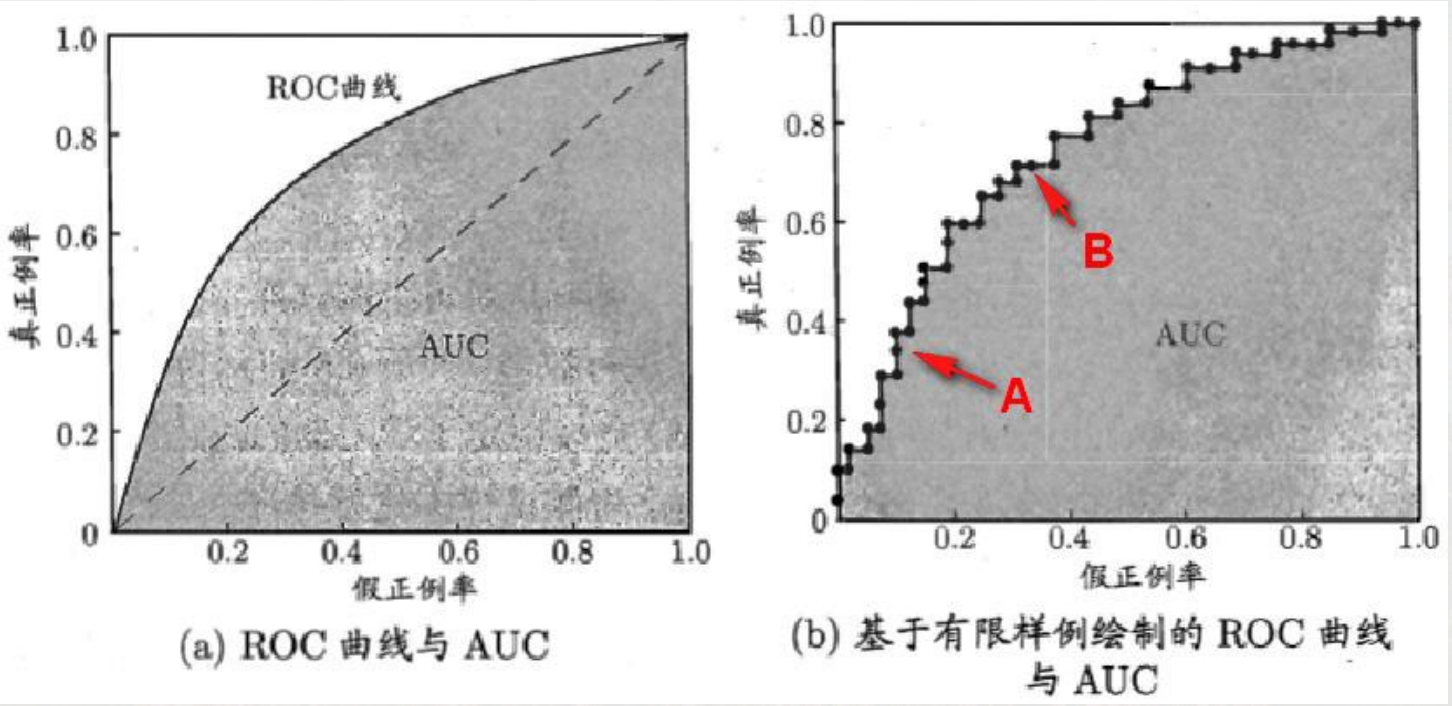
ROC曲线的纵轴是“真正例率” （True Positive Rate，TPR） $TPR = \frac{TP}{TP + FN}$

ROC曲线的横轴是“假正例率” （False Positive Rate，FPR） $FPR = \frac{FP}{TN + FP}$

1、模型选择

2. ROC与AUC

ROC曲线的纵轴是真正例率（TPR），横轴是假正例率（FPR），将分类阈值从预测值的最大值逐步取到最小值，可得到ROC曲线



ROC曲线与AUC示意图

- AUC（Area Under ROC Curve）表示的是ROC曲线下的面积，可用于比较不同模型泛化性能

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

1、模型选择

3. 代价敏感错误率

现实任务的不同错误导致的后果不同，会产生不同的代价

真实类别	预测类别	
	正例	反例
正例	0	$cost_{01}$
反例	$cost_{10}$	0

二分类代价矩阵

此时，对应的**“代价敏感”错误率**为

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

1、模型选择

- **测验：**在一个指纹门禁系统中，有999，990个正常用户的正例样本$y_i = +1$，10个入侵者的反例样本$y_i = -1$，若采用的模型$g(x)$总是返回常数+1，那么该模型的代价敏感错误率为多少?

		$g(x)$	
		+1	-1
y	+1	0	1
	-1	1000	0

A. 0.001

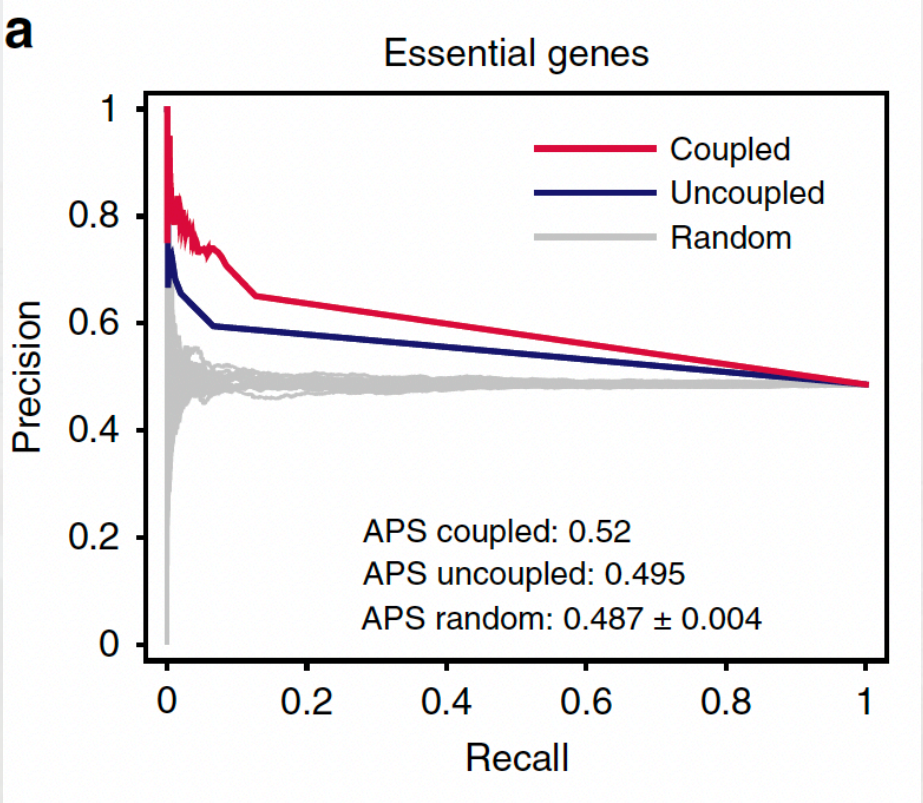
B. 0.01 ✓

C. 0.1

D. 1

1、模型选择

- 前面的模型性能度量可用于度量不同模型的性能，如右图



Xueming Liu et al., *Nature Communications* (2020) 11:6043

- 在测试集数据上发现模型A比模型B好，若想说明A的泛化性能在统计意义上优于B，还需进行统计检验分析。

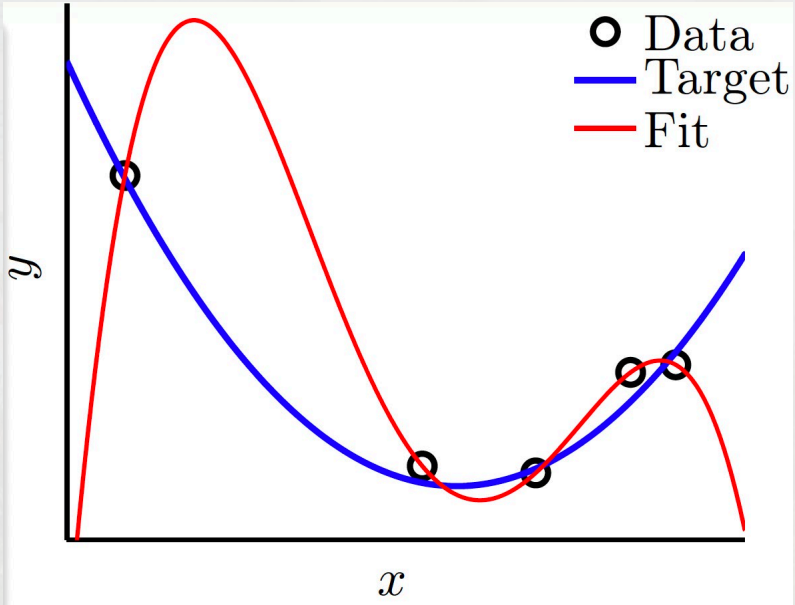
- **自学内容：**西瓜书P37，2.4比较检验

2、过拟合问题

- 在模型将训练样本学得“太好”的时候，会把训练样本本身特点当成一般性质，从而导致泛化性能下降，这中现象叫做**过拟合(Overfitting)**

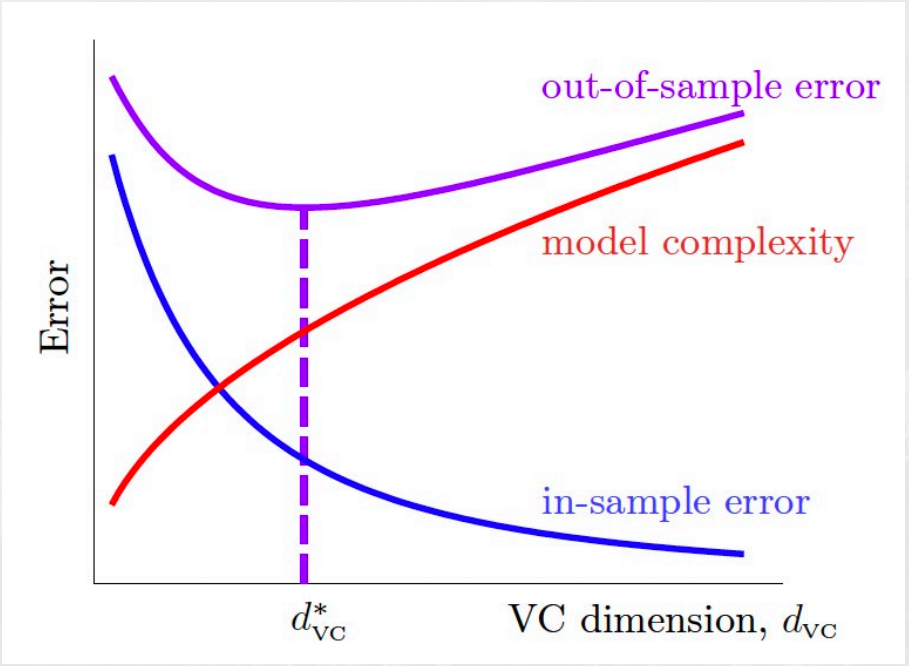
- 如右图给定目标函数为2阶多项式，从中取5个点并加上少量的噪声作为训练样本集$\mathcal{D}$

  - ✓ 学习到的模型为4阶多项式，穿过这5个数据点$E_{\text{in}} = 0$
  - ✓ 而对应的泛化误差E_{out}却很大



2、过拟合问题

➢ 回顾VC维、模型复杂度与误差间的关系



✓ 在最优d^*_{VC}的右侧，随着VC维的增大，E_{in} ↓，E_{out} ↑，发生过拟合现象

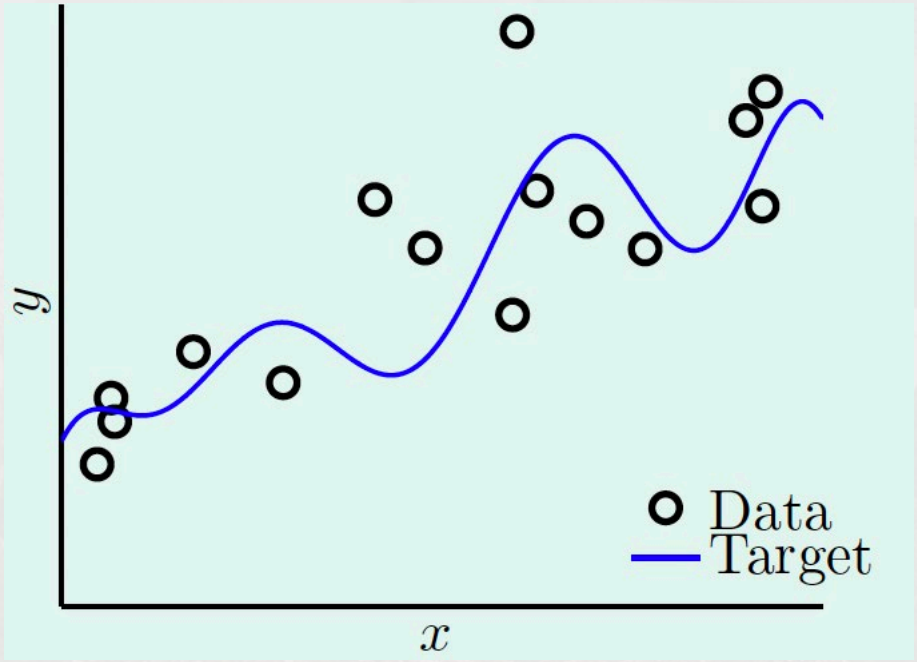
✓ 在最优d^*_{VC}的左侧，随着VC维的减小，E_{in} ↑，E_{out} ↑，发生欠拟合（underfitting）现象

VC维的大小会影响过拟合现象

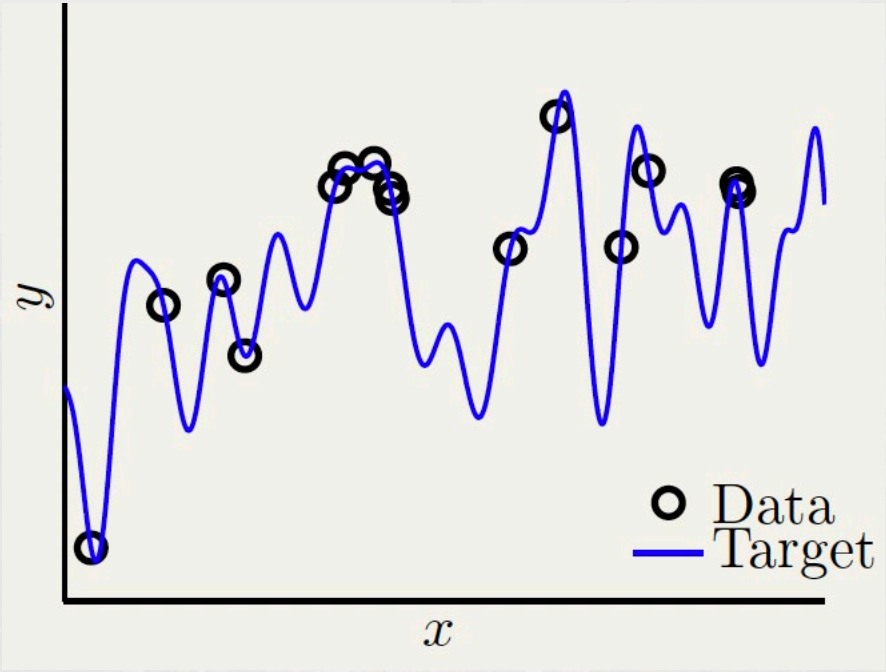
2、过拟合问题

- 影响过拟合现象的因素分析。给定如下两组数据点

目标函数为10次多项式+噪声



目标函数为50次多项式、无噪声



用2次多项式和10次多项式来对这两组数据进行拟合

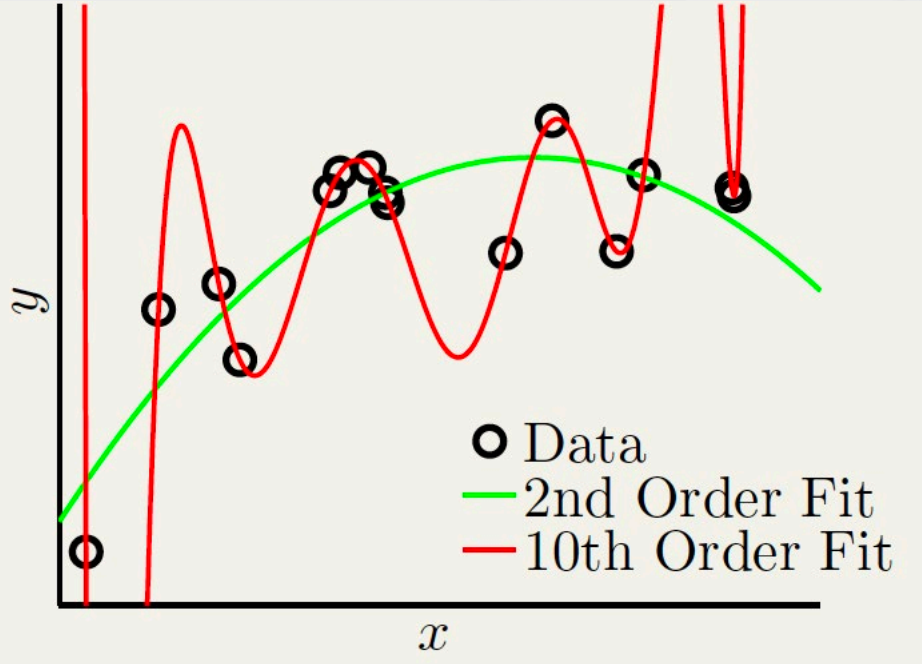
2、过拟合问题

- 影响过拟合现象的因素分析。给定如下两组数据点

目标函数为10次多项式+噪声

	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.050	0.034
E_{out}	0.127	**9.00**

目标函数为50次多项式、无噪声



	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.029	0.00001
E_{out}	0.120	**7680**

2、过拟合问题

- 影响过拟合现象的因素分析。给定如下两组数据点

目标函数为10次多项式+噪声

	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.050	0.034
E_{out}	0.127	**9.00**

- ✓ 影响因素：噪声

目标函数为50次多项式、无噪声

	$g_2 \in \mathcal{H}_2$	$g_{10} \in \mathcal{H}_{10}$
E_{in}	0.029	0.00001
E_{out}	0.120	**7680**

- ✓ 影响因素：问题复杂度（维度灾难）

- ✓ 此外，数据集的规模也是一重要的影响因素

- 影响过拟合现象的因素：

数据集规模、噪声、问题复杂度、VC维

2、过拟合问题

- 测验：给定一个数据集，下面那种情形发生过拟合现象的风险最小？

A. 小噪声，学习函数的VC维从小增大到中等大小 ✓

B. 小噪声，学习函数的VC维从小增大到很大

C. 大噪声，学习函数的VC维从小增大到中等大小

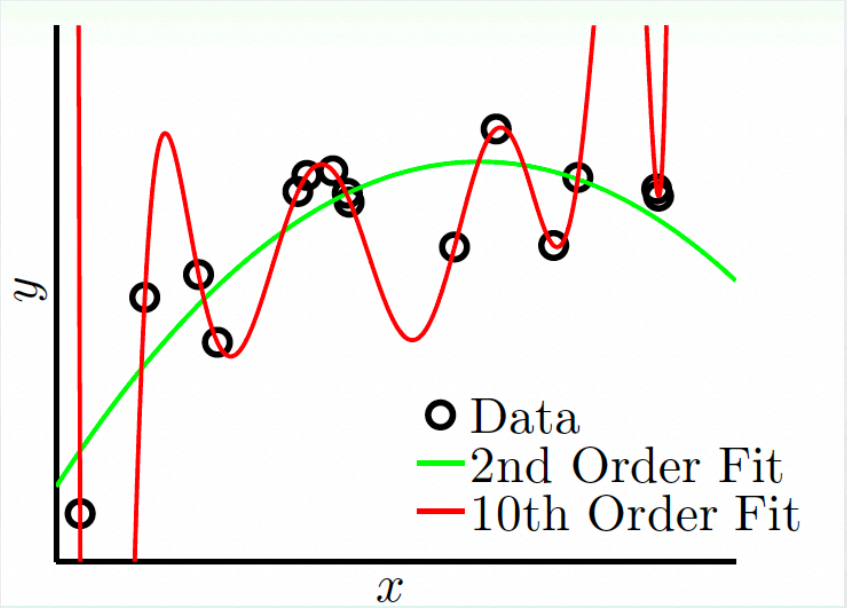
D. 大噪声，学习函数的VC维从小增大到很大

2、过拟合问题

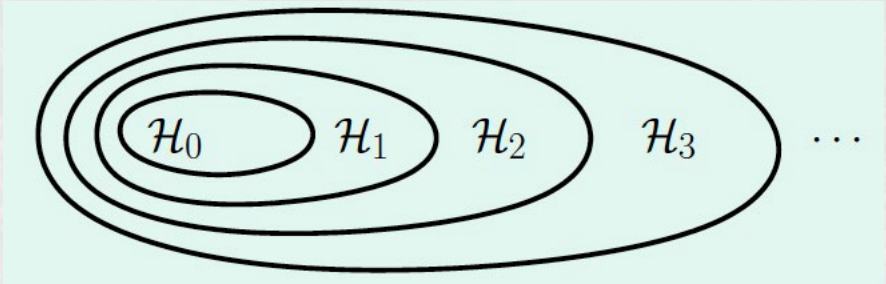
- 如何减少过拟合现象的产生？
  - ✓ 从简单的学习函数开始进行拟合
  - ✓ 对训练数据集里label明显错误的样本进行修正或剔除（data cleaning/pruning）
  - ✓ 对样本数少的情形，可对已知样本进行简单处理、变换，从而获得更多的样本（Data hinting）
  - ✓ 正则化（Regularization）等

3、正则化方法

- 下面以多项式拟合为例来介绍正则化（Regularization）方法



正则化思想：从高次函数假设空间$\mathcal{H}_{10}$退回低次函数空间$\mathcal{H}_2$



3、正则化方法

- 如何从高次函数假设空间$\mathcal{H}_{10}$退回低次函数空间$\mathcal{H}_2$?

  假设空间$\mathcal{H}_{10}$的函数：$h(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_{10} x^{10}$

  假设空间$\mathcal{H}_2$的函数：$h(x) = w_0 + w_1 x + w_2 x^2$

  - 当权重满足$w_3 = w_4 = \cdots = w_{10} = 0$时，$\mathcal{H}_{10} = \mathcal{H}_2$
  - 定义权重向量$\boldsymbol{w}^T = [w_0, w_1, \ldots, w_n]$，即$\boldsymbol{w} \in \mathbb{R}^{n+1}$

假设空间$\mathcal{H}_{10}$拟合：

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

假设空间$\mathcal{H}_2$拟合：

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

$$\text{s.t.} \quad w_3 = w_4 = \cdots = w_{10} = 0$$

3、正则化方法

➢ 对限制条件进行放松，推广到更一般的情形：

假设空间$\mathcal{H}_2$拟合：

$$\min_{\boldsymbol{w}\in\mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

$$\text{s.t.} \quad w_3 = w_4 = \cdots = w_{10} = 0$$

→

假设空间$\mathcal{H}'_2$拟合：

$$\min_{\boldsymbol{w}\in\mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

$$\text{s.t.} \quad \sum_{q=0}^{10} \mathbb{I}(w_q \neq 0) \leq 3$$

- 假设空间$\mathcal{H}'_2$比$\mathcal{H}_2$更灵活：$\mathcal{H}_2 \subset \mathcal{H}'_2$
- 假设空间$\mathcal{H}'_2$比$\mathcal{H}_{10}$过拟合风险更低：$\mathcal{H}'_2 \subset \mathcal{H}_{10}$
- 缺点：$\sum_{q=0}^{10} \mathbb{I}(w_q \neq 0) \leq 3$是离散不等式，不易求解

3、正则化方法

- 对于向量$\boldsymbol{w}$，$\|\boldsymbol{w}\|_0 = \sum_{q=0}^{10} \mathbb{I}(w_q \neq 0)$ 表示其0-范数

$$\|\boldsymbol{w}\|_1 = \sum_{q=0}^{10} |w_q|$$ 表示其1-范数

$$\|\boldsymbol{w}\|_2 = \sqrt{\sum_{q=0}^{10} w_q^2}$$ 表示其**2-范数**

- 将假设空间$\mathcal{H}'_2$的限制条件进行变换:

假设空间$\mathcal{H}'_2$拟合:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

$$\text{s.t.} \quad \sum_{q=0}^{10} \mathbb{I}(w_q \neq 0) \leq 3$$

→

假设空间$\mathcal{H}(C)$拟合:

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

$$\text{s.t.} \quad \|\boldsymbol{w}\|_2{}^2 = \sum_{q=0}^{10} w_q^2 \leq C$$

- $\mathcal{H}(0) \subset \mathcal{H}(1.024) \subset \ldots. \subset \mathcal{H}(1024) \subset \ldots. \subset \mathcal{H}(\infty) = \mathcal{H}_{10}$

3、正则化方法

假设空间 $\mathcal{H}(C) \equiv \{\boldsymbol{w} \in \mathbb{R}^{10+1}, 且 \|\boldsymbol{w}\|_2 \leq C\}$ 拟合：

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w})$$

$$\text{s.t. } \|\boldsymbol{w}\|_2{}^2 = \sum_{q=0}^{10} w_q^2 \leq C$$

- 满足上述优化问题的假设称为正则化假设$\boldsymbol{w}_{\text{REG}}$。且该优化问题可等价转化为：

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad E_{\text{in}}(\boldsymbol{w}) + \lambda \sum_{q=0}^{10} w_q^2$$

其中，正则化参数 λ为一可调系数。对于多项式拟合，该优化问题为

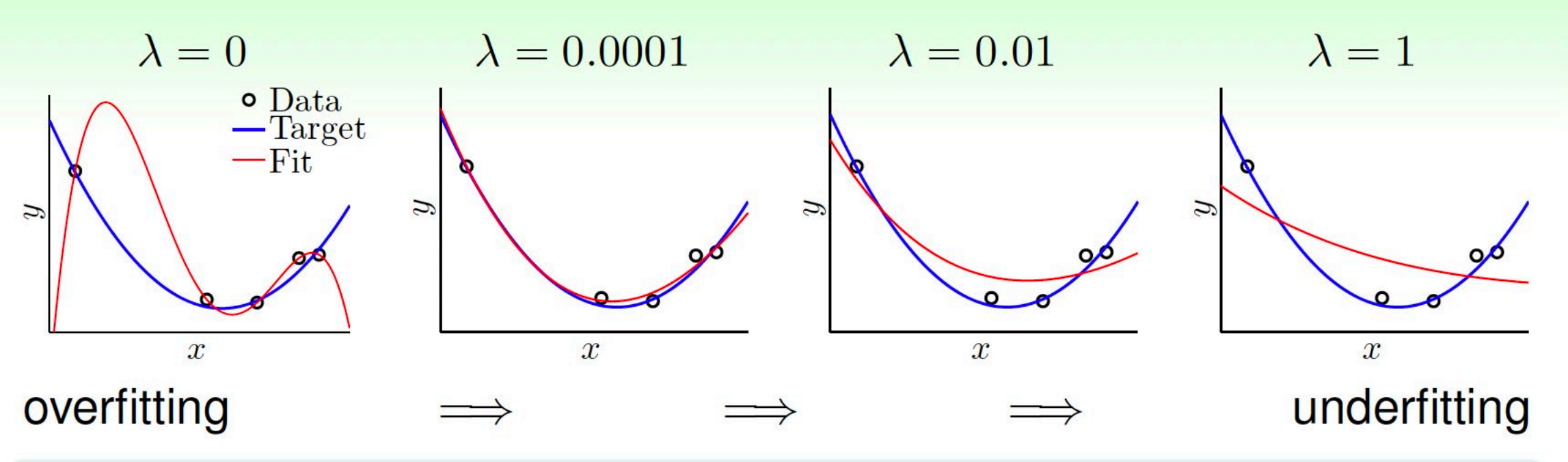
$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad \sum_{i=0}^{m} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w}$$

$\mathcal{L}_2$正则项

3、正则化方法

$$\min_{\boldsymbol{w} \in \mathbb{R}^{10+1}} \quad \sum_{i=0}^{m}(\boldsymbol{w}^T\boldsymbol{x} - y_i)^2 + \lambda \boldsymbol{w}^T\boldsymbol{w}$$

- 通过调节正则化参数λ的值，可实现从过拟合到欠拟合的调节：



3、正则化方法

- 测验：对于$Q \geq 1$，下面哪个假设权重向量（$\boldsymbol{w} \in \mathbb{R}^{Q+1}$）不属于正则化假设集合$\mathcal{H}(1)$

A. $\boldsymbol{w}^T = [0, 0, \ldots, 0]$

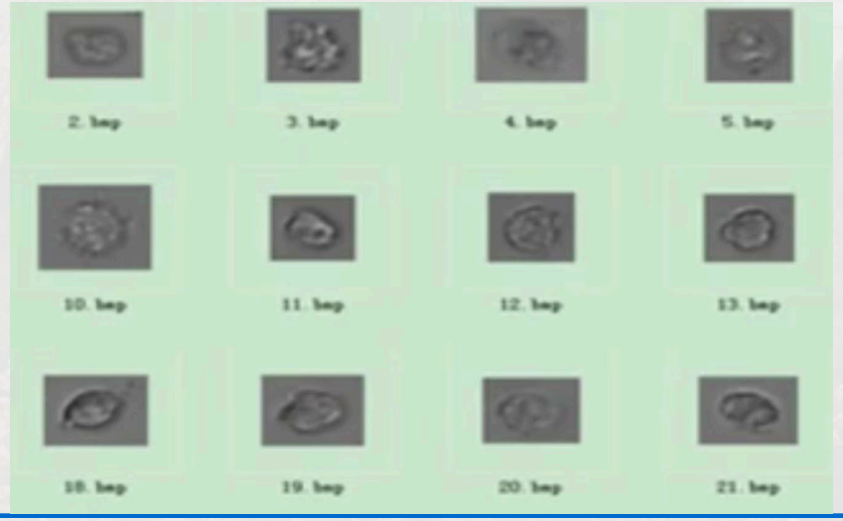
B. $\boldsymbol{w}^T = [1, 0, \ldots, 0]$

✓ C. $\boldsymbol{w}^T = [1, 1, \ldots, 1]$

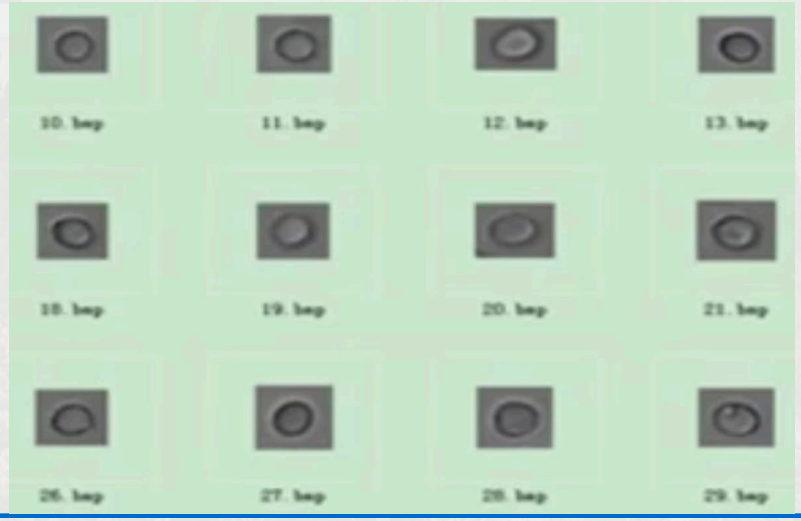
D. $\boldsymbol{w}^T = [\sqrt{\frac{1}{Q+1}}, \sqrt{\frac{1}{Q+1}}, \ldots, \sqrt{\frac{1}{Q+1}}]$

4、特征选择

- 特征是在观测现象中的独立、可测量的属性
  - 对当前学习任务有用的属性称为“相关特征”
  - 对当前学习任务没用的属性称为“无关特征”
- 从给定的特征集合中选择相关特征子集的过程，叫**特征选择**
- 例子：基于SVM的红、白细胞识别



白细胞图像



红细胞图像

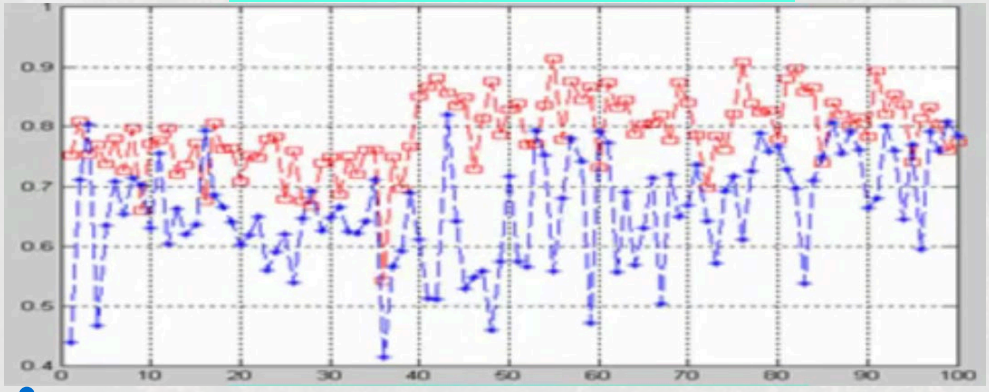
4、特征选择

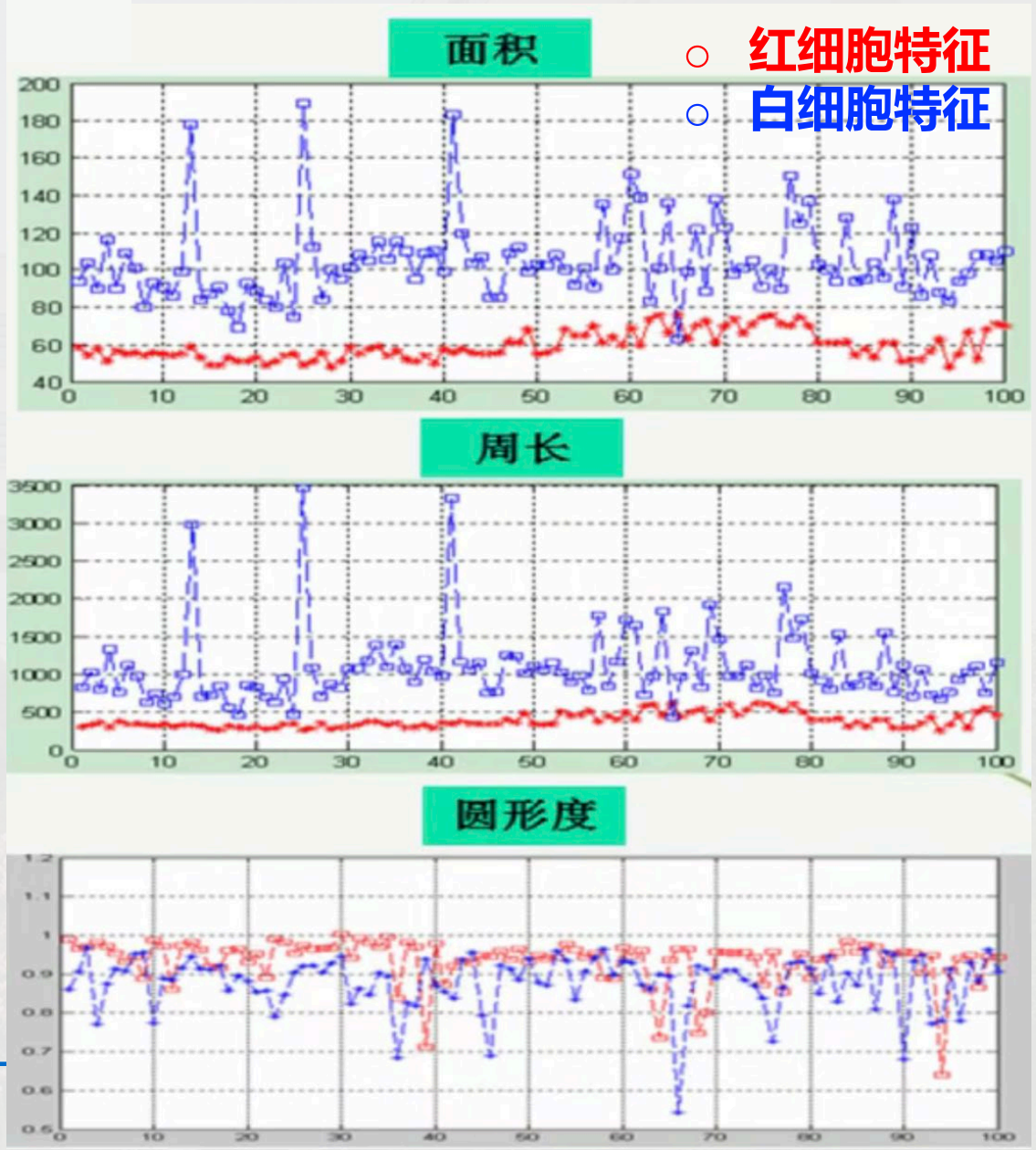
- 基于SVM的红、白细胞识别

特征集合：

- 面积
- 周长
- 圆形度
- 纹理特征：灰度共生矩阵等

灰度共生矩阵相关性



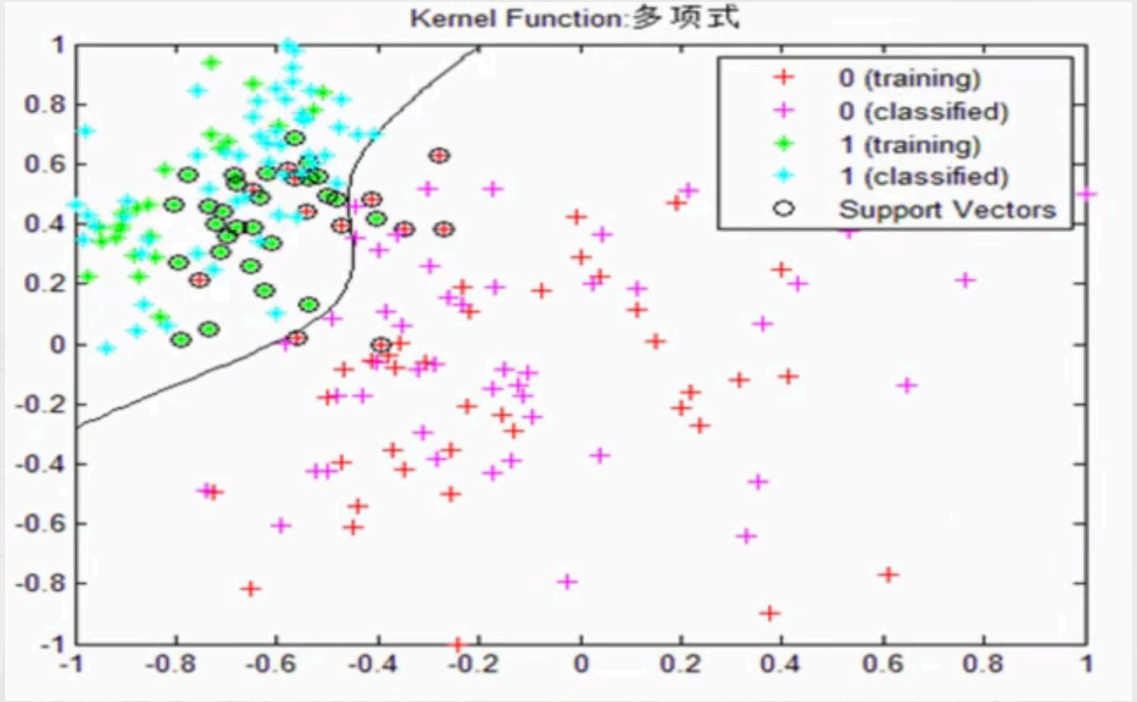


4、特征选择

- 例子：红、白细胞识别

特征集合：

- 面积
- 周长 （冗余特征）
- 圆形度
- 纹理特征：灰度共生矩阵等



- 特征选择是一个重要的“数据预处理”过程
  - 避免维数灾难，降低学习任务难度，减少计算开销
  - 特征选择可能会降低模型的预测能力

4、特征选择

➢ 特征选择包含特征子集搜索机制与子集评价机制

- 子集搜索

✓ 前向搜索：1. 将每一特征作为一子集，选定最优的单个特征$\{x_i\}$
2. 在上一轮选定集中增加一个特征，选定最优$\{x_i, x_j\}$
3. …，直到最优$k+1$特征子集不如上轮选定集结束

✓ 后向搜索：从完整特征子集开始，每次去掉一个无关特征

✓ 双向搜索：每一轮增加选定特征（此后不删除），并减少无关特征

4、特征选择

- 特征选择包含特征子集搜索机制与子集评价机制

- 子集评价

  - ✓ 给定数据集$\mathcal{D}$，假定$\mathcal{D}$中第i类样本所占比例为$p_i, (i = 1,2, \dots, |\mathcal{Y}|)$，其信息熵定义为

$$\mathrm{Ent}(\mathcal{D}) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

  - ✓ 对属性子集A，根据其取值将$\mathcal{D}$分成V个子集$\{\mathcal{D}^1, \mathcal{D}^2, \dots \mathcal{D}^V\}$，每个子集中样本在$A$上取值相同，则属性$A$的信息增益为

$$\mathrm{Gain}(A) = \mathrm{Ent}(\mathcal{D}) - \sum_{v=1}^{|V|} \frac{|\mathcal{D}^v|}{|\mathcal{D}|} \mathrm{Ent}(\mathcal{D}^v)$$

信息增益$\mathrm{Gain}(A)$越大，特征子集A包含的有助于分类的信息越多

4、特征选择

- 特征选择法大致可分为：过滤式、包裹式、嵌入式
  - 过滤式选择
    - ✓ 特征选择与后续学习**无关**
    - ✓ 设计一个相关统计量分量δ^j来度量每个特征j的重要性，如Relief法

对样本$\boldsymbol{x}_i$，在其同类样本中选择最近邻样本$\boldsymbol{x}_{i,nh}$，称为“猜中近邻”

对样本$\boldsymbol{x}_i$，在其异类样本中选择最近邻样本$\boldsymbol{x}_{i,nm}$，称为“猜错近邻”

$$\delta^j = \sum_i \left[-\mathrm{diff}(x_i^j, x_{i,nh}^j)^2 + \mathrm{diff}(x_i^j, x_{i,nm}^j)^2 \right]$$

x_a^j表示样本$\boldsymbol{x}_a$在属性j上的取值； 若$x_i^j = x_{i,nh}^j$，则$\mathrm{diff}\left(x_a^j, x_b^j\right) = 0$，*否则为1(或*$\left|x_a^j - x_b^j\right|$*)*

- ✓ 特征子集的重要性则由子集中每个特征的δ^j之和决定

4、特征选择

- 特征选择法大致可分为：过滤式、包裹式、嵌入式

- 包裹式选择
  - ✓ 将最终模型的性能作为特征子集的评价标准，如LVW(Las Vegas Wrapper)方法

1. 初始化最优特征子集$A^* = A$， A为初始完整特征子集，训练模型，用交叉验证法估计其误差
2. 随机产生特征子集A'，训练模型并计算误差，若它比当前A^*对应的误差更小，则令$A^* = A'$
3. 重复步骤2，直至运行时间达到上限，输出A^*

- *最终模型性能比过滤式选择的更好，但计算开销大，若时间限制，有可能给不出解*

4、特征选择

- 特征选择法大致可分为：过滤式、包裹式、嵌入式

- 嵌入式选择
  - ✓ 特征选择过程与模型训练过程融为一体，在模型训练的过程中自动进行特征选择。如L_1正则化

用L_1范数替代多项式拟合中的L_2范数的平方，模型优化函数为

$$\min_{\boldsymbol{w}} \quad \sum_{i=0}^{m}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2 + \lambda\|\boldsymbol{w}\|_1 \qquad \text{其中}\|\boldsymbol{w}\|_1 = \sum_q w_q$$

用梯度下降法求使得$\sum_{i=0}^{m}(\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2$最小的$\boldsymbol{w}$，同时考虑$L_1$范数的最小化

西瓜书P253

5、偏差-方差平衡

- 为进一步理解模型的泛化性能，下面对模型泛化错误率$E(g;\mathcal{D})$进行分解，即“偏差-方差分解”

  - 不同训练集得到的学习模型不同，对测试样本$\boldsymbol{x}$，令
    - $y_{\mathcal{D}}$为$\boldsymbol{x}$在数据集中的标记；
    - y为$\boldsymbol{x}$的真实标记；
    - $g(x;\mathcal{D})$为训练集$\mathcal{D}$上学得的模型g在$\boldsymbol{x}$上的预测输出

✓ 数据集本身噪声为 $\varepsilon^2 = \mathbb{E}_{\mathcal{D}}[(y_{\mathcal{D}} - y)^2]$

5、偏差-方差平衡

- ✓ 训练集$\mathcal{D}$得到模型的**期望预测**为 $\bar{g}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}[g(\boldsymbol{x};\mathcal{D})]$
- ✓ 使用样本数相同的不同训练集$\mathcal{D}$产生的方差（variance）为

$$var(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}[(g(\boldsymbol{x};\mathcal{D}) - \bar{g}(\boldsymbol{x}))^2]$$

- ✓ 期望输出与真实标记的差别称为偏差（bias），即

$$bias^2(\boldsymbol{x}) = (\bar{g}(\boldsymbol{x}) - y)^2$$

- ➢ 假定噪声期望为零，即$\mathbb{E}_{\mathcal{D}}[y_{\mathcal{D}} - y] = 0$，对期望泛化误差分解得

$$E(g;\mathcal{D}) = \mathbb{E}_{\mathcal{D}}[(g(\boldsymbol{x};\mathcal{D}) - y_{\mathcal{D}})^2] = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \varepsilon^2$$

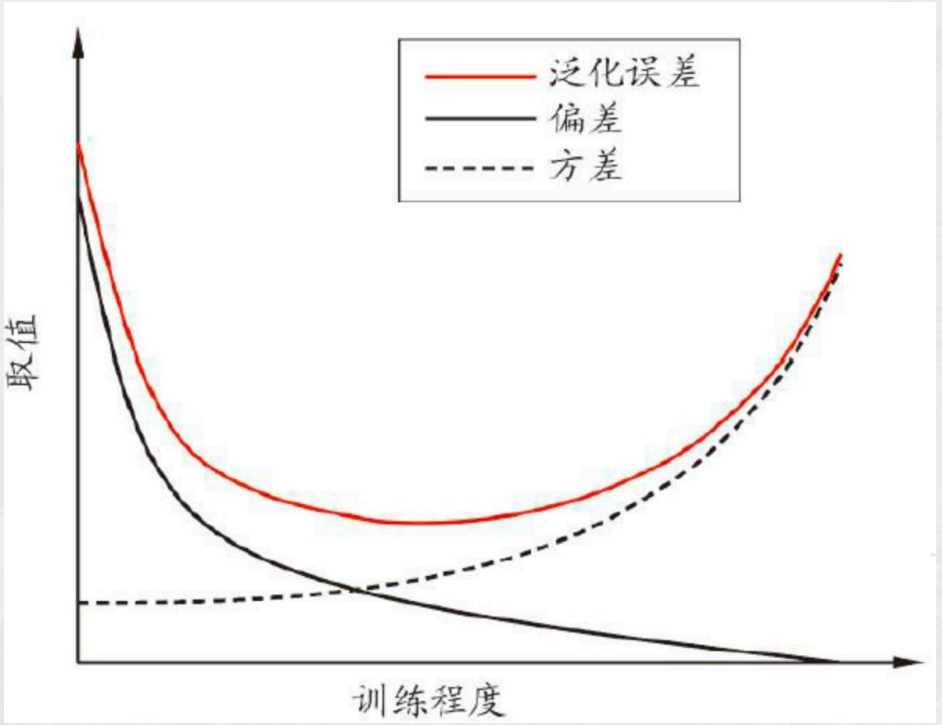
5、偏差-方差平衡

$$E(g;\mathcal{D}) = \mathbb{E}_{\mathcal{D}}[(g(\boldsymbol{x};\mathcal{D}) - y_{\mathcal{D}})^2] = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \varepsilon^2$$

- 泛化误差可分解为偏差、方差与噪声之和
  - ✓ 偏差度量了学习算法的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的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响
  - ✓ 噪声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的下界，即刻画了学习问题本身的难度

5、偏差-方差平衡

- 偏差与方差是有冲突的，称为“偏差-方差窘境”



- 训练不足时，模型拟合能力差，偏差主导了泛化误差
- 训练充足后，模型拟合能力强，数据本身特点会被学到，发生过拟合，对数据扰动敏感，方差主导泛化误差

本章小结

- ✓ 模型选择方法：留出法、交叉验证法、自助法
- ✓ 模型性能度量指标：
  - 回归任务：均方误差
  - 分类任务：错误率（或精度）；查准率、查全率与$F1$；ROC与AUC；代价敏感错误率
- ✓ 过拟合产生因素：数据集规模、噪声、问题复杂度、VC维
- ✓ $\mathcal{L}_2$正则化方法来减少过拟合
- ✓ 特征选择方法（子集搜索与评价）：过滤式、包裹式、嵌入式
- ✓ 模型偏差-方差分解与平衡