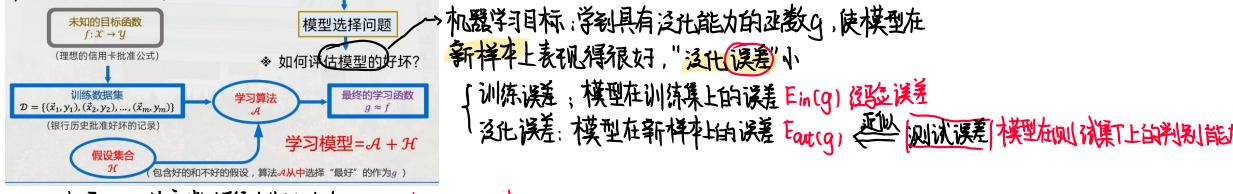


模型选择

机器学习工作流程



本测试集 T 从给定数据集中划分出来的, 训练集 S 与测试集 T 尽量相斥

划分方法

1. 留出法

- 直接将数据集 D 划分为两个互斥的集合, 即 $D = S \cup T$, 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, 分重叠进行试验评估, 再取平均值
- 存在的矛盾: S 和 T 大小的分配问题
 - S 太小, 评估结果不够稳定准确;
 - S 太大, 训练出的模型与用 D 训练出的模型可能有较大差别。

2. 交叉验证法, 也称“k折交叉验证”

- 将数据集 D 划分为分布尽可能一致 k 个互斥子集, 即 $D = D_1 \cup D_2 \dots \cup D_k$, 且 $D_i \cap D_j = \emptyset (i \neq j)$;
 - 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;
 - 返回这 k 个测试结果的均值
 - 随机使用不同的划分重复 k 次取均值
 - 令 $k = m$, 得到留一法 (Leave-one-out)
- 评估结果比较准确, k 大计算开销大

5折交叉验证示意图

3. 自助法

- 自助法 (bootstrapping)
- 给定 m 个样本的数据集 D , 有放回地采样得到包含 m 个样本的数据集 D'
- 用 D' 作为训练集, $D \setminus D'$ 用作测试集
- 有部分样本在 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 $\lim_{m \rightarrow \infty} (1 - \frac{1}{m})^m = e^{-1} \approx 0.368$
- 在数据集小, 难以划分训练集和测试集有影响
- 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

留出与 k -fold 中 S 比 k 小引入训练样本规模不同导致的误差
 留一法计算复杂度高

调查 \Rightarrow 对每个参数选定一个范围和变化步长 (参数配置导致模型性能差别显著)

最终模型 \Rightarrow 上述评估过程仅使用部分数据, 在学习算法与参数配置完成后, 应使用整个数据集重新训练

性能度量 泛化能力评价标准 $\leftarrow E_{out} \rightarrow E_{in} \rightarrow$ 预测结果

$$(1) \text{ 回归: } E(g; D) = E_{in} = \frac{1}{m} \sum_{i=1}^m (g(x_i) - y_i)^2 \rightarrow \text{真实标注}$$

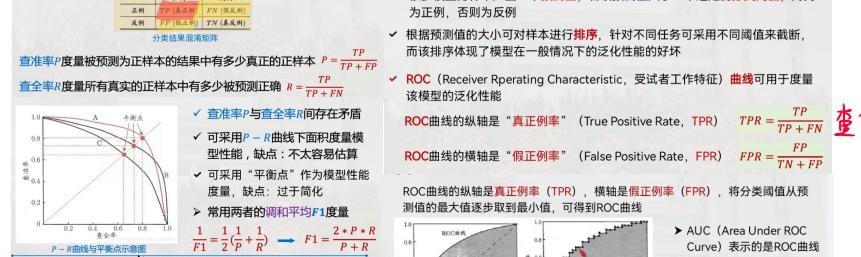
$$| \text{分类: } E(g; D) = E_{in} = \frac{1}{m} \sum_{i=1}^m I(g(x_i) \neq y_i)$$

$$\text{精度 } acc(g; D) = 1 - E(g; D)$$

对分类任务

(1). 准确率 (Precision)、查全率 (Recall) 与 F1

(2) ROC 与 AUC



(3) 代价敏感错误率 \Rightarrow 不同错误后果不同, 代价不同

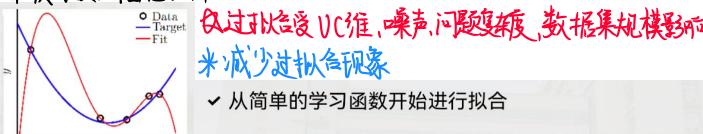
真实类别	预测类别	正例	反例
正例	正例	0	Cost ₀₁
反例	反例	Cost ₁₀	0

$$E(g; D; cost) = \frac{1}{m} \sum_{(x_i, y_i)} I(g(x_i) \neq y_i) \times cost_{01} + \sum_{(x_i, y_i)} I(g(x_i) \neq y_i) \times cost_{10}$$

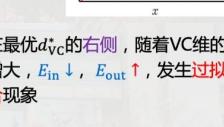
仅可比较不同模型性能; 测试集性能优不代表泛化性能优, 还需流经检验

过拟合问题 \Rightarrow 将训练样本本身特点当作一般性质导致泛化能力下降

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 D



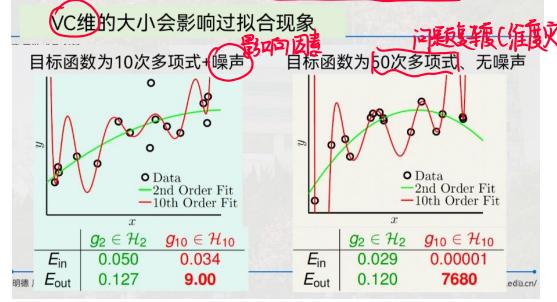
- 学习到的模型为 4 阶多项式, 穿过这 5 个数据点 $E_{in} = 0$
- 而对应的泛化误差 E_{out} 却很大



- 在最优 d_{vc}^* 的右侧, 随着 VC 维的增大, $E_{in} \downarrow$, $E_{out} \uparrow$, 发生过拟合现象
- 在最优 d_{vc}^* 的左侧, 随着 VC 维的减小, $E_{in} \uparrow$, $E_{out} \uparrow$, 发生欠拟合 (underfitting) 现象

- 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

从高次函数段落空间 H_n 拉回到低次函数空间 H_l .



正则化方法

$$g_{10} \neq h(x) = w_0 + w_1 x + w_2 x^2 + \dots + w_{10} x^{10}$$

$$g_{10} \neq h(x) = w_0 + w_1 x + w_2 x^2$$

$w_0 = w_1 = \dots = w_{10} = 0$ 时 $E_{in} = E_{out}$, 完成退化

$$\min_{w \in \mathbb{R}^{n+1}} E_{in}(w) \quad \text{退化} \quad \min_{w \in \mathbb{R}^{n+1}} E_{in}(w) \quad \text{W.R.T.} \quad w \in \mathbb{R}^{n+1} \quad \text{s.t.} \quad w^T w = 1 = w_0^2 + w_1^2 + \dots + w_{10}^2 = 0$$

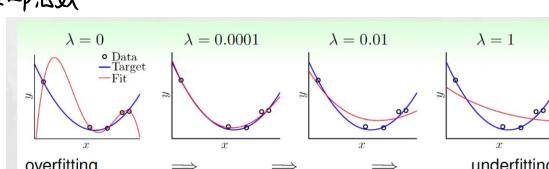
$$\min_{w \in \mathbb{R}^{n+1}} E_{in}(w) \quad \text{W.R.T.} \quad w \in \mathbb{R}^{n+1} \quad \text{s.t.} \quad \sum_{i=0}^{10} (w_i)^2 \leq 1 \quad \Rightarrow \text{此时更灵活, 但增加过拟化风险}$$

$$\min_{w \in \mathbb{R}^{n+1}} E_{in}(w) \quad \text{W.R.T.} \quad w \in \mathbb{R}^{n+1} \quad \text{s.t.} \quad \|w\|_2^2 \leq C \quad \Rightarrow \text{离散, 不易解}$$

$$\min_{w \in \mathbb{R}^{n+1}} E_{in}(w) \quad \text{W.R.T.} \quad w \in \mathbb{R}^{n+1} \quad \text{s.t.} \quad \|w\|_1 \leq C \quad \Rightarrow \text{可解}$$

等价
由此回归

对多项式拟合优化问题
 $\min_{w \in \mathbb{R}^{n+1}} \sum_{i=0}^m (w^T x_i - y_i)^2 + \lambda \|w\|^2$



特征选择: 从特征集中选择相关特征子集的过程

特征: 观测到现象中独立、可测量的属性

相关特征: 对学习任务有用的属性

无关特征: 对学习任务无用的属性

权重系数处理过程

避免维度灾难, 降低学习任务难度, 减少计算开销

特征选择可能降低模型预测能力

前向搜索: 1. 将每一个特征作为一子集, 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

给定数据集 D , 假设 D 中第 i 类样本所占比率为 p_i , ($i = 1, 2, \dots, |Y|$)

其信息熵定义为 $Ent(D) = - \sum_{i=1}^{|Y|} p_i \log(p_i)$

对属性子集 A , 根据其值将 D 分成 k 个子集 $\{D^1, D^2, \dots, D^k\}$, 每个子集在 A 上取值相同, 则属性 A 的信息增益为

$$Gain(A) = Ent(D) - \sum_{j=1}^k \frac{|D^j|}{|D|} Ent(D^j)$$

信息增益 $Gain(A)$ 越大, 特征子集包含的有助于分类的信息越多

特征选择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

$\min_w \sum_{i=0}^m (w^T x_i - y_i)^2 + \lambda \|w\|_1$ 其中 $|w|_1 = \sum_i |w_i|$

用梯度下降法求解 $\sum_{i=0}^m (w^T x_i - y_i)^2$ 的最小值 w , 同时考虑 λ 的最小化

通过过滤、包装或嵌入选择好, 但计算开销大, 有限时间可能无法给出解

选择方法

过滤式

包装式

嵌入式

训练集 D 得到模型期望预测 $y(x) = E_{in}(g(x; D))$

对泛化误差 $E(g; D)$ 分解, 即“偏差-方差分解”

$Var(x) = E_D[(g(x; D) - g(x))^2]$ 差 \Rightarrow 样本数不同数据集产生的方差

$bias^2(x) = [g(x) - y(x)]^2$ 偏差 \Rightarrow 期望输出与真实输出的差别

测试样本 x \rightarrow 真实标注 $y(x)$ \rightarrow 在数据集中挑选 $E = E_{in}(y(x) - y(x))$ 假定噪声期望 $E_{in}(y(x) - y(x))$

对期望泛化误差分解

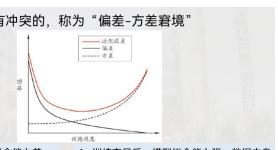
$E(g; D) = E_D[(g(x; D) - y(x))^2] = bias^2(x) + Var(x) + \epsilon^2$

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



◆ 训练不足时, 模型拟合能力差, 数据本身特点会被学到, 发生过拟合, 对数据扰动敏感, 偏差主导泛化误差