步逻辑Linao 2022.10.29

02模型评估与选择

2.1经验误差与过拟合

- 误差
  - 错误率(error rate) —— 分类错误的样本占总样本的比例 —— $E = \frac{a}{m}$
  - 精度(accuracy) —— 精度=1-错误率 —— $1 - \frac{a}{m}$
  - 误差(error) —— 学习器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)/经验误差(empirical error) —— 学习器在训练集上的误差
    - 泛化误差(generalization) —— 学习器在新样本上的误差
- 拟合
  - 过拟合(overfitting)/过配 —— 把训练样本自身的一些特点当作所有潜在样本的一般性质，导致泛化性能下降 —— 原因：学习能力过于强大，把训练样本包含的不太一般性的学到 —— 无法彻底避免，只能缓解
  - 欠拟合(underfitting)/欠配 —— 对训练样本的一般性质尚未学好 —— 原因：学习能力低下

2.2评估方法

- 测试集(testing set) —— 用于测试学习器对新样本的识别能力 —— 测试集应尽可能与训练集互斥
- 测试误差(testing error) —— 一般将其作为泛化误差的近似
- 2.2.1留出法(hold-out) —— 概念：将数据集D划分为两个互斥的集合：训练集S和测试集T
  - 注意
    - 1.训练/测试集的划分要尽可能保持数据分布的一致性，避免因数据划分过程中引入额外的偏差对最终结果产生影响
    - 2.在使用留出法时一般采用若干次随机划分，重复进行实验评估后取平均值作为评估结果
  - 训练集与测试集的比例的多数取为训练效果和评估结果的矛盾 —— 一般将大约2/3~4/5的样本用于训练，剩下的作为测试集
- 2.2.2交叉验证法(cross validation)/k折交验证(k-fold cross validation)
  - 步骤：先将数据集D划分为k个大小相近的互斥子集D，每个子集保持数据分布一致性(分层采样)，然后每次用k-1个子集的并集作为训练集，余下的子集作测试集，进行k次训练和测试，最终返回k个测试结果的均值。
  - 为减小因样本划分不同引入差别，通常要重随机划分预不同的划分重复p次，最终评估结果为p次k ——10次10折交验证
  - 留一法(Leave-One-Out, LOO) —— 数据集中有m个样本，令k=m进行交叉验证
    - 优点：不受随机样本划分方式的影响，评估结果比较准确
    - 缺陷：数据集较大时训练m个模型的计算开销很大
- 2.2.3自助法(bootstrap)
  - 基础：自助采样法(bootstrap sampling)可重复采样/有放回采样
  - 步骤：每次随机从D中挑选一个样本将其拷贝放入 D' 然后再将该样本放回初始数据集D中，使得该样本在下次采样时仍有可能被采到;这个过程重复执行m次后最终到了包含m个样本的数据集D'，初始数据集中约 36.8% 的样本未出现在采样数据集D中,于是可将 D' 用作训练集, D\D' 用作测试集
  - 适用情况：数据集较小、难以有效划分训练集和测试集
  - 优点：能从初始数据集产生多个不同训练集，对集成学习等方法有好处
  - 缺点：却产生的数据集改变了初始数据集的分布，引入估计偏差
- 2.2.4调参与最终模型
  - 参数(parameter)
    - 超参数 —— 算法的参数，数目在10以内，通常由人工设定多个参数候选量产生模型
    - 模型的参数 —— 数目可能很多，通过学习来产生多个候选模型
  - 调参(parameter tuning)
    - 对算法参数进行设定
    - 现实中常对每个参数选定范围和变化步长
  - 验证集(validation set) —— 模型评估与选择中用于评估测试的数据集

2.3性能度量

- 性能度量(performance measure) —— 衡量模型泛化能力的评价标准 —— 回归任务：均方误差(mean squared error)
  - (2.22) $E(f;D) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$
  - (2.3) $E(f;\mathcal{D}) = \int_{x\sim\mathcal{D}}(f(x) - y)^2 p(x)dx$
- 2.3.1错误率与精度
  - 错误率 —— 分类错误的样本数占样本总数的比例
    - (2.4) $E(f;D) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(f(\boldsymbol{x}_i) \neq \boldsymbol{y}_i)$
    - (2.6) $E(f;\mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}} \mathbb{I}(f(\boldsymbol{x}) \neq \boldsymbol{y})p(\boldsymbol{x})d\boldsymbol{x}$
  - 精度 —— 分类正确的样本数占样本总数的比例
    - (2.5) $acc(f;D) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(f(\boldsymbol{x}_i) = \boldsymbol{y}_i) = 1 - E(f;D)$
    - (2.7) $acc(f;\mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}} \mathbb{I}(f(\boldsymbol{x}) = \boldsymbol{y})p(\boldsymbol{x})d\boldsymbol{x} = 1 - E(f;\mathcal{D})$
- 2.3.2查准率、查全率与F1
  - 查准率(precision)/准确率 —— 检索出的信息有多少比例使用户感兴趣的 —— (2.8) $P = \frac{TP}{TP + FP}$
  - 查全率(recall)/召回率 —— 用户感兴趣的信息中有多少被检索出来了 —— (2.9) $R = \frac{TP}{TP + FN}$
  - P-R曲线

图 2.3 P-R曲线及平衡点示意图

  - 真正例(true positive)、假正例(false positive)、真反例(true negative)、假反例(false negative) —— TP+FP+TN+FN=样例总数
  - 混淆矩阵(confusion matrix) ——

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)

  - 查准率和查全率为一对矛盾的度量
    - 增加选取数量能尽可能多地将好瓜选出，但查准率低
    - 只挑选最有把握的瓜来推选出的话中好瓜比例尽可能高，但会漏掉不少好瓜，查全率低。
  - 查准率-查全率曲线(P-R曲线)
    - 1.一个学习器的曲线被另一个完全包住，则后者性能优于前者
    - 2.平衡点：查准率=查全率时的取值 —— (2.10) $F1 = \frac{2 \times P \times R}{P + R} = \underbrace{\frac{2 \times TP}{\text{样例总数} + TP - TN}}$
      - (2.10)推导
        - $\begin{aligned} &\text{令}\ P=R \\ &P=\frac{TP}{TP+FP} \\ &R=\frac{TP}{TP+FN} \\ &\text{化一下}\ FP=FN \\ &\frac{2\times P\times R}{P+R}=\frac{2\times\frac{TP}{TP+FN}\times\frac{TP}{TP+FN}}{\frac{TP}{TP+FN}+\frac{TP}{TP+FN}} \\ &=\frac{2\times\frac{TP}{TP+FN}\times\frac{TP}{TP+FN}}{\frac{2TP+FN+FN}{TP+FN}} \\ &=\frac{2TP}{\text{样例总数}+TP-TN} \end{aligned}$
  - 查准率/查全率偏好 —— (2.11) $F_\beta = \frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}$ —— $\begin{cases} \beta<1 & \text{查准率影响大} \\ \beta=1 & \text{退化为标准 F1} \\ \beta>1 & \text{查全率影响大} \end{cases}$
  - 多分类
    - 先计算查准率和查全率，再计算平均值
      - 宏查准率(macro-P) $macro-P = \frac{1}{n}\sum_{i=1}^{n} P_i$
      - 宏查全率(macro-R) $macro-R = \frac{1}{n}\sum_{i=1}^{n} R_i$
      - 宏F1 $macro-F1 = \frac{2\times macro-P\times macro-R}{macro-P+macro-R}$
    - 先平均再计算
      - 微查准率(micro-P) $micro-P = \frac{TP}{TP+FP}$
      - 微查全率(micro-R) $micro-R = \frac{TP}{TP+FN}$
      - 微F1(micro-F1) $micro-F1 = \frac{2\times micro-P\times micro-R}{micro-P+micro-R}$
- 2.3.3ROC与AUC
  - 分类阈值(threshold) —— 将学习器的预测值与阈值比较，大于则为正类，否则为反类 —— 决定了学习器的泛化能力
  - 截断点(cut point) —— 根据预测结果对测试样本排序，将截断点前作正例，后一部分作反例
    - 重视查准率 —— 在靠后位置截断
    - 重视查全率 —— 在靠前位置截断
  - 接试者工作特征(Receiver Operating Characteristic, ROC)曲线
    - 纵轴：真正例率(True Positive Rate, TPR) $TPR = \frac{TP}{TP+FN}$
    - 横轴：假正例率(False Positive Rate, FPR) $FPR = \frac{FP}{TN+FP}$

  - AUC(Area Under ROC Curve) —— ROC曲线下面积 —— (2.20) $AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i) \cdot (y_i + y_{i+1})$ —— 反映了样本的排序质量，越大越好
  - 排序损失(loss)/排序误差 —— (2.21) $\ell_{rank} = \frac{1}{m^+m^-}\sum_{\boldsymbol{x}^+\in D^+}\sum_{\boldsymbol{x}^-\in D^-}\left(\mathbb{I}(f(\boldsymbol{x}^+) < f(\boldsymbol{x}^-)) + \frac{1}{2}\mathbb{I}(f(\boldsymbol{x}^+) = f(\boldsymbol{x}^-))\right)$
    - 正例预测值小于反例记一个罚分，等于则记0.5个
    - (2.22)推导
      - $\begin{aligned} &\text{令}\ m^+m^-=\sum_{x^+\in D^+}\sum_{x^-\in D^-}\mathbb{I}(f(x)<f(x^-))+\sum_{x^+\in D^+}\sum_{x^-\in D^-}\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-)) \\ &=\sum_{x^+\in D^+}\sum_{x^-\in D^-}\mathbb{I}(f(x^+)<f(x^-))+\frac{1}{2}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\mathbb{I}(f(x^+)=f(x^-)) \\ &=\sum_{x^+\in D^+}\sum_{x^-\in D^-}\mathbb{I}(f(x^+)<f(x^-))+\frac{1}{2}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\mathbb{I}(f(x^+)=f(x^-)) \\ &\text{令}\ 1-\ell_{rank} \\ &\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)<f(x^-))+\mathbb{I}(f(x^+)=f(x^-))\right) \\ &=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)>f(x^-))+\mathbb{I}(f(x^+)=f(x^-))\right) \\ &=1-\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)<f(x^-))+\mathbb{I}(f(x^+)=f(x^-))\right) \\ &=1-\ell_{rank} \end{aligned}$
    - (2.22) $AUC = 1 - \ell_{rank}$
- 2.3.4代价敏感错误率与代价曲线
  - 非均等代价(unequal cost) —— 权衡不同类型错误所造成的不同损失 —— 期望最小化总体代价(total cost)而不是错误数
  - 二分类：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

  - 代价敏感(cost-sensitive)错误率 —— (2.23) $E(f;D;cost) = \frac{1}{m}\left(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}(f(\boldsymbol{x}_i) \neq \boldsymbol{y}_i) \times cost_{01} + \sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}(f(\boldsymbol{x}_i) \neq \boldsymbol{y}_i) \times cost_{10}\right)$
    - (2.23) 演进
      - $\begin{aligned} &\text{令}\ E(f;D;cost)=\frac{1}{m}\left(\sum_{x_i\in D^+}\mathbb{I}(f(x_i)\neq y_i)\times cost_{01}+\sum_{x_i\in D^-}\mathbb{I}(f(x_i)\neq y_i)\times cost_{10}\right) \\ &=\frac{1}{m}\left(\sum_{x_i\in D^+}\mathbb{I}(f(x_i)=y_i)\times cost_{01}+\sum_{x_i\in D^+}\mathbb{I}(f(x_i)\neq y_i)\times cost_{01}+\sum_{x_i\in D^-}\mathbb{I}(f(x_i)=y_i)\times cost_{10}+\sum_{x_i\in D^-}\mathbb{I}(f(x_i)\neq y_i)\times cost_{10}\right) \end{aligned}$
  - 代价曲线(cost curve) —— 反映学习器的期望总体代价
    - 横轴：正例概率代价[0,1] —— (2.24) $P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}}$
    - 纵轴：归一化代价[0,1] —— (2.25) $cost_{norm} = \frac{FNR \times p \times cost_{01} + FPR \times (1-p) \times cost_{10}}{p \times cost_{01} + (1-p) \times cost_{10}}$
      - 期望总体代价(normalization)：将不同变化范围的横纵轴同时调整到范围

2.4比较检验

- 学习器性能的比较
  - 1.测试集上的性能并不一定等于泛化性能
  - 2.不同样例的检测大小的测试集的测试结果也会不同
  - 3.相同的参数在一个测试集运行多次的结果也会不同
  - 统计假设检验(hypothesis test) —— 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，以及这个结论的把握有多大
- 2.4.1假设检验
  - 在包含m个样本的测试集上，泛化错误率为ε，epsilon-δ的学习器被测得测试错误率为ε'hat\ epsilon-δ的概率为
    - (2.26) $P(\hat{\epsilon};\epsilon;\epsilon) = \binom{m}{\hat{\epsilon}\cdot m}\epsilon^{\hat{\epsilon}\cdot m}(1-\epsilon)^{m-\hat{\epsilon}\cdot m}$ —— 概率在二者相等时最大
      - (2.26)推演
        - $\begin{aligned} &\text{令}\ \frac{\partial P(\hat{\epsilon};\epsilon;\epsilon)}{\partial\epsilon}=\frac{\partial\binom{m}{\hat{\epsilon}\cdot m}(1-\epsilon)^{m-\hat{\epsilon}\cdot m}\epsilon^{\hat{\epsilon}\cdot m}}{\partial\epsilon} \\ &=\binom{m}{\hat{\epsilon}\cdot m}\left((1-\epsilon)^{m-\hat{\epsilon}\cdot m}\cdot\frac{\partial\epsilon^{\hat{\epsilon}\cdot m}}{\partial\epsilon}+\epsilon^{\hat{\epsilon}\cdot m}\cdot\frac{\partial(1-\epsilon)^{m-\hat{\epsilon}\cdot m}}{\partial\epsilon}\right) \\ &=\binom{m}{\hat{\epsilon}\cdot m}\left(\epsilon^{\hat{\epsilon}\cdot m}\cdot(m-\hat{\epsilon}\cdot m)\cdot(1-\epsilon)^{m-\hat{\epsilon}\cdot m-1}-\epsilon^{\hat{\epsilon}\cdot m-1}\cdot\hat{\epsilon}\cdot m\cdot(1-\epsilon)^{m-\hat{\epsilon}\cdot m}\right) \\ &=\binom{m}{\hat{\epsilon}\cdot m}\epsilon^{\hat{\epsilon}\cdot m-1}\cdot(1-\epsilon)^{m-\hat{\epsilon}\cdot m-1}\cdot(\epsilon-\hat{\epsilon}\cdot m) \\ &\text{令}\ \frac{\partial P(\hat{\epsilon};\epsilon;\epsilon)}{\partial\epsilon}=0 \\ &\binom{m}{\hat{\epsilon}\cdot m}\epsilon^{\hat{\epsilon}\cdot m-1}\cdot(1-\epsilon)^{m-\hat{\epsilon}\cdot m-1}\cdot(\epsilon-\hat{\epsilon}\cdot m)=0 \\ &\epsilon=\hat{\epsilon}\cdot m \end{aligned}$
  - 自1-α概率内能观测到的最大错误率
    - (2.27) $\bar{\epsilon} = \min\epsilon \quad s.t. \quad \sum_{i=\lceil\epsilon\cdot m\rceil}^{m}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i} < \alpha$
      - (2.27)推导
        - $\begin{aligned} &\text{令}\ \bar{\epsilon}=\min\epsilon\quad s.t.\quad \sum_{i=\lceil\epsilon\cdot m\rceil}^{m}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}<\alpha \\ &\text{令}\ \sum_{i=\lceil\epsilon\cdot m\rceil}^{m}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=\alpha \\ &\text{令}\ \sum_{i=\lceil\epsilon\cdot m\rceil}^{m}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=\sum_{i=0}^{m}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}-\sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i} \\ &\text{令}\ \sum_{i=0}^{m}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=1 \\ &\text{令}\ \sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=1-\alpha \\ &\text{令}\ \sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=\sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i} \\ &\text{令}\ \sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=\sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i} \\ &\text{令}\ \sum_{i=0}^{\lceil\epsilon\cdot m\rceil-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}=1-\alpha \end{aligned}$
  - 若测试错误率小于临界值，则能以1-α的置信度(confidence)认为学习器的泛化错误率不大于ε0, epsilon_0(5)
- 2.4.2交叉验证检验
  - 基本思想：若两个学习器的性能相同，则它们使用相同的训练集、测试集得到的测试错误率应相同
- 2.4.3McNemar检验
  - 对二分类问题使用留出法估计学习器的测试错误率，获得分类结果矩阵
- 2.4.4Friedman检验与Nemenyi检验
  -

2.5偏差与方差

- 偏差-方差分解(bias-variance decomposition) —— 是解释学习算法性能的一种重要工具
- 回归任务
  - 学习算法的期望预测 —— (2.37) $\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x};D)]$
  - 相同样本数的不同训练集产生的方差和噪声
    - (2.38) $var(\boldsymbol{x}) = \mathbb{E}_D\left[(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x}))^2\right]$ —— 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，刻画了数据扰动所造成的影响
    - (2.39) $\epsilon^2 = \mathbb{E}_D\left[(y_D - y)^2\right]$ —— 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，刻画了学习问题本身的难度
  - 期望输出与真实标记的差别，也即偏差(bias) —— (2.40) $bias^2(\boldsymbol{x}) = (\bar{f}(\boldsymbol{x}) - y)^2$ —— 度量了学习算法与期望的真实结果的偏离程度，刻画了学习算法本身的拟合能力
  - 泛化误差可分解为偏差、方差与噪声之和 —— (2.42) $E(f;D) = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \epsilon^2$ —— 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难度共同决定的
- 给定任务较好的泛化性能
  - 较小的偏差 —— 充分拟合数据
  - 较小的方差 —— 使数据扰动的影响小
- 偏差-方差窘境(bias-variance dilemma) —— 偏差与方差是有冲突的 —— 1.训练不足，拟合不强，扰动影响不强，偏差主导——2.训练充足，拟合增强，扰动影响增大，期望预测变深度，过拟合