

模型评估与选择

一、一种训练集一种算法

2.1 经验误差与过拟合

m 为样本数量（如 $m=10000$ 张照片）， Y 为样本正确的结果， Y' 为预测结果，其中有 a 个错了，则error rate错误率是 $E=a/m$, accuracy精度为 $1-E$, error误差为 $|Y-Y'|$

过拟合与欠拟合 过拟合(拟合过头):由两片都有锯齿的叶子，认为没有锯齿的叶子不是树叶，这就是拟合过头了 欠拟合(特征不够):由两片绿叶，认为绿色的都是树叶，这就是欠拟合

解决以上问题一般就是选泛化误差小的模型

2.2 评估方法【训练集与测试集】

1. 泛化能力

即模型对没有见过数据的预测能力，训练集vs预测集

2. training set 训练集


用于估计模型

3. testing set 测试集的保留方法

用于检验魔心复杂程度

留出法: e.g 1.10年中，训练7年数据，预测后3年数据 2.10年中抽出7成的东西做训练集，3成做预测集

交叉验证: k折交叉验证: 一份数据分成k份，每次一个训练集对应一个测试集做出一个结果，最后结果取均值做为最终结果

自助法: 原理: 在包含 m 个样本的数据集 D 中，抽取 m 次样本(放回抽取)形成数据集 D' ,那么一个数据一直没抽取到的概率为 $(1-\frac{1}{m})^m$,取极限 m 趋近于无穷时:  36.8%的数据未出现在 D' 中，于是可以用 D' 做训练集， $D \setminus D'$ 做测试集 适用: 适用于数据集较小，难以划分的时候 缺点: 改变初始数据集分布，容易引起偏差

4. validation set 验证集

- 调参很难，很多参数都是人为规定的
 - 比如三个参数，没个参数有5个候选值，对于一个训练集/测试集就有 $5^3=125$ 个模型需要考察
 - 为了调参，经常会加一个数据集、验证集
 - 训练集训练，验证集看结果，调参，再看验证结果，参数调完，最后再在测试集上看结果
-


2.3 性能度量 performance management

原理:


给定例集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, 其中 y_i 是 x_i 的真实标记。要评估学习器 f 的性能, 就要把预测结果 $f(x)$ 和 y 做比较。


均方误差 **mean squared error** (最常用的性能度量)



若对于每个样例有不同的概率密度 $p(x)$, 则: 

错误率与精度

error rate 错误率:  其中 l 是指成立返回 1, 不成立返回 0

accuracy 精度: 

查准率、查全率与 **F1** 度量


I. 二分类问题


confusion matrix 混淆矩阵 $\begin{pmatrix} \text{true positive 真正例 (TP)} & \text{false positive 假正例 (FP)} \\ \text{true negative 真反例 (TN)} & \text{false negative 假反例 (FN)} \end{pmatrix}$



查准率 (**Precision**): $P = \frac{TP}{TP+FP}$ 查全率 (**Recall**): $R = \frac{TP}{TP+FN}$

一般来说, 查准率高时查全率低, 查全率高时查准率低

P-R 反向变动关系原理 

P-R 曲线 (查准率查全率曲线) 

A 完全包住了 C, 因此 A 优于 C A 和 B 有交点, 不好判断 AB 的高低, 因此有一些综合考虑查准率查全率的性能度量。

最优阈值的确定

- 平衡点 (Break-Event Point, 简称 BEP) $P=R$ 时的取值, 如图 A 优于 B
- F1 度量 $F_1 = \frac{1}{\frac{1}{2}(\frac{1}{P} + \frac{1}{R})}$ 得: $F_1 = \frac{2PR}{P+R} = \frac{2TP}{\text{样例总数} + TP - TN}$

F1 度量的一般形式: $F_\beta = \frac{1}{1 + \beta^2 (\frac{1}{P} + \frac{\beta^2}{R})}$ 得: $F_\beta = \frac{(1 + \beta^2)PR}{(\beta^2 P + R)}$

其中 $\beta > 0$ 度量了查全率对查准率的相对重要性, $\beta \left\{ \begin{aligned} &\text{< 1 --- 查准率有更大影响} \\ &= 1 \text{ --- 退化为 F1} \\ &> 1 \text{ --- 查全率有更大影响} \end{aligned} \right.$

II.n个二分类实现的多分类问题

多分类问题解决方法 \begin{aligned} & \text{直接使用算法} \\ & \text{分解为n个二分类问题:OvsO、OvsR} \end{aligned}

1. 先分别计算，再求平均值: 假设多个二分类得到多组查准率与查全率的组合: $(P_1, R_1), (P_2, R_2), \dots, (P_n, R_n)$. 得到: 宏查准率 $\text{macro-P} = \frac{1}{n} \sum_{i=1}^n P_i$, 宏查全率 $\text{macro-R} = \frac{1}{n} \sum_{i=1}^n R_i$, 带入F1公式得宏F1: $\text{macro-F}_1 = \frac{2 * \text{macro-P} * \text{macro-R}}{\text{macro-P} + \text{macro-R}}$

2. 先平均再计算 先将几个要素求品均值 $\bar{TP}, \bar{FP}, \bar{TN}, \bar{FN}$, 得到: 微查准率 $\text{micro-P} = \frac{\bar{TP}}{\bar{TP} + \bar{FP}}$, 微查全率 $\text{micro-R} = \frac{\bar{TP}}{\bar{TP} + \bar{FN}}$, 带入F1公式得微F1: $\text{micro-F}_1 = \frac{2 * \text{micro-P} * \text{micro-R}}{\text{micro-P} + \text{micro-R}}$