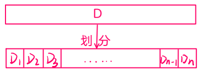
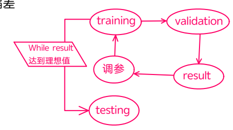
模型评估与选择

一種訓練集一種算法

- 經驗誤差与过拟合
  - 使用模型對m样本進行預測
  - 模型得到一个预测函数 $f(x)=\hat{y}$（预测值），而样本的正确结果为Y
  - 则有错误率 E=a/m（a：错误个数）
  - 精度1-E
  - 误差 $|Y-\hat{y}|$
- 評估方法
  - 泛化能力
    - 模型对没有见过的数据的预测能力
    - 训练集vs测试集
  - 训练集
  - 测试集的保留方法
    - 留出法
      - 很简单的三七分、二八分
      - 注意训练集与测试集同分布
      - 或进行多次随机划分、训练出多个模型，最后取均值
    - 交叉验证法
      - K折交叉验证
      - 缺点：数据量较大时，对算力要求高



      - 将D划分为n个数据集
      - 每次取其中一个作为测试集，其它用于训练
      - 取n个测试结果的平均值作为该模型的结果
    - 自助法
      - 原理：通过自助采样，初始数据集D中约有36.8%的样本未出现在采样数据集D中，可将D'用作训练集，D\D'作为测试集（$\lim_{m\to\infty}(1-\frac{1}{m})^m \to \frac{1}{e} \approx 0.368$）
      - 适用于数据集较小，难以划分时
      - 缺点：改变初始数据集分布，会引入估计偏差
  - 验证集　为了调参，经常在数据集中再划分一部分作为验证集，流程为

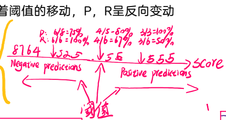


- 性能度量
  - 任务描述
    - 给定样本例集 $D=\{(x_1,y_1),(x_2,y_2),\dots,(x_m,y_m)\}$
    - 其中 y_i 是 x_i 的真实标记。要评价学习器f的性能，就要把学习器预测结果 $f(x)$ 与真实标记y进行比较
  - 均方誤差
    - 回归任务最常用的性能度量是“均方误差”
    - $E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$ [概率分布一致时]
    - 更一般地，对于数据分布 D 和概率密度函数 $p(\cdot)$，均方误差可描述为 $E(f;D)=\int_{x\sim D}(f(x)-y)^2p(x)dx$
  - 錯誤率与精度
    - 错误率：$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)\neq y_i)$　其中 $\mathbb{I}(\cdots)$ 为指示函数，括号内值为真时取1，否则为0
    - 精度：$acc(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i)=y_i)=1-E(f;D)$
  - 混淆矩陣
    - True Positive 真正例
    - False Positive 假正例
    - True Negative 真反例
    - False Negative 假反例

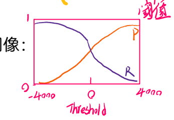
对于二分类

真实值	预测结果：正例	预测结果：反例
正例	TP	FN
反例	FP	TN

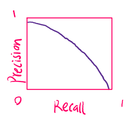
  - 查准率 查全率和F1值
    - 查准率 $P=\frac{TP}{TP+FP}$ 也叫精确率
    - 查全率 $R=\frac{TP}{TP+FN}$ 也叫召回率
    - P-R曲綫
      - 舉例说明：
        - 手写数字识别：使用机器分类是5还是非5（机器对每个样本进行打分，分最高的约容易被判断为正例）
        - 实际上是随着阈值的移动，P，R呈反向变动
      - P-R反向變動关系原理



        - 模型生成了一个值用于判断樣本是真还是假。随著該值的变化，得到的R、R會呈反向變動
      - 閾值：PR圖像：



      - P-R曲线



    - 最优阈值的確定
      1. 使用平衡點（即P=R时）
      2. F1度量
         - 多数时候，往往只需TP的占比情况，我们希望能提高TP的权重，故设定了一个调和平均数 F_1
         - 其中 $\frac{1}{F_1}=\frac{1}{2}(\frac{1}{P}+\frac{1}{R}) \Rightarrow F_1=\frac{2PR}{P+R}=\frac{2TP}{m+TP-TN}$

真实值	预测结果：正例	预测结果：反例
正例	TP	FN
反例	FP	TN

      3. F_β
         - F1是基于查准率和查全率的调和平均定义的
         - F_β 则是加权调和平均：$\frac{1}{F_\beta}=\frac{1}{1+\beta^2}(\frac{1}{P}+\frac{\beta^2}{R})$ $[\frac{1}{F_1}=\frac{1}{2}(\frac{1}{P}+\frac{1}{R})]$
         - 其中β＞0度量了查全率对查准率的相对重要性，β＝1时退化为标准的F1;β＞1时查全率有更大的影响；β＜1时查准率有更大影响
         - 与算数平均 $\frac{P+R}{2}$ 和几何平均 $\sqrt{P^2\times R^2}$ 相比，调和平均更重视较小值
    - n个二分类实现的多分类问题
      - 先分别计算，再求平均值
        - $(P_1,R_1),(P_2,R_2)\dots(P_n,R_n)$
        - marco-P宏查准率：$marco\text{-}P=\frac{1}{n}\sum_{i=1}^{n}P_i$
        - marco-R宏查全率：$marco\text{-}R=\frac{1}{n}\sum_{i=1}^{n}R_i$
        - marco-F1宏F1：$marco\text{-}F_1=\frac{2\times marco\text{-}P\times marco\text{-}R}{marco\text{-}P+marco\text{-}R}$
      - 先求 $\overline{TP}$、$\overline{FP}$、$\overline{TN}$、$\overline{FN}$

n个二分类实现的多分类问题

$marco\text{-}R$ 宏查全率：$marco\text{-}R=\frac{1}{n}\sum_{i=1}^{n}R_i$

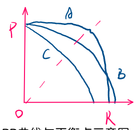
$marco\text{-}F1$ 宏F1：$marco\text{-}F1=\frac{2\times marco\text{-}P\times marco\text{-}R}{marco\text{-}P+marco\text{-}R}$

先平均再计算

- 先求 $\overline{TP}$、$\overline{FP}$、$\overline{TN}$、$\overline{FN}$
- $marco\text{-}P=\frac{\overline{TP}}{\overline{TP}+\overline{FP}}$
- $marco\text{-}R=\frac{\overline{TP}}{\overline{TP}+\overline{FN}}$
- $marco\text{-}F1=\frac{2\times marco\text{-}P\times marco\text{-}R}{marco\text{-}P+marco\text{-}R}$

一种训练多种算法

P-R曲线



PR曲线与平衡点示意图

比较ABC 3个模型的好坏

- 首先可以确定B和A优于C，BA之间有交叉，无法确定
- 对比AB
  - 方法1：比较AB的面积大小，在一定程度上表征了模型的优劣，但是这个值不容易估算
  - 方法2：F_1
  - 方法3：F_β

ROC与AUC

ROC

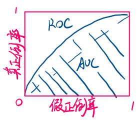
True Positive Rate: $TPR=\frac{TP}{TP+FN}$

False Positive Rate: $FPR=\frac{FP}{TN+FP}$

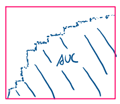
横着比对

真实值	预测值	
	正例	反例
正例	TP	FN
反例	FP	TN

示意图



基于有限样例绘制



AUC

AUC: Area Under ROC Curve 即ROC曲线下的面积

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

排序损失

- 形式化地看，AUC考虑的是样本预测的排序质量，因此它与排序误差有紧密联系。
- 给定m^+个正例和 m^- 个反例，令 D^+,D^-表示正反例集合
- 则排序损失定义为 $\ell_{rank}=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\underbrace{\mathbb{I}(f(x^+)<f(x^-))}_{\text{类似于线性代数中求逆序数}}+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right)$
- 以手写数字为例（机器对样本进行排序，越往左分值越低，反之越高），得到逆序数为4

  8764 ↓ 525 ↓ 56 ↓ 555 → score
  Negative predictions ← various thresholds → Positive predictions

  即 $\ell_{rank}=\frac{4}{6\times6}=\frac{4}{36}$
  $m=12, m^+=6, m^-=6$

真实值	预测值	
	正例	反例
正例	TP	FN
反例	FP	TN

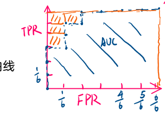
- $AUC=1-\ell_{rank}$ 证明：

  position: 12 11 10 9 8 7 6 5 4 3 2 1 0
  8764 ↓ 525 ↓ 56 ↓ 555 → score
  Negative predictions ← Various thresholds → Positive predictions

  对于书写数字寻找数字5，阈值 Various thresholds 有12种取法

  1. position = 0时，认为样本中不存在正例，$TPR=\frac{0}{6}=0$，$FPR=\frac{0}{6}=0$
  2. … =1，认为只有1個正例 $TPR=\frac{1}{6}$，$FPR=\frac{0}{6}$
  3. … =2，… 2 …，$TPR=\frac{2}{6}$，$FPR=\frac{0}{6}$
  4. … =3 … 3 …，$TPR=\frac{3}{6}$，$FPR=\frac{0}{6}$
  5. … =4 … 4 …，$TPR=\frac{3}{6}$，$FPR=\frac{1}{6}$
  6. … =5 … 5 …，$TPR=\frac{4}{6}$，$FPR=\frac{1}{6}$
  7. … =6 … 6 …，$TPR=\frac{5}{6}$，$FPR=\frac{1}{6}$
  8. … =7，… 7 …，$TPR=\frac{5}{6}$，$FPR=\frac{2}{6}$
  9. … =8，… 8，… $TPR=\frac{6}{6}$，$FPR=\frac{2}{6}$
  10. … =9，… 9，… $TPR=1$，$FPR=\frac{3}{6}$
  11. =10 … … $FPR=\frac{4}{6}$
  12. =11 … … $FPR=\frac{5}{6}$
  13. … =12 … … $FPR=\frac{6}{6}$

  绘制ROC曲线



橙色块共有4个，即 $\frac{4}{36}=\ell_{rank}$
$\therefore AUC=1-\ell_{rank}$

多种训练集一种算法

代价敏感错误率与代价曲线

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

代价敏感错误率为

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

代价曲线

- 使用参数方程，使得图像是线性的
- 目的：对于一个模型，根据p（概率）不同，找到使得代价总期望最小的模型的阈值
- 集合中正例比例，即 $\frac{m^+}{m}$，阈值是预测值，和给定样本中的 $\frac{m^+}{m}$ 无关
- 横轴：归一化的正概率代价期望 $P(+)\,cost=\frac{p\times cost_{10}}{p\times cost_{01}+(1-p)\times cost_{10}}$
- 归一化用来统一横轴纵轴两边的单位
- False Negative Rate & False Positive Rate
- 纵轴：归一化的总代价期望 $cost_{normal}=\frac{[FNR\times p\times cost_{01}+FPR\times(1-p)\times cost_{10}]}{p\times cost_{01}+(1-p)\times cost_{10}}$ ← 总损失期望

TP	FN
FP	TN

偏差与方差

比较检验

测试集上的性能在多大程度上保证真实的性能

问题

1.测试集上的性能与真正的泛化性能未必相同
2.测试集不同反映出来的性能不同
3.机器学习算法本身具有随机性，同一个测试集上多次运行，可能会有不同的结果

一个测试集一种算法

假设测试集有m个样本，其中判断错误的有m'个，可知错误率$\epsilon=\frac{m'}{m}$
在所有的真实数据中，假设其错误率为ϵ_0，若测试集中m=10，则m'=?

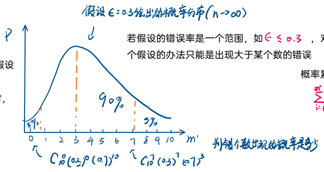
二项分布

常用的离散型随机变量
（1）0-1随机分布
若随机变量X只有两个可能的取值0和1，其概率分布为
$P(X=x_i)=p^{x_i}(1-p)$，$x_i=0,1$
则称X服从0-1分别
（2）二项分布
设事件A在任意一次实验中出现的概率是p（0<p<1）,X表示n重伯努利试验中发生的次数，则X的所有可能取值为01234;··n，且相应概率为
$P(X=k)=C_n^k p^k(1-p)^{n-k}$，$k=0,1,2\cdots n$

假设检验

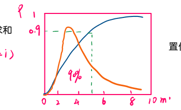
从测试集中找出m'的个数，若与图像出入太大，说明假设错误，可以拒绝假设
若与图像接近，则不能拒绝假设；
具体的，若置信区间为90%,则当测试集中测得的错误样本数量落在置信区间时，则无法拒绝假设

之所以要归到二项分布、正态分布等标准分布，是因为它们的面积好求



若假设的错误率是一个范围，如$\epsilon\leq 0.3$，对于这个假设，实际只设了上限，拒绝这个假设的办法只能是出现大于某个数的错误

概率累加求和
$\sum_{i=1}^{m'}P(n=i)$



置信区间90%

多个测试集一种算法

t检验
- 离散型随机变量的函数分布
- 卡方分布
- t分布
- 正态分布与处理为t分布

假设同一个模型有3个测试集T_1,T_2,T_3，样本m=10，它们分别有各自的错误率$\hat{\epsilon}_1,\hat{\epsilon}_2,\hat{\epsilon}_3$，对它们进行排列组合
则它们每个的概率分布都是

二项分布

各自错误样本数

Test1	0	0	0		10
Test2	0	0	1	···	10
Test3	0	1	1		10

由于是独立事件，每个之间都是用乘法求的概率
得到的错误率ϵ与方差σ共有11^3个，重复的需要合并，得到新的概率分布
又回归到1个测试集1种算法的问题

其他

多个测试集两种算法	交叉验证t检验
一个测试集两种算法	McNemar检验
多个测试集多种算法	Friedman与Nemenyi后续检验