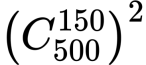
**李韫琪 2020213063004 20智媒**

**2.1数据集包含 1000 个样本，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，试估算共有多少种划分方式.**

留出法：

S：正例350反例350

T：正例150反例150

故共有种划分方式

**2.2 数据集包含100 个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别(训练样本数相同时进行随机猜测），试给出用 10 折交叉验证法和留一法分别对错误率进行评估所得的结果**

10折交叉验证法：由于划分子集与数据集保持数据分布一致，故判断为正例和反例的概率为50%、50%，错误率为50%

留一法：如果使用的训练集缺失一个正例，则训练集中反例样本比正例多一个，留出的样本会被判断为反例，错误率100%，同样的若缺失一个反例，错误率同理可得100%

**2.3 若学习器A的F1值比学习器B高，试析A的 BEP 值是否也比B高.**

无法确定A的 BEP 值是否也比B高，因为F1值的大小与BEP值无明确关系。

BEP是查全率P = 查准率R时的取值；

F1值是2\*P\*R / (P+R) 计算得来；相同BEP值的学习器可能有不同F1值，BEP值与F1值间无必然联系。

故若学习器A的F1值比学习器B高，A的 BEP 值不一定比B高，需要具体情况具体的PR值具体分析

**2.4 试述真正例率(TPR)、假正例率(FPR)与查准率(P）、查全率(R)之间的联系**

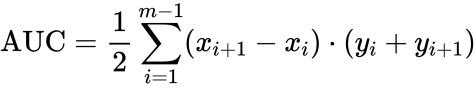
真正例率：真实正例被预测为正例的比例

查全率：真实正例被预测为正例的比例，等于真正例率。

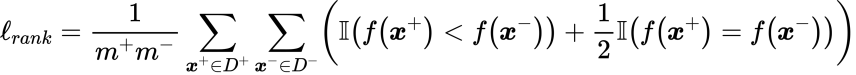
查准率：预测为正例的实例中真实正例的比例，与真正例率、假正例率均没有直接的数值关系。

假正例率：真实反例被预测为正例的比例，与查全率、查准率两者并没有直接的数值关系。

**2.5 试证明式(2.22).**

AUC定义为ROC曲线下的面积，AUC通过对ROC 下面积积分求得，即等价于将小梯形面积累加，结果为yi为上底长，yi+1为下底长，xi+1-xi为高

形式化看，AUC考虑的是样本预测的排序质量，给定m+个正例、m—个反例，令D+和D—分别表示正反例集合，则排序损失可表示为lrank



容易看出，对应是ROC之上的面积，故有52451665731921_.pic

**2.6 试述错误率与 ROC 曲线的联系**

ROC曲线每个点对应了一个TPR与FPR，此时有唯一对应的错误率：

**2.7 试证明任意一条 ROC 曲线都有一条代价曲线与之对应，反之亦然.**

（1）任意一条ROC曲线都有一条代价曲线：

ROC曲线上每一点对应了代价平面上的一条线段，设ROC曲线上点的坐标(FPR, TPR)，则可以相应计算出FNR，然后在代价平面上绘制一条从(0, FPR)到(1, FNR)的线段。故无数个点组成的ROC曲线对应了，无数条代价平面上线段取极限所形成的曲线。

线段下的面积表示该条件下的期望总体代价。如此可以将ROC曲线上的每个点转化为代价平面上的一条线段，然后取所有线段的下界，围成的面积即为所在条件下学习器期望总体代价。

（2）任意一条代价曲线对应一条ROC曲线：

在有限个样本情况下，ROC是一条折线，此时根据代价曲结无法还原ROC曲线。但若是有无限个样本，代价曲线连续，每个点的切线可以求出该点所对应的TPR和FNR，从而得到唯一的ROC曲线。

**2.8 Min-max规范化和z-score规范化是两种常用的规范化方法. 令x和x’分别表示变量在规范化前后的取值，相应的，令xmin和xmax表示规范化前的最小值和最大值，x’min和 x’max表示规范化后的最小值和最大值，x-和6x分别表示规范化前的均值和标准差，则min-max 规范化、z-score规范化分别如式(2.43)和(2.44)所示.试析二者的优缺点.**

（1）min-max 规范化：

优点：简单，并且规范后所有元素均为正值。当加入的新值超过当前最大和最小值范围时，才要重新计算所有之前的结果。

缺点：容易受异常值影响，易受杠杆点和离群点影响

1. z-score规范化：

优点：对异常值敏感度低，不易受离群点影响。

缺点：计算量相对较大。每加入新值都要重新计算所有之前结果。

**2.9 试述x2检验过程．**

（1）提出原假设：

H0：总体X的分布函数为F(x).

如果总体分布为离散型，则假设具体为

H0：总体X的分布律为P{X=xi}=pi， i=1，2，...

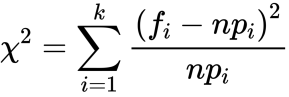
（2）将总体X的取值范围分成k个互不相交的小区间A1，A2，A3，…，Ak，如可取

A1=（a0，a1]，A2=(a1，a2]，...，Ak=(ak-1,ak)，

其中a0可取-∞，ak可取+∞，区间的划分视具体情况而定，但要使每个小区间所含的样本值个数不小于5，而区间个数k不要太大也不要太小。

（3）把落入第i个小区间的Ai的样本值的个数记作fi，成为组频数（真实值），所有组频数之和f1+f2+...+fk等于样本容量n。

（4）当H0为真时，根据所假设的总体理论分布，可算出总体X的值落入第i 个小区间Ai的概率pi，于是，npi就是落入第i个小区间Ai的样本值的理论频数（理论值）。

（5）当H0为真时，n次试验中样本值落入第i个小区间Ai的频率fi/n与概率pi应很接近，当H0不真时，则fi/n与pi相差很大。基于这种思想，皮尔逊引进如下检验统计量 ，在0假设成立的情况下服从自由度为k-1的卡方分布。