习题一

Boardwell, Nanjing University

2017 年 7 月 10 日

Problem 1

若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设，此时的版本空间是什么？在此情形下，试设计一种归纳偏好用于假设选择。

Solution.
版本空间：因为版本空间对应的是所有被预测为真的假设的集合，测试集中所有被预测为真的元素都应该是版本空间的元素。而噪声导致了假设空间中有可能不存在与所有训练样本都一致的假设，故此时假设空间中的任何元素都不被包括，即版本空间是空集。
归纳偏好：首先估计噪声的原因并由此决定噪声的分布（通常情况下系统误差可以认为是正态分布的），然后利用假设检验决定哪一项属性出现噪声（即不符合某种可预测的趋势，分布呈现随机化）的可能性比较大，在得到每种属性的可信度之后，将属性从最可信（出现噪声的可能性最少）到不可信排序，依次赋予不同的权重来得到一个预测的偏好：尽量包含更多一致的属性，倾向于放弃使测试集与版本空间中可信度最低的属性一致。

Problem 2

对于有限样例，请证明

$$\mathrm{AUC} = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right)$$

Proof.
因为*AUC*是*ROC*曲线下方的面积，考虑任取一个在*ROC*曲线上反例，其坐标为（x_i,y_i）,那么y_i恰好是排在其之前的正例所占的比例，所以下方长条的面积$y_i(x_{i+1}-x_i)$也代表了排在其之前的正例所占的比例（$(x_{i+1}-x_i)$相当于一个单位）。如果下一个是真正例，那么反例个数并没有增加，x_i也就没有变，此时还是同一个反例，既不增加计数也不增加面积。在当前坐标为(x_i,y_i)的情况下，如果考虑下一个是正例与反例相等的情况，那么下一个点坐标就是$(x_i+\frac{1}{m^-},y_i+\frac{1}{m^+})$，此时根据梯形面积公式修正可得$S=\frac{1}{2}(y_i+y_{i+1})(x_{i+1}-x_i)$。如果下一个是真正例，那么反例个数并没有增加，$x_i$也就没有变，此时还是同一个反例，既不增加对反例个数累加的过程也不增加面积，与事实相符。所以，一方面，将这些面积累加起来就可以得到AUC在（2.20）的表达式，也就是ROC曲线下方面积。另一方面，

(x_i, y_i)与(x_{i+1}, y_{i+1})之间的下方面积除以归一化系数$\frac{1}{m^+m^-}$以后表示了排在其之前的正例个数（在如果正反例相等则计一半的个数的情况下），将排在所有反例之前的正例个数全部累加起来，根据定义，就等于需要证明公式中的$\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)>f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right)$

因为这两种表达都在描述同一种事物，区别在于一个是以比例为单位，一个是以真实个数为单位，故两者在乘上系数之后是相等的。即证明了原命题：

$$\mathrm{AUC}=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)>f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right)$$

□

Problem 3

在某个西瓜分类任务的验证集中，共有10个示例，其中有3个类别标记为“1”，表示该示例是好瓜；有7个类别标记为“0”，表示该示例不是好瓜。由于学习方法能力有限，我们只能产生在验证集上精度(accuracy)为0.8的分类器。

(a) 如果想要在验证集上得到最佳查准率(precision)，该分类器应该作出何种预测？

此时的查全率(recall)和F1分别是多少？

(b) 如果想要在验证集上得到最佳查全率(recall)，该分类器应该作出何种预测？

此时的查准率(precision)和F1分别是多少？

Solution.

（a）因为精度是分类正确的样本占样本总数的比例，在此题中精度为0.8的意思为必须有两个分类错误的样本。如果这两个错误分类中任一个出现在了7个标记为“0”的样本中，则FP就一定不为0，根据$\mathrm{P}=\frac{TP}{TP+FP}$可知$P$肯定不会达到最大值（也就是1），所以错误分类全部出现在标记为“1”的样本中，此时$FN=2$，$R=\frac{1}{3}$，$F1=\frac{1}{2}$。

（b）同理，若想达到最大的查全率，则错误分类应该全部出现在标记为“0”的样本中。此时$FP=2$，$P=\frac{3}{5}$，$F1=\frac{3}{4}$。

Problem 4

在数据集D_1, D_2, D_3, D_4, D_5运行了A, B, C, D, E五种算法，算法比较序值表如表1所示:

使用Friedman检验($\alpha=0.05$)判断这些算法是否性能都相同。若不相同，进行Nemenyi后续检验($\alpha=0.05$)，并说明性能最好的算法与哪些算法有显著差别。

Solution.

Friedman检验:利用公式(2.34)并令k=N=5，带入r_i(各算法的平均序值)，则可以算出$\tau_{\chi^2}=9.92$,再带入(2.35)的公式则可得$\tau_F=3.94$,而$\alpha=0.05$时F检验的临界值为3.007，所以拒绝

表 1: 算法比较序值表

数据集	算法A	算法B	算法C	算法D	算法E
D_1	2	3	1	5	4
D_2	5	4	2	3	1
D_3	4	5	1	2	3
D_4	2	3	1	5	4
D_5	3	4	1	5	2
平均序值	3.2	3.8	1.2	4	2.8

了原假设，即五种算法的性能有显著差别。

Nemenyi后续检验:利用(2.36)并带入$\alpha = 0.05$时的$q_\alpha = 2.728$算出临界值域$CD = 2.728$，两两比较之后发现性能最好的算法C只与性能最差的算法D有显著差别，和别的算法都没有显著差别。