习题一

141180016,丁俊峰

2017 年 3 月 13 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. 版本空间是与大部分数据一致但与少数数据不同的假设组成的集合。
归纳偏好是选择尽可能多的与训练数据一致的假设。

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. ROC曲线的绘制是将所有样例按照概率预测值降序排列后，从概率最大的点开始计算TPR和FPR，依次绘制。每次当前样例为正样本时TPR增加$\frac{1}{m^+}$，曲线才会爬升1个单位的$\frac{1}{m^+}$。每次当前样例为负样本时FPR增加$\frac{1}{m^-}$，曲线水平增加一个单位$\frac{1}{m^-}$。而排序时若是正样本与负样本预测值相同，则曲线向x，y轴各增加一个单位的$\frac{1}{m^-}$和$\frac{1}{m^+}$。所以要计算曲线下面积AUC，可以将曲线分为m^-个小矩形（梯形）。每个矩形的面积是底×高，底=$\frac{1}{m^-}$。因为只有遇到正样本曲线才会上升一个单位$\frac{1}{m^+}$，所以高="当前矩形对应的负样本之前所有的正样本的个数"$\times\frac{1}{m^+}$，也就是"预测值大于当前负样本的正样本的个数"$\times\frac{1}{m^+}$。而梯形面积是小矩形+小三角形，小矩形的计算方法与前面相同，而小三角形面积为底×高/2，也就是$\frac{1}{2}\times\frac{1}{m^-}$。将所有梯形和矩形的面积加和即能得到

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

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)分类器输出概率值最大的预测为好瓜，其他预测为坏瓜。查准率为1,$F1 = \frac{2\times\frac{2}{3}}{1+\frac{1}{3}} = 1$
(b)全都预测为好瓜。查准率为$\frac{3}{3+7} = 0.3$,$F1 = \frac{2\times0.3}{1+0.3} = 0.46$

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:

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

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. N=5,k=5
根据公式$\mathcal{T}_{x^2} = \frac{k-1}{k} \cdot \frac{12N}{k^2-1} \sum_{i=1}^{k} \left(r_i - \frac{k+1}{2}\right)^2$，得到$\mathcal{T}_{x^2} = 9.92$，
根据公式$\mathcal{T}_F = \frac{(N-1)\mathcal{T}_{x^2}}{N(k-1)-\mathcal{T}_{x^2}}$，得到$\mathcal{T}_F = 3.937$
大于$\alpha = 0.05$时的F检验临界值3.007，故所有算法不相同。
然后使用Nemenyi后续检验，根据公式$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}}$，k=5,N=5时的$q_\alpha = 2.728$，得到CD=2.7。

性能最好的算法D，只有与算法C的平均序值相差为2.8 $>CD$，所以只与算法C有显著差别.