习题一

学号：141130077
作者姓名：邱梓豪

2017 年 3 月 15 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. 此处用于写解答(中英文均可)

若假设空间中不存在与所有训练样本都一致的假设，那么按照对版本空间的定义，此时该训练集对应的版本空间是空集∅。在这种情况下，我觉得应该选择尽可能多的使训练样本得到满足的假设，也就是在回归学习图中这个假设对应的曲线应该穿过最多的点。

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. 此处用于写证明(中英文均可)

设ROC曲线中前一个标记点坐标为（x_i，y_i），则若当前标记点（x_{i+1}，y_{i+1}）为真正例，则坐标为（x，$y+\frac{1}{m^+}$），若为假正例，则坐标为（$x+\frac{1}{m^-}$，y），所以有：

$$x_{i+1} - x_i = \begin{cases} 0, & i+1 \in D^+ \\ \frac{1}{m^-}, & i+1 \in D^- \end{cases} \tag{1}$$

$$y_i + y_{i+1} = \begin{cases} 2y_i + \frac{1}{m^+}, & i+1 \in D^+ \\ 2y_i, & i+1 \in D^- \end{cases} \tag{2}$$

在将上面的式子带入公式（2.20）中，可得：

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)*(y_i+y_{i+1}) = \frac{1}{2}\sum_{i=1,i\in D^-}^{m-1}\frac{1}{m^-}2y_i = \frac{1}{m^-}\sum_{i=1,i\in D^-}^{m-1}y_i \tag{3}$$

对于y_i而言，由定义可知它是由一个个$\frac{1}{m^+}$在$i \in D^+$时积累起来的，因此有：

$$y_i = \sum_{j=1, j \in D^+}^{i} \frac{1}{m^+} \tag{4}$$

将（4）带入（3）中，则有：

$$AUC = \frac{1}{m^-} \sum_{i=1, i \in D^-}^{m-1} \sum_{j=1, j \in D^+}^{i} \frac{1}{m^+} = \frac{1}{m^- m^+} \sum_{i=1, i \in D^-}^{m-1} \sum_{j=1}^{i} \mathbb{I}(j \in D^+) \tag{5}$$

下面对（5）式进行讨论：

(1)若f(j)>f(i)，则$\mathbb{I}$(f(j)>f(i))=1，所以此时$j \in D^+$

(2)若f(j)<f(i)，则$\mathbb{I}$(f(j)>f(i))=0，所以此时j一定$\in D^-$

(3)若f(j)=f(i)，则$\mathbb{I}$(f(j)=f(i))=1，因为此时j可以判定为D^+也可以判定为D^-，所以I($j \in D^+$)=$\frac{1}{2}\mathbb{I}$(f(j)=f(i))

综上所述，可得：

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

□

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. 此处用于写解答(中英文均可)

由于该分类器的精度为0.8，所以分类正确的样本数占样本总数的0.8，在本题中也就是8个，所以有2个被错误地分类了。可以考虑两种极端情况：1、两个好瓜被错误的

标记为坏瓜，这样数据样本中可能最多有5个好瓜；2、两个坏瓜被标记为好瓜，这样数据样本中可能只有1个好瓜。

(a) 由以上分析可知，若想获得最佳查准率，应该仅选取阈值最高的1个瓜为好瓜，此时的查全率R=$\frac{1}{3}$，查准率P=1，所以F1=$\frac{2PR}{P+R}$=$\frac{1}{2}$。

(b) 若想获得最佳查全率，应该仅选取阈值最高的5个瓜为好瓜，此时的查准率P=$\frac{3}{5}$，查全率R=1，所以F1=$\frac{2PR}{P+R}$=$\frac{3}{4}$。

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:

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. 此处用于写解答(中英文均可)

由题目可知N=5，k=5，r_1=3.2，r_2=3.8，r_3=1.2，r_4=4，r_5=2.8，将这些数带入公式（2.34）中，计算出$\tau_{\chi^2} = 9.92$；再带入公式（2.35）中可得$\tau_F = 3.936$，查表2.6可知，它大于$\alpha = 0.05$时的F检验临界值3.007，因此拒绝“所有算法性能相同”这个假设。然后使用Nemenyi后续检验，在表2.7中找到k=5时$q_{0.05}$=2.728，然后根据公式（2.36）算出临界值域CD=2.728，由表中的平均序值可知，算法C和算法D之间有显著差别，其他算法之间没有显著差别。