习题一

141220120, 徐世坚

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. 此时的版本空间可以定义为：能够正确判断样本最多的假设，或者是设定一个阈值，判断正确达到阈值及以上的假设。
此时的归纳偏好可以定为：在一定正确率的前提下，采用奥卡姆剃刀原则，选择最简单的假设。

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. AUC$=\frac{1}{2}\sum_i^{m-1}(x_{i+1}-x_i)(y_{i+1}+y_i)$
$=\frac{1}{2}\sum_{(x_i,y_i)\in D^-}\frac{1}{m^-}(2y_i)$
$=\frac{1}{m^-}\sum_{(x_i,y_i)\in D^-}y_i$
$=\frac{1}{m^-}\sum_{(x_i,y_i)\in D^-}\frac{k}{m^+}$
$=\frac{1}{m^-m^+}\sum_{x^-\in D^-}$(当前样本之前的真正例个数)
将样例按预测值从大到小排序，因为当前的样本是个反例，所以该样本之前的真正例的个数等于预测值比它大的正例的个数，对应于$\mathbb{I}(f(x^+) > f(x^-))$，当因为可能存在正例的预测值和当前的反例预测值相等，那么在排序中可能排在前也可能排在后，取平均就对应于$\frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))$
$\therefore$上式$=\frac{1}{m^-m^+}\sum_{x^-\in D^-}\sum_{x^+\in D^+}(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-)))$ □

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. $\because$ accuracy=0.8
$\therefore$ 有8个样本分类正确
设$TP = x$, $x \in$ 0,1,2,3; 设$TN = 8 - x$, $FN = 3 - x$, $FP = 7 - (8 - x) = x - 1$.
P=$\frac{TP}{TP+FP} = \frac{x}{2x-1}$,R=$\frac{TP}{TP+FN} = \frac{x}{3}$.
(a) 当x=1,查准率最高，P=1
TP=1, FN=2, TP=0, TN=7
R=$\frac{1}{3}$
F1=$\frac{2\times P\times R}{P+R}$=0.5
(b) 当x=3，查全率最高，R=1
TP=3, FN=0, FP=2, TN=5
p=$\frac{3}{5}$
F1=$\frac{2\times P\times R}{P+R}$=0.75

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. Friedman检验：$\tau_{\chi^2} = \frac{248}{25}$，$\tau_F = 3.937$.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。
Nemenyi检验：$CD = 2.728$, 算法C最好，它与算法D的差距超过临界值域，所以，最好的算法C和算法D有显著差别。