习题一

参考解答

2017 年 3 月 27 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.
此时的版本空间是空集。
归纳偏好：采用与训练样本一致数量最多的假设。（言之有理即可）

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.
考虑ROC曲线的绘制过程，设前一个样例在ROC曲线上的坐标为(x, y)，

1) 若当前样例为真正例，则对应在ROC曲线上的坐标为$(x, y + \frac{1}{m^+})$;

2) 若当前样例为假正例，则对应在ROC曲线上的坐标为$(x + \frac{1}{m^-}, y)$。

由此可知，考虑任何一对正例和负例对，

1) 若其中正例预测值小于反例，则x先增加，y后增加，曲线下方的面积(即AUC)将不会因此而增加；

2) 若其中正例预测值大于反例，则y值会先增加，x后增加，曲线下方的面积(即AUC)将增加一个矩形格子，其面积为$\frac{1}{m^+m^-}$；

3) 若一个正例预测值等于反例，对应标记点x, y坐标值同时增加，曲线下方的面积(即AUC)将增加一个三角形，其面积为$\frac{1}{2}\frac{1}{m^+m^-}$.

考虑所有正例和负例对，AUC的面积即为曲线下方的面积，根据上述情况进行累加，则有

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

□

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.

由精度可知，在验证集上，TP+TN=8，FP+FN=2。

(a) 该分类器应该输出预测1个好瓜。

此时TP=1，FP=0，FN=2，TN=7。P=1，R=1/3=0.33，F1=0.5。

(b) 该分类器应该输出预测5个好瓜。

此时TP=3，FP=2，FN=0，TN=5。R=1，P=0.6，F1=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.

$k = 5, N = 5$。由42页式(2.34)与(2.35)可得：

$\tau_{\chi^2} = 9.92, \tau_F = 3.9365 > 3.007$。因此拒绝“所有算法性能相同”假设。

$CD = 2.728, 1.2 + 2.728 = 3.928 < 4$。因此$C$与$D$有显著区别。$C$与其余算法之间没有显著区别。