习题一

141220132 银琦141220132@smail.nju.edu.cn

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：因为含有噪声，所以对于那些只与极少数样本不一致却与极大多数样本一致的假设的集合作为版本空间。

归纳偏好：如果两个数据的属性越相近，则更倾向于将他们分为同一类。若相同属性出现了两种不同的分类，则认为它属于与他最临近几个数据的属性。

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. 由AUC的定义可知，

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1})$$

所有样例可被划分为两个集合，其中正例$m^+ \in D^+$，反例$m^- \in D^-$.
在求和过程中，

若$m \in D^+$，那么$x_{i+1} - x_i = 0$，所以D^+集合中的元素全部不需要计算；

若$m \in D^-$，那么$x_{i+1} - x_i = \frac{1}{m^-}$，$\frac{y_i+y_{i+1}}{2} = y_i$ 所以AUC可以简化为

$$\mathrm{AUC} = \frac{1}{m^-}\sum_{y\in D^-} y$$

接下来计算AUC即转化为计算y，而y恰是排序在其之前的正例所占的比例，即真正例率，因此有

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

综上

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

□

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. 由题可得，精度为0.8，在课本表2.1中，

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} = 0.8$$

而在题目中TP+FP+TN+FN = 10，所以TP+TN = 8，且TP+FN = 3，FP+TN = 7.

(a)

$$\mathrm{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+\frac{FP}{TP}}(TP \neq 0)$$

要得到最佳查准率，即FP取0，此时得到TN = 7，TP = 1，FN = 2，计算可得：

$$\mathrm{Recall} = \frac{TP}{TP+FN} = \frac{1}{3}$$

$$\mathrm{F1} = \frac{2\times P\times R}{P+R} = \frac{2\times 1\times \frac{1}{3}}{1+\frac{1}{3}} = \frac{1}{2}$$

即此时的查全率为$\frac{1}{3}$，$F1 = \frac{1}{2}$.预测结果如下：

真实情况	预测结果	
	正例	反例
正例	1	2
反例	0	7

(b)

$$\mathrm{Recall} = \frac{TP}{TP+FN} = \frac{TP}{3}$$

要得到最佳查全率，即TP取3，此时得到FN = 0，TN = 5，FP = 2，计算可得：

$$\mathrm{Precision} = \frac{TP}{TP+FP} = \frac{3}{5}$$

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} = \frac{2 \times 1 \times \frac{3}{5}}{1 + \frac{3}{5}} = \frac{3}{4}$$

即此时的查全率为$\frac{3}{5}$，$F1 = \frac{3}{4}$.预测结果如下：

真实情况	预测结果	
	正例	反例
正例	3	0
反例	2	5

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:

Table 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. 根据式子

$$\tau_{x^2} = \frac{12N}{k(k+1)} \left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4} \right)$$

及式子

$$\tau_F = \frac{(N-1)\tau_{x^2}}{N(k-1) - \tau_{x^2}}$$

计算出$\tau_F = 9.92$，由课本表2.6可得，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，在课本表2.7中找到$k = 5$时$q_{0.05} = 2.728$，根据式子

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

计算出临界值域$CD = 2.728$，由Table1中的平均序值可知，算法之间的差距如下表：

	算法A	算法B	算法C	算法D	算法E
算法A	-	0.6	2	0.8	0.4
算法B	0.6	-	2.6	0.2	1
算法C	2	2.6	-	2.8	1.6
算法D	0.8	0.2	2.8	-	1.2
算法E	0.4	1	1.6	1.2	-

由表格可知，只有算法C和算法E的差距超过临界值域，因此检验结果认为算法C与算法E的性能显著不同，而算法A、算法B、算法C、算法D之间以及算法A、算法B、算法D、算法E之间的性能没有显著差别。