习题一

151220023, 段建辉

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. 此时的版本空间应该依旧为现有的假设集合。因为最后的模型都是服务于生活，所以假设空间不应该因为噪声的影响而改变。在有噪声的时候需要预处理，适当减少参数空间，让参数变得简单一些。对样本空间放宽约束，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，仍将其保留在版本空间中。此时的归纳偏好最好是选取结果更为平滑的曲线，模型复杂度不要特别高，因为有噪声，所以需要防止出现过拟合的情况。

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. 当有多个测试样本的score相等的时候，我们调整一下阈值，得到的不是曲线一个阶梯往上或者往右的延展，而是斜着向上形成一个梯形。此时，我们就需要计算这个梯形的面积。因此将梯形分为两部分：

Part 1.两相邻坐标点之间以靠前坐标纵坐标为高的矩形

Part 2.矩形以上曲线以下的三角形（因为曲线是折线连接）

计算Part 1:

$$S_1 = \Delta x * y_1$$

上下同乘 m^+m^- 之后得到(因FPR乘m^-之后结果为FP，TPR乘m^+后为TP)：

$$S_1 = \frac{\Delta FP * TP_1}{m^+m^-}$$

其中$\Delta FP * TP_1$表示新增正负样本正、反例对数目，因为前一个的正样本数TP_1和本次新增的负样本数ΔFP组成的正、反例对都满足正样本的判正率高于负样本,权重为1，即：

$$S_1 = \frac{1}{m^+m^-}\left(\mathbb{I}(f(x^+) > f(x^-))\right)$$

计算Part 2:

$$S_2 = \frac{1}{2}\Delta x * \Delta y$$

上下同乘 m^+m^- 之后得到(因FPR乘m^-之后结果为FP，TPR乘m^+后为TP)：

$$S_2 = \frac{1}{2}\frac{\Delta FP * \Delta TP}{m^+m^-}$$

其中$\Delta FP * \Delta TP$表示此次新增的同判正率的正负样本正、反例对数目，因为二者判正率相同，刚好增加的数目相同，因此权重刚好为0.5即：

$$S_2 = \frac{1}{m^+m^-}\left(\frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right)$$

由此可知每一个梯形面积都表示此时正负样本对满足正样本判正率大于等于负样本的加权计数值占全体正负样本的占比。从S_0累计到最后一个S_n整体表示样本整体满足条件的正负样本的占比，进行求和即为：

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

此时等于AUC的面积计算值。

□

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) 根据查准率的计算公式可得预测：TP = 1, FN = 2, FP = 0, TN = 7，精度为0.8
此时为最佳查准率 P = $\frac{1}{1+0} = 1$
此时的查全率为 R = $\frac{1}{1+2} = \frac{1}{3} = 33.3\%$
此时的 F1 = $\frac{2*P*R}{P+R} = \frac{2*1*\frac{1}{3}}{1+\frac{1}{3}} = \frac{1}{2}$

(b) 根据查全率的计算公式可得预测：TP = 3, FN = 0, FP = 2, TN = 5，精度为0.8
此时为最佳查全率为 R = $\frac{3}{3+0} = 1$
此时的查准率 P = $\frac{3}{3+2} = 60.0\%$
此时的 F1 = $\frac{2*P*R}{P+R} = \frac{2*0.6*1}{0.6+1} = \frac{3}{4}$

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. 此题中拥有5个数据集，5种算法，因此 N = 5, k = 5
根据公式(2.34)计算得：

$$\tau_{x^2} = 9.920$$

根据公式(2.35)计算得：

$$\tau_F = 3.937$$

由表2.6可以得知，它大于$\alpha = 0.5$时的F检验临界值为3.007，因此拒绝“所有算法性能都相同”这个假设。
使用Nemenyi后续检验，在表2.7中找到 k = 5时的 $q_0.05 = 2.728$,根据式(3.36)计算出临街值域：

$$CD = 2.728$$

由表中的平均序值可以知道，算法A与算法B、C、D、E差距没超过临界值，算法B与算法C、D、E之间差距也没有超过临界值，算法D与算法E，算法C与算法E之间差距也没有超过临界值。但是算法C与算法D之间差距超过了临界值，则算法C与算法D的性能有显著差别。