习题一

151242041, 王昊庭

2017 年 3 月 12 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. 版本空间是指与假设空间中与训练集完全一致的假设的集合, 因此版本空间为空。此时可以选择一个尽可能符合最多训练样本的假设。

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. 设所有的 $f(x)$ 的值构成 n 元集合, 这些值是本质不同的截断点。将它们从大到小标记为 $p_i, i = 1, 2, \ldots, n$。对于某个截断点 p_i,

$$\begin{aligned}
\mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
&= \frac{1}{m^+}\sum_{x^+\in D^+}\mathbb{I}(f(x^+) > p_i) + \frac{1}{2}\sum_{x^+\in D^+}\mathbb{I}(f(x^+) = p_i)\,, \\
\mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \\
&= \frac{1}{m^-}\sum_{x^-\in D^-}\mathbb{I}(f(x^-) > p_i) + \frac{1}{2}\sum_{x^-\in D^+}\mathbb{I}(f(x^-) = p_i)\,.
\end{aligned}$$

当截断点从 p_i 变化到 p_{i-1} 时, FPR 的变化值为

$$\Delta\mathrm{FPR} = \frac{1}{m^-}\sum_{x^-\in D^-}\mathbb{I}(p_{i-1} < f(x^-) < p_i) + \frac{1}{2}\sum_{x^-\in D^+}\left(\mathbb{I}(f(x^-) = p_i) + \mathbb{I}(f(x^-) = p_{i-1})\right)\,.$$

这一区间上对 AUC 的贡献为,

$$
\begin{aligned}
\mathrm{AUC}_i &= \Delta\mathrm{FPR} \cdot \mathrm{TPR} \\
&= \left(\frac{1}{m^-} \sum_{x^- \in D^-} \mathbb{I}(p_{i-1} < f(x^-) < p_i) + \frac{1}{2} \sum_{x^- \in D^+} \left(\mathbb{I}(f(x^-) = p_i) + \mathbb{I}(f(x^-) = p_{i-1}) \right) \right) \\
&\left(\frac{1}{m^+} \sum_{x^+ \in D^+} \mathbb{I}(f(x^+) \geq p_i) \right) \\
&= \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \mathbb{I}(p_{i-1} < f(x^-) < p_i) \mathbb{I}(f(x^+) \geq p_i) \\
&+ \left(\mathbb{I}(f(x^-) = p_i) + \mathbb{I}(f(x^-) = p_{i-1}) \right) \mathbb{I}(f(x^+) \geq p_i) \\
&= \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \mathbb{I}(f(x^+) > f(x^-)) \mathbb{I}(f(x^-) > p_{i-1}) \\
&+ \frac{1}{2} \mathbb{I}(f(x^+) = p_i) \mathbb{I}(f(x^-) = p_{i-1}) \,.
\end{aligned}
$$

将截断点取遍 p_i，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. (a) 因为模型在测试集上的准确率是 0.8，因此有两个瓜被错误分类。由于有 3 个正例，最坏情况下将 2 个正例标记为反例，因此至少有一个 True Positive。挑选最有把握的那个瓜标记为好瓜，其他标记为坏瓜。此时查准率为 1，查全率为 $\frac{1}{3}$，F1 为 $\frac{1}{2}$。

(b) 应该全部标记为“1”，此时查全率为 1，查准率为 0.3，F1 为 $\frac{6}{13}$。

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.

$$\tau_{\chi^2} = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4}\right) = 9.92.$$

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} = 3.93\,.$$

查表知 $N = 5, k = 5$ 时 F 检验的临界值为 3.007，故这 5 个算法有显著区别。

$$\mathrm{CD} = q_\alpha\sqrt{\frac{k(k+1)}{6N}} = 2.728\,.$$

最好的算法是算法 C,若一个算法显著弱于算法 C,则它的平均序值需要高于 r_C+CD = 3.928，可见只有算法 D 显著弱于算法 C。