机器学习导论
习题一

181220031 李惟康 liwk@smail.nju.edu.cn

2020 年 3 月 13 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是署你名字的工作必须由你完成，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。对于完成作业中有关键作用的公开资料，应予以明显引用；

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，抄袭和被抄袭双方的成绩都将被取消。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在LaTeX模板中第一页填写个人的姓名、学号、邮箱信息；

(2) 本次作业需提交该pdf文件、 问题2问题4可直接运行的源码(两个.py文件)、 作业2用到的数据文件 (为了保证问题2代码可以运行)，将以上四个文件压缩成zip文件后上传，例如181221001.zip;

(3) 未按照要求提交作业，或提交作业格式不正确，将会被扣除部分作业分数；

(4) 本次作业提交截止时间为3月15日23:59:59。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

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.

直觉上，只要训练集样例当中没有出现矛盾(即噪声)：属性值相同但相应标记不相同的情况，则假设空间当中至少能找到一个使得训练错误为0的假设。由于不存在训练错误为0的假设，故此时的版本空间为空集。

归纳假设: 一个简单粗暴的想法是考虑从数据集中剔除所有具有相同属性而不同分类的数据，则对于剩余数据集存在训练误差为0的假设，但同时可能会丢失部分信息。为避免这一情况，可以考虑将相同属性标记不同的样本全部归为这些样本中较多的一类。

Problem 2 [编程]

现有500个测试样例，其对应的真实标记和学习器的输出值如表1所示 (完整数据见data.csv文件)。该任务是一个二分类任务，1表示正例，0表示负例。学习器的输出越接近1表明学习器认为该样例越可能是正例，越接近0表明学习器认为该样例越可能是负例。

表 1: 测试样例表

样本	x_1	x_2	x_3	x_4	x_5	...	x_{496}	x_{497}	x_{498}	x_{499}	x_{500}
标记	1	1	0	0	0	...	0	1	0	1	1
输出值	0.206	0.662	0.219	0.126	0.450	...	0.184	0.505	0.445	0.994	0.602

(1) 请编程绘制P-R曲线

(2) 请编程绘制ROC曲线，并计算AUC

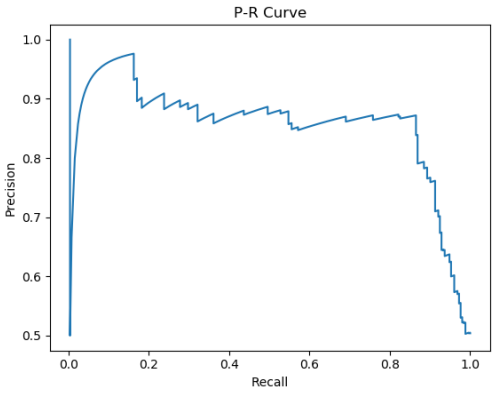

图 1: P-R曲线图

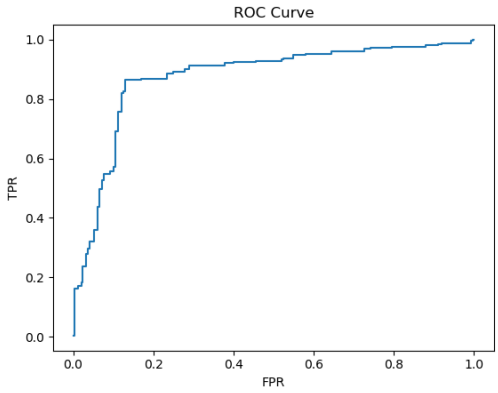

图 2: ROC曲线图

Solution.

(1) P-R曲线见上图1.

首先对原始数据集进行预处理，按输出值降序排序，随后依次把样本作为正例预测，即依次将输出值作为分类的阈值进行处理，同时计算出相应的参数。

```
def gen_params(threshold):
    TN = dataset[(dataset["output"]<threshold) & (dataset["label"]==0)]["Index"].count()
    FP =dataset[(dataset["output"]>=threshold) & (dataset["label"]==0)]["Index"].count()
    FN =dataset[(dataset["output"]<threshold) & (dataset["label"]==1)]["Index"].count()
    TP = dataset[(dataset["output"]>=threshold) & (dataset["label"]==1)]["Index"].count()
    P = TP / (TP + FP)
    R = TP / (TP + FN)
    precision.append(P)
    recall.append(R)
    TPR = TP / (TP + FN)
    FPR = FP / (TN + FP)
    TP_Rate.append(TPR)
    FP_Rate.append(FPR)
```

将每个输出值作为阈值通过上述函数运行之后，即可相应的绘制出 P-R 以及 ROC 曲线图。

(2) ROC曲线见上图2

```
for idx in range(len(output)-1):
    AUC += (FP_Rate[idx+1]-FP_Rate[idx])*(TP_Rate[idx+1]+TP_Rate[idx])/2
```

根据 AUC 估算公式：

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

计算得出：$\mathrm{AUC}=0.874$

Problem 3

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.

参照课本公式(2.21)以及(2.22)：

$$\begin{aligned}
\ell_{rank}=&\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)<f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right)\\
\mathrm{AUC}=1-\ell_{rank}=&\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)\neq f(x^-))+\mathbb{I}(f(x^+)=f(x^-))\right)-\\
&\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)<f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right)\\
=&\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+)>f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-))\right)
\end{aligned}$$

□

Problem 4 [编程]

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2所示:

表 2: 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。本题需编程实现Friedman检验和Nemenyi后续检验。(预计代码行数小于50行)

Solution. 根据课本公式(2.34) (2.35):

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

计算出：$\tau_F = 3.937$，查表2.6可知大于 $\alpha = 0.05$ 时的 F 检验临界值 3.007，因此不符合“所有算法性能相同这一假设”。
之后进行 Nemenyi 后续检验，查表2.7可知 $k = 5$ 时 $q_{0.005} = 2.728$。根据式(2.36):

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

计算出临界值域 $CD = 2.728$，根据表2中的平均序值可知算法C和算法D性能显著不同。