机器学习导论
习题一

181220010, 丁豪, 181220010@smail.nju.edu.cn

2020 年 4 月 20 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. 此处用于写解答 (中英文均可)

由于不存在与所有训练样本都一致的假设，因此版本空间为 $\emptyset$
此时一种可行的归纳偏好是，将精确符合放宽到允许一定差异存在的近似符合，得到新的非空版本空间，并在此使用奥卡姆提到原则选择一个简单的模型

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**??**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

本题需结合关键代码说明思路，并贴上最终绘制的曲线。建议使用 Python 语言编程实现。(预计代码行数小于 100 行)

提示:

- 需要注意数据中存在输出值相同的样例。
- 在 Python 中，数值计算通常使用 Numpy, 表格数据操作通常使用 Pandas, 画图可以使用 Matplotlib (Seaborn), 同学们可以通过上网查找相关资料学习使用这些工具。未来同学们会接触到更多的 Python 扩展库，如集成了众多机器学习方法的 Sklearn, 深度学习工具包 Tensorflow, Pytorch 等。

Solution. 此处用于写解答 (中英文均可)

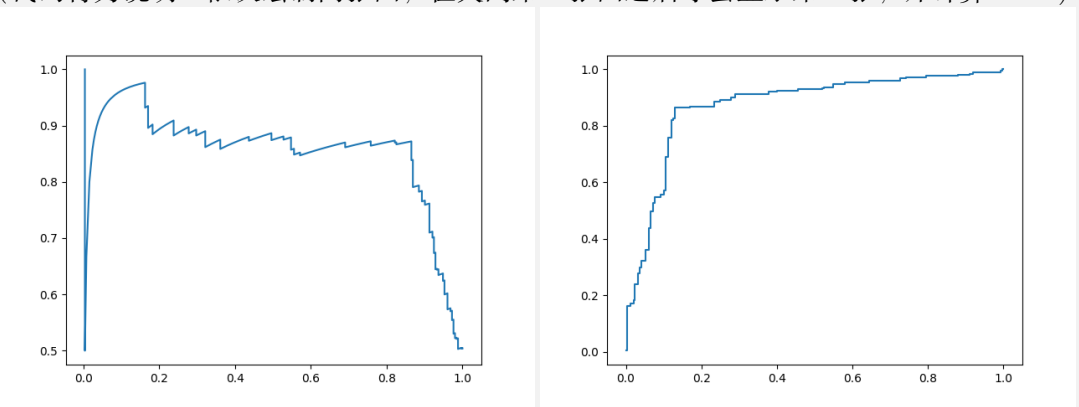
先将结果按照 output 从高到低排序，从第一个开始依次选定阈值为下一个 output 值，每次统计 TP FP TN FN 四个量，然后用公式直接计算所需量，并将得到的结果作为图上的一个点加以存储。具体见代码及注释

```
def draw_pr(csv_data,output_name,label_name):
    csv_data.sort_values(by=output_name, ascending=False, inplace=True)
    output = csv_data[output_name].values
    label = csv_data[label_name].values
    x = list()
    y = list()
    for i in range(len(label)):
        tp=fp=fn=tn=0
        threshold = output[i]
        # 统计4个量
        for j in range(len(label)):
            if output[j]>=threshold:
                if label[j]==0:
                    fp+=1
                else:
                    tp+=1
            else:
                if label[j]==0:
                    tn+=1
                else:
                    fn+=1
        # 在图上建立一个点
        p=tp/(tp+fp)
        r=tp/(tp+fn)
        x.append(r)
        y.append(p)

    plt.plot(x,y)
    plt.show()
```

绘图获得的 P-R 曲线与 ROC 曲线如下图，为了节省时间没有做表头和其他标示，均为默认。其中计算的到的 AUC=0.8737199180747567

(代码行为说明：依次绘制两张图，在关闭第一张图之后才会显示第二张，并计算 AUC)



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. 此处用于写证明 (中英文均可)

记 $\{a_i\}$ 为按照机器学习输出结果从高到低排序的样例集合，与坐标 x 以示区别

$$\begin{aligned}
AUC &= \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_{i+1}+y_i)\\
&= \frac{1}{2m^-}\sum_{i=1}^{m-1}\mathbb{I}(a_i\in D^-)\cdot(2y_i+\mathbb{I}(a_i\in D^+))\\
&= \frac{1}{m^-}\sum_{i=1}^{m-1}\mathbb{I}(a_i\in D^-)\cdot y_i\\
&= \frac{1}{m^-m^+}\sum_{i=1}^{m-1}\mathbb{I}(a_i\in D^-)\cdot\sum_{j=1}^{i-1}\mathbb{I}(a_j\in D^+)\\
&= \frac{1}{m^-m^+}\sum_{i=1}^{m-1}\mathbb{I}(a_i\in D^-)\cdot\sum_{a_j\in D^+}\left(\mathbb{I}(f(a_j)>f(a_i))+\frac{1}{2}\mathbb{I}(f(a_j)=f(a_i))\right)\\
&= \frac{1}{m^+m^-}\sum_{a^+\in D^+}\sum_{a^-\in D^-}\left(\mathbb{I}(f(a^+)>f(a^-))+\frac{1}{2}\mathbb{I}(f(a^+)=f(a^-))\right)
\end{aligned}$$

将上一行中的 a 替换为 x 即可得到题目结论

□

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**??**所示:

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. 此处用于写解答 (中英文均可)
详细说明见代码注释，将公式编程实现得到如下结果

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