机器学习导论
习题一

171860027, 卫歆, 2523058044@qq.com

2020 年 3 月 15 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. 噪声干扰数据的属性值，但不影响它的标签，因此会出现属性值一致但标签不同的情况，即正例反例重合，可以通过去除这样的数据来减小噪声干扰，求出版本空间，但是这样会导致信息丢失。

归纳偏好：选择能够满足最多的训练样本的算法。

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

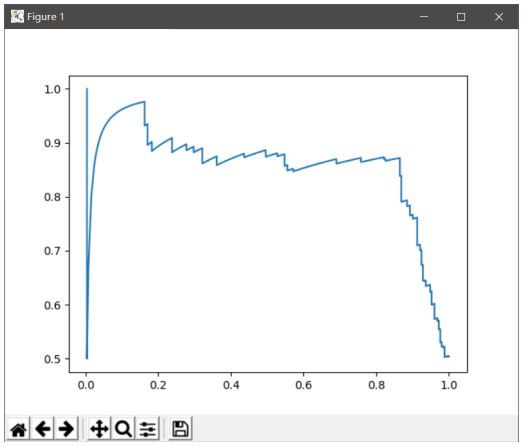
(2) 请编程绘制 ROC 曲线，并计算 AUC

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)

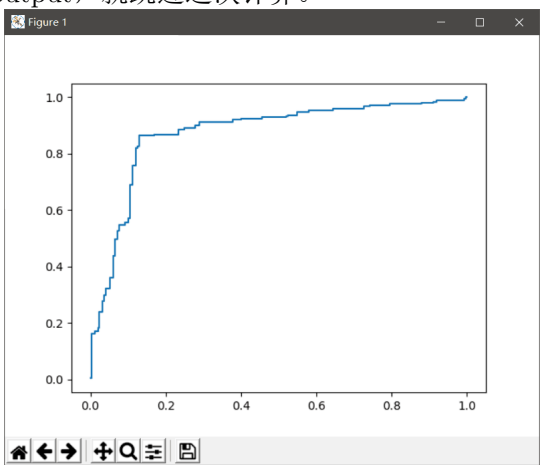
提示:

- 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. 绘制 P-R 曲线首先将数据读入，按照 output 的值降序排序，设定阈值为第一个 output 的值，然后遍历排序完的数组，以当前数据的 output 为阈值，计算数据集的 TP，FP，TN，FN，计算出此时的查准率和查全率存入数组。最后将数组中的点坐标绘制出来。记录一个 lastoutput，如果当前 output 等于 lastoutput，就跳过这次计算。



绘制 ROC 曲线首先将数据读入，按照 output 的值降序排序，设定阈值为第一个 output 的值，然后遍历排序完的数组，以当前数据的 output 为阈值，计算数据集的 TP，FP，TN，FN，计算出此时的 tpr 和 fpr 存入数组。最后将数组中的点坐标绘制出来。记录一个 lastoutput，如果当前 output 等于 lastoutput，就跳过这次计算。



AUC 等于绘制的 tpr-fpr 图线与 x 轴、x=1 所围的面积，用梯形公式累加得 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. AUC 为 $\begin{cases} x= & FP/m- \\ y= & TP/m+ \end{cases}$ 形成的图线与 x 轴、x=1 之间的面积，等价于 1− 图线与 y 轴和 y=1 所围的面积。在计算 AUC 前先将数据集按 output 降序排序，从最大的 output 开始，若当前数据是正例，则 TP+1，若是反例，则 FP+1。下一个数据的 output 小于上一个时，若它为反例，则 FP+1，若它为正例，则 TP+1；output 等于上一个时，若上一个是正例而当前为反例，则 TP+1，FP+1。

将 y 坐标轴放大 m+ 倍，将 x 坐标轴放大 m-倍，此时的图线为 $\begin{cases} x= & FP \\ y= & TP \end{cases}$，将图线与 y

轴和 y=m+ 所围的面积记为 S, 对每个 $x^+ \in D^+$, 它在计算过程中遇到后来的反例时，FP 每增长 1*TP，S 就会增加一个单位面积，TP 每增长 1，S 增加 1/2*FP 个单位面积。可以看做一个上下底平行于 x 轴的直角梯形，它的下底是 TP 没有增加时 FP 的增量，上底是 FP 的全部增量，高为 TP 的增量。因此这个面积可以表示为

$$s = \sum_{x^- \in D^-} \left(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-)) \right)$$

那么 S=$\sum_{x^+ \in D^+} s$.

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

□

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**??**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. 不同。用 Friedman 检验计算出 τ_F=3.874 大于临界值 3.007。因此拒绝所有算法性能相同。用 Nemenyi 检验计算出 CD=2.728。根据平均序值可知，算法 C 与算法 D 的差距超过了 CD，所有认为 C 与 D 有显著差别。