姓名：麻超
学号：201300066

一. (30 points) **概率论基础**

教材附录 C 介绍了常见的概率分布. 给定随机变量 X 的概率密度函数如下,

$$f_X(x) = \begin{cases} \frac{1}{4} & 0 < x < 1; \\ \frac{3}{8} & 3 < x < 5; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

1. 请计算随机变量 X 的累积分布函数 $F_X(x)$;
2. 随机变量 Y 定义为 $Y = 1/X$, 求随机变量 Y 对应的概率密度函数 $f_Y(y)$;
3. 试证明, 对于非负随机变量 Z, 如下两种计算期望的公式是等价的.

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} z f(z) \mathrm{d}z. \tag{2}$$

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} \Pr[Z \geq z] \mathrm{d}z. \tag{3}$$

同时, 请分别利用上述两种期望公式计算随机变量 X 和 Y 的期望, 验证你的结论.

解:

1. 当 $x \leq 0$ 时,$F_X(x) = 0$.

当 $x > 5$ 时,$F_X(x) = 1$.

当 $0 < x < 1$ 时,$F_X(x) = \int_{-\infty}^{x} f(x)\ dx = \int_{-\infty}^{0} 0\ dx + \int_{0}^{x} \frac{1}{4}\ dx = \frac{1}{4}x$.

当 $1 \leq x \leq 3$ 时,$F_X(x) = \int_{-\infty}^{1} \frac{1}{4}\ dx = \frac{1}{4}$

当 $3 < x < 5$ 时,$F_X(x) = \int_{-\infty}^{x} f(x)\ dx = \int_{-\infty}^{0} 0\ dx + \int_{0}^{1} \frac{1}{4}\ dx + \int_{1}^{3} 0\ dx + \int_{3}^{x} \frac{3}{8}\ dx = \frac{3}{8}x - \frac{7}{8}$.

所以所求分布函数是

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{4}x, & 0 < x < 1 \\ \frac{1}{4}, & 1 \leq x \leq 3 \\ \frac{3}{8}x - \frac{7}{8}, & 3 < x < 5 \\ 1, & x \geq 5. \end{cases}$$

2. Y 的取值范围为 $(0,\infty)$, 先求 Y 的分布函数, 由于 $X > 0$, 所以 $Y > 0$. 故当 $y \leq 0$ 时,$F_Y(y) = 0, f_Y(y) = 0$. 当 $y > 0$ 时,

$F_Y(y) = P(\frac{1}{X} \leq y) = P(X \geq \frac{1}{y}) = 1 - P(X < \frac{1}{y}) = 1 - F_X(\frac{1}{y})$

$f_Y(y) = -f_X(\frac{1}{y})(-\frac{1}{y^2}) = f_X(\frac{1}{y})(\frac{1}{y^2})$

所以所求概率密度函数为

$$f_Y(y) = \begin{cases} \frac{1}{4}\frac{1}{y^2}, & y > 1 \\ \frac{3}{8}\frac{1}{y^2}, & \frac{1}{5} < y < \frac{1}{3} \\ 0, & otherwise \end{cases}$$

3. 随机变量 Z 的概率密度为 $f(z)$, 有:

$$\begin{aligned} \mathbb{E}(Z) &= \int_0^\infty zf(z)dz \\ &= \int_0^\infty \int_0^z dxf(z)dz \end{aligned}$$

交换积分次序得到

$$\begin{aligned} \mathbb{E}(Z) &= \int_0^\infty \int_x^\infty f(z)dzdx \\ &= \int_0^\infty P(Z \geq x)dx \end{aligned}$$

由于积分与积分符号无关, 所以有

$\mathbb{E}(Z) = \int_0^\infty P(Z \geq z)dz$

得证.

对于随机变量 X: 第一种方法:

$$\mathbb{E}[X] = \int_0^{\infty} x f_X(x)\ dx = \int_0^1 \frac{1}{4}x\ dx + \int_3^5 \frac{3}{8}x\ dx = \frac{25}{8}$$

第二种方法:

$$\begin{aligned}\mathbb{E}[X] &= \int_0^{\infty} Pr[X \geq x]\ dx \\ &= \int_0^{\infty} (1 - F_X(x))\ dx \\ &= \int_0^1 (1 - \frac{x}{4})\ dx + \int_1^3 \frac{3}{4}\ dx + \int_3^5 (\frac{15}{8} - \frac{3}{8}x)\ dx \\ &= \frac{25}{8}\end{aligned}$$

对于随机变量 Y: 第一种方法:

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^{\infty} y f_Y(y)\ dy \\ &= \int_{\frac{1}{5}}^{\frac{1}{3}} \frac{3}{8}\frac{1}{y}\ dy + \int_1^{\infty} \frac{1}{4}\frac{1}{y}\ dy \\ &= +\infty\end{aligned}$$

第二种方法:

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^{\infty} Pr[Y \geq y]\ dy \\ &= \int_0^{\infty} (1 - F_Y(y))\ dy \\ &= \int_{\frac{1}{5}}^{\frac{1}{3}} (\frac{3}{8y} - \frac{7}{8})\ dy + \int_{\frac{1}{3}}^{1} \frac{1}{4}\ dy + \int_1^{\infty} (\frac{1}{4y})\ dy \\ &= +\infty\end{aligned}$$

二. (40 points) **评估方法**

教材 2.2.3 节描述了自助法 (bootstrapping), 下面考虑将自助法用于对统计量估计这一场景, 并对自助法做进一步分析. 考虑 m 个从分布 $p(x)$ 中独立同分布抽取的（互不相等的）观测值 $x_1, x_2, \ldots, x_m$, $p(x)$ 的均值为 μ, 方差为 σ^2. 通过 m 个样本, 可使用如下方式估计分布的均值

$$\bar{x}_m = \frac{1}{m}\sum_{i=1}^{m} x_i \ , \tag{4}$$

和方差

$$\bar{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m} (x_i - \bar{x}_m)^2 \tag{5}$$

设 $x_1^*, x_2^*, \ldots, x_m^*$ 为通过自助法采样得到的结果, 且

$$\bar{x}_m^* = \frac{1}{m}\sum_{i=1}^{m} x_i^* \ , \tag{6}$$

1. 请证明 $\mathbb{E}[\bar{x}_m] = \mu$ 且 $\mathbb{E}[\bar{\sigma}_m^2] = \sigma^2$;
2. 计算 $var[\bar{x}_m]$;
3. 计算 $\mathbb{E}[\bar{x}_m^* \mid x_1, \ldots, x_m]$ 和 $var[\bar{x}_m^* \mid x_1, \ldots, x_m]$;
4. 计算 $\mathbb{E}[\bar{x}_m^*]$ 和 $var[\bar{x}_m^*]$;
5. 针对上述证明分析自助法和交叉验证法的不同.

解:

1.

$$
\begin{aligned}
\mathbb{E}[\bar{x}_m] &= \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} x_i] \\
&= \frac{1}{m}\mathbb{E}[\sum_{i=1}^{m} x_i] \\
&= \frac{1}{m} m\mu = \mu \\
\mathbb{E}[\bar{\sigma}_m^2] &= \frac{1}{m-1}\mathbb{E}[\sum_{i=1}^{m}(x_i - \bar{x}_m)^2] \\
&= \frac{1}{m-1}\mathbb{E}[\sum_{i=1}^{m} x_i^2 + m\bar{x}_m^2 - 2\sum_{i=1}^{m} x_i \bar{x}_m] \\
&= \frac{1}{m-1}\mathbb{E}[\sum_{i=1}^{m} x_i^2 - m\bar{x}_m^2] \\
\because var[x] &= \mathbb{E}[x^2] - \mathbb{E}^2[x] \\
\therefore \mathbb{E}(x_i^2) &= \mu^2 + \sigma^2 \\
\mathbb{E}(x_m^2) &= var(\bar{x}_m) + \mathbb{E}^2(\bar{x}_m) \\
&\text{由第二问的 } var(\bar{x}_m^2) \text{ 可以得到} \\
\mathbb{E}(\bar{x}_m^2) &= \frac{\sigma^2}{m} + \mu^2 \\
\mathbb{E}(\bar{\sigma}_m^2) &= \frac{1}{m-1}(m\mu^2 + m\sigma^2 - \sigma^2 - m\mu^2) = \sigma^2
\end{aligned}
$$

2.

$$
var(\bar{x}_m) = var(\frac{1}{m}\sum_{i=1}^{m} x_i) = \frac{1}{m^2}var(\sum_{i=1}^{m} x_i) = \frac{\sigma^2}{m}
$$

3. 由独立同分布可得:

$$
\begin{aligned}
\mathbb{E}[\bar{x}_m^* \mid x_1, \ldots, x_m] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[x_i^*] = \bar{x}_m \\
var[\bar{x}_m^* \mid x_1, \ldots, x_m] &= var[\frac{1}{m}\sum_{i=1}^{m}\bar{x}_i^* \mid x_1, \ldots, x_m] \\
&= \frac{1}{m^2}var[\sum_{i=1}^{m}\bar{x}_i^* \mid x_1, \ldots, x_m] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}var[\bar{x}_i^* \mid x_1, \ldots, x_m] \\
&= \frac{1}{m^2}\cdot m \cdot \frac{m-1}{m}\cdot \bar{\sigma}_m^2 \\
&= \frac{m-1}{m^2}\bar{\sigma}_m^2
\end{aligned}
$$

4.

$$
\begin{aligned}
\mathbb{E}[\bar{x}_m^*] &= \mathbb{E}(\mathbb{E}[\bar{x}_m^* \mid x_1, \ldots, x_m]) = \mathbb{E}[\bar{x}_m] = \mu \\
var(\bar{x}_m^*) &= \mathbb{E}[var[\bar{x}_m^* \mid x_1, \ldots, x_m]] + var[\mathbb{E}[\bar{x}_m^* \mid x_1, \ldots, x_m]] \\
&= \mathbb{E}[\frac{m-1}{m^2}\bar{\sigma}_m^2] + var[\bar{x}_m] \\
&= \frac{m-1}{m^2}\sigma^2 + \frac{\sigma^2}{m} \\
&= \frac{2m-1}{m^2}\sigma^2
\end{aligned}
$$

5. 交叉验证法从原数据集中分层采样, 可以认为交叉验证法得到的训练集, 测试集与原数据集分布保持一致, 但是自助法为放回采样, 不能保证分布保持一致.

三. (30 points) **性能度量**

教材 2.3 节介绍了机器学习中常用的性能度量. 假设数据集包含 8 个样例, 其对应的真实标记和学习器的输出值（从大到小排列）如表 1所示. 该任务是一个二分类任务, 标记 1 和 0 表示真实标记为正例或负例. 学

习器的输出值代表学习器认为该样例是正例的概率.

Table 1: 样例表

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
标记	1	1	0	1	0	1	0	0
分类器输出值	0.81	0.74	0.62	0.55	0.44	0.35	0.25	0.21

1. 计算 P-R 曲线每一个端点的坐标并绘图;
2. 计算 ROC 曲线每一个端点的坐标并绘图, 计算 AUC;

解:

1.

可以看出, 该数据集中正例一共有 4 个, 所以查全率 = 正例的样本/4, 查准率 = 正例样本/学习器认为是正例的个数, 以每个样例对应的分类器输出值为阈值, 列出表格如下

样例对应分类器输出值	TP	FP	FN	TN	P	R
x1	1	0	3	4	1.0	0.25
x2	2	0	2	4	1.0	0.5
x3	2	1	2	3	0.67	0.5
x4	3	1	1	3	0.75	0.75
x5	3	2	1	2	0.6	0.75
x6	4	2	0	2	0.67	1.0
x7	4	3	0	1	0.57	1.0
x8	4	4	0	0	0.5	1.0

绘图如下:

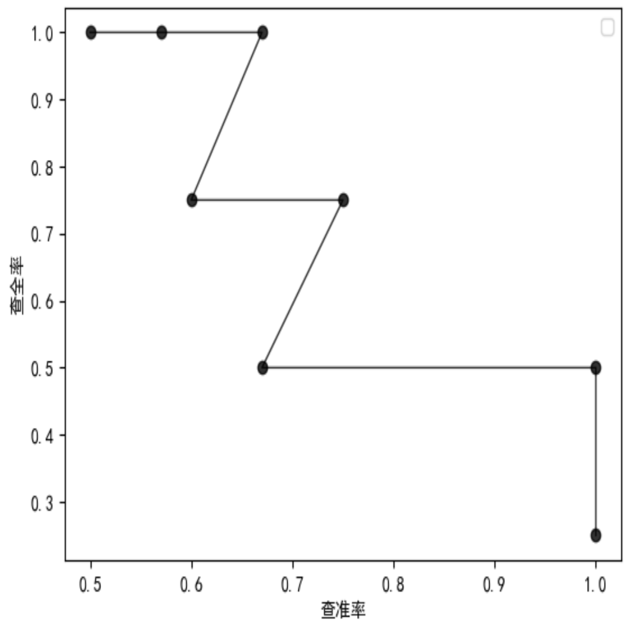

Figure 1: P-R 曲线

2.

根据上一问中得到的结论, 可以得到根据划分不同阈值得到的 TPR 和 FPR 列表为

TPR:[0.25,0.5,0.5,0.75,0.75,1.0,1.0,1.0]

FPR:[0,0,0.25,0.25,0.5,0.5,0.75,1.0]

绘图如下:

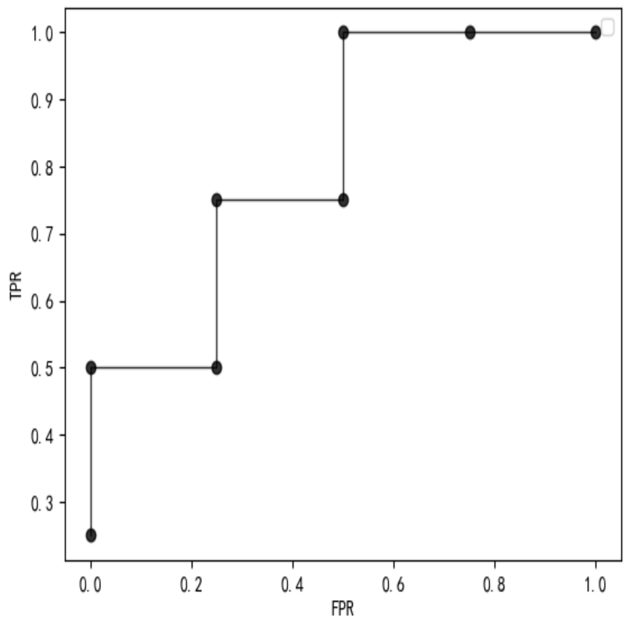

Figure 2: ROC 曲线

计算得到:AUC=0.8125