姓名：杜兴豪
学号：201300096

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.

$$F_X(x) = \int_{-\infty}^{x} f_X(t) dt = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4} & 0 < x < 1 \\ \frac{1}{4} & 1 \leq x \leq 3 \\ \frac{3x-7}{8} & 3 < x < 5 \\ 1 & O.T.W. \end{cases}$$

2.

$$\because F_Y(y) = P(Y \leq y) = P(\frac{1}{X} < y) = P(X > \frac{1}{y}) = 1 - P(X \leq \frac{1}{y})$$

$$\therefore F_Y(y) = 1 - F_X(\frac{1}{y}) = \begin{cases} 0 & \frac{1}{y} \leq 0 \\ \frac{1}{4y} & 0 < \frac{1}{y} < 1 \\ \frac{1}{4} & 1 \leq \frac{1}{y} \leq 3 \\ \frac{3-7y}{8y} & 3 < \frac{1}{y} < 5 \\ 1 & O.T.W. \end{cases}$$

$$\therefore F_Y(y) = \begin{cases} 0 & y \leq 0 \\ \frac{3-7y}{8y} & \frac{1}{5} < y < \frac{1}{3} \\ \frac{1}{4} & \frac{1}{3} \leq y \leq 1 \\ \frac{1}{4y} & y > 1 \\ 1 & O.T.W. \end{cases}$$

$$\therefore f_Y(y) = \begin{cases} -\frac{3}{8y^2} & \frac{1}{5} < y < \frac{1}{3} \\ -\frac{1}{4y^2} & 1 \leq \frac{1}{y} \leq 3 \\ 0 & O.T.W. \end{cases}$$

3.

$$(2) \rightarrow E[Z] = \int_0^\infty zf(z)dz = \int_0^\infty zdF(z) = zF(z)|_0^\infty - \int_0^\infty F(z)dz$$

$$\because F(\infty) = 1, F(0) = 0$$

$$\therefore E[Z] = z|_0^\infty - \int_0^\infty F(z)dz = \int_0^\infty dz - \int_0^\infty F(z)dz$$

$$\therefore E[Z] = \int_0^\infty (1 - F(z))dz \qquad (4)$$

$$(3) \rightarrow E[Z] = \int_0^\infty \Pr[Z > z]dz = \int_0^\infty (1 - \Pr[Z < z])dz$$

$$\because F(z) = \Pr[Z < z]$$

$$\therefore E[Z] = \int_0^\infty (1 - F(z))dz \qquad (5)$$

$$(4) = (5)$$

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.

$$E[\overline{x}_m] = E[\frac{1}{m}\sum_{i=1}^{n} x_i] = \frac{1}{m}\sum_{i=1}^{m} E[x_i] = \frac{1}{m}\sum_{i=1}^{m} \mu = \mu$$

$$\because Var(\overline{x}_m) = \frac{\sum_{i=1}^{m} Var(x_i)}{m^2} = \frac{\sigma^2}{m}$$

$$\therefore E[\overline{x}_m^2] = Var(\overline{x}_m) + E^2[\overline{x}_m] = \frac{\sigma^2}{m} + \mu^2$$

$$\therefore E[\sigma_m^2] = \frac{1}{m-1}\sum_{i=1}^{m} E[(x_i - \overline{x}_m)^2]$$

$$= \frac{1}{m-1}[\sum_{i=1}^{m} E[x_i^2] - 2E\sum_{i=1}^{m} x_i\overline{x}_m + \sum_{i=1}^{m} E[\overline{x}_m^2]]$$
$$= \frac{1}{m-1}[\sum_{i=1}^{m} E[x_i^2] - 2mE[\overline{x}_m^2] + mE[\overline{x}_m^2]]$$
$$= \frac{1}{m-1}[m(\sigma^2+\mu^2) - m(\frac{\sigma^2}{m}+\mu^2)] = \frac{1}{m-1}(m-1)\sigma^2 = \sigma^2$$

2.

Learn from 1. :$Var(\overline{x}_m) = \dfrac{\sigma^2}{m}$

3.

$$E[\overline{x}_m^*|x_1,\ldots,x_m] = \frac{1}{m}\sum_{i=1}^{m} E[x_i^*|x_1,\cdots,x_m] = \frac{\sum\limits_{i=1}^{m} x_i}{m} = \overline{x}_m$$
$$Var[\overline{x}_m^*|x_1,\ldots,x_m] = \frac{1}{m-1}\sum_{i=1}^{m}(x_i^* - E[\overline{x}_m^*|x_1,\ldots,x_m])^2 = \overline{\sigma}_m^2$$

4.

$$E[\overline{x}_m^*] = \frac{1}{m}\sum_{i=1}^{m} E[x_i^*] = \frac{1}{m}\sum_{i=1}^{m}\mu = \mu$$
$$Var[x_i^*] = Var(x) = \sigma^2$$
$$Var[\overline{x}_m^*] = \frac{1}{m^2}Var[\sum_{i=1}^{m} x_i^*] = \frac{\sum_{i=1}^{m} Var(x_i^*)}{m^2} = \frac{\sigma^2}{m}$$

5.
自助法是基于样本总体进行多次有放回抽样的新数据集进行学习，而交叉验证法则通过严格区分训练集和测试集的方式，多次划分取平均值来进行学习。

在以 p(x) 的分布作为数据集进行学习时，交叉验证法的某一次划分中，得到的结果可能是具有较大误差的，然而在进行平均值计算后可以得到一个比较好的水平；而在自助法中，我们计算的模型是通过自助取样得到的，与初始数据集的分布有一定的偏差，导致我们计算结果即使再精确，也有可能难以达到预期的效果。
但是在第三问中，以 x1 到 xm 这个比较小的数据集进行学习时，可以看出计算结果和其真正均值和方差相同，自助法可以发挥出比较好的水平。

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：

$$P=\frac{TP}{TP+FP}, \qquad R=\frac{TP}{TP+FN}$$

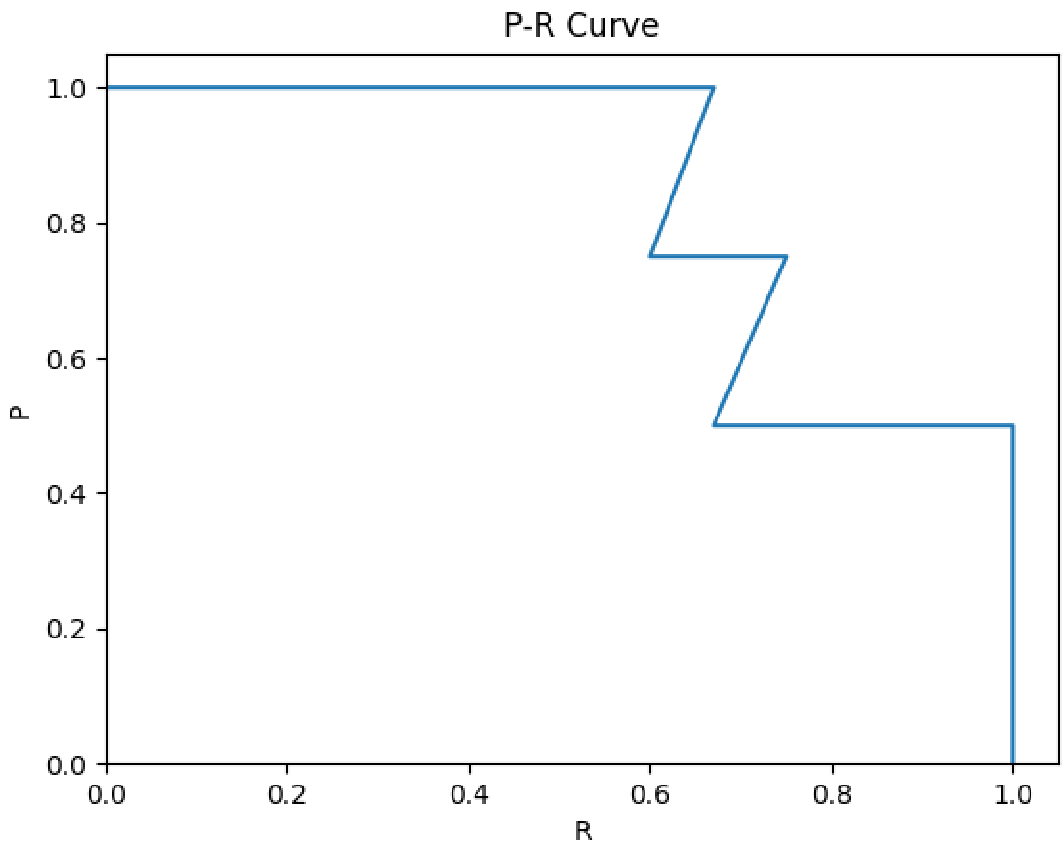
$$\therefore P_1=\frac{1}{1+0}=1, \qquad R_1=\frac{1}{1+3}=\frac{1}{4}=0.25$$

$$P_2=\frac{2}{2+0}=1, \qquad R_2=\frac{2}{2+2}=\frac{1}{2}=0.5$$

$$P_3=\frac{2}{2+1}=\frac{2}{3}=0.67, \qquad R_3=\frac{2}{2+2}=\frac{1}{2}=0.5$$

$$P_4=\frac{3}{3+1}=\frac{3}{4}=0.75, \qquad R_4=\frac{3}{3+1}=\frac{3}{4}=0.75$$

$$
\begin{aligned}
P_5 &= \frac{3}{3+2} = \frac{3}{5} = 0.6, \quad & R_5 &= \frac{3}{3+1} = \frac{3}{4} = 0.75 \\
P_6 &= \frac{4}{4+2} = \frac{2}{3} = 0.67, \quad & R_6 &= \frac{4}{4} = 1 \\
P_7 &= \frac{4}{4+3} = \frac{4}{7} = 0.57, \quad & R_7 &= \frac{4}{4} = 1 \\
R_8 &= \frac{4}{4+4} = \frac{1}{2} = 0.5, \quad & R_8 &= \frac{4}{4} = 1
\end{aligned}
$$



$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{TN+FP}$$

$$\therefore TPR_1 = \frac{1}{4} = 0.25, \qquad FPR_1 = \frac{0}{4} = 0$$

$$TPR_2 = \frac{2}{4} = 0.5, \qquad FPR_2 = \frac{0}{4} = 0$$

$$TPR_3 = \frac{2}{4} = 0.5, \qquad FPR_3 = \frac{1}{4} = 0.25$$

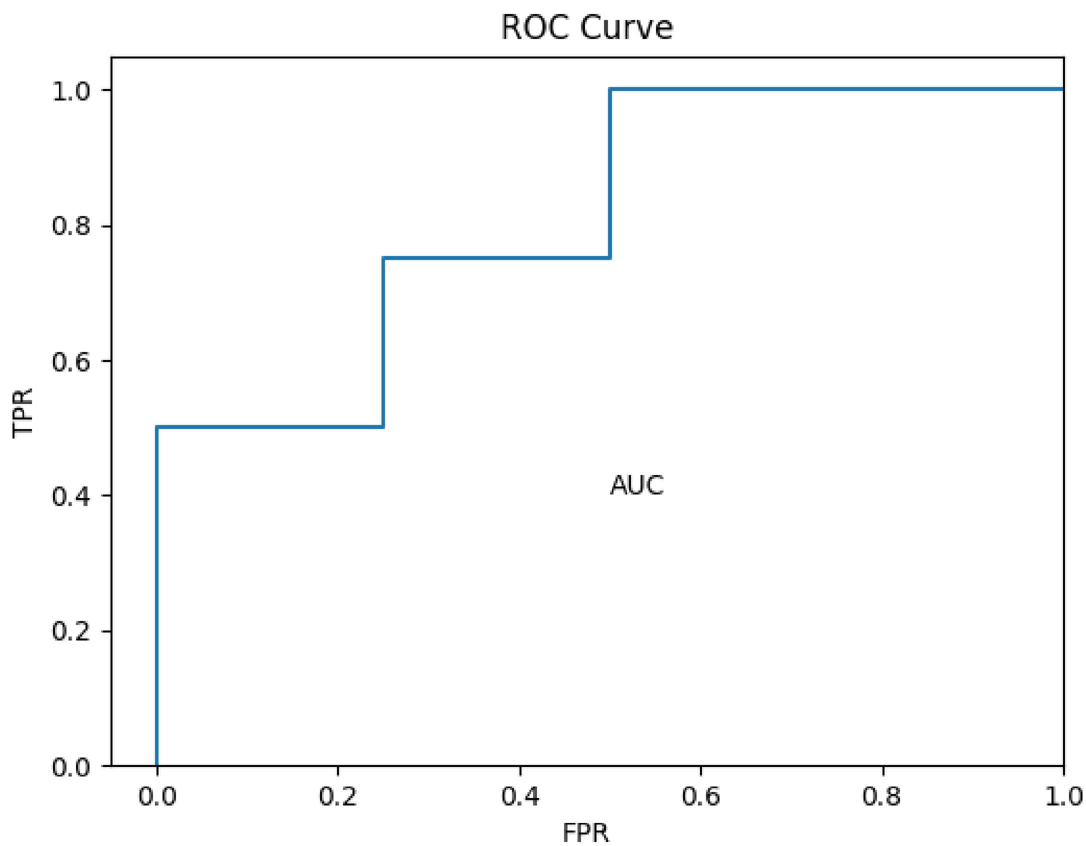
$$TPR_4 = \frac{3}{4} = 0.75, \qquad FPR_4 = \frac{1}{4} = 0.25$$

$$TPR_5 = \frac{3}{4} = 0.75, \qquad FPR_5 = \frac{2}{4} = 0.5$$

$$TPR_6 = \frac{4}{4} = 1, \qquad FPR_6 = 24 = 0.5$$

$$TPR_7 = \frac{4}{4} = 1, \qquad FPR_7 = \frac{3}{4} = 0.75$$

$$TPR_8 = \frac{4}{4} = 1, \qquad FPR_8 = \frac{4}{4} = 1$$



$$AUC = 0.5 \times 0.25 + 0.75 \times 0.25 + 1 \times 0.5 = 0.8125$$