姓名：方盛俊
学号：201300035

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 $0 < x \leq 1$ 时, $F_X(x) = \int_0^x \frac{1}{4} \mathrm{d}x = \frac{x}{4}$

   当 $1 < x \leq 3$ 时, $F_X(x) = F_X(1) = \frac{1}{4}$

   当 $3 < x \leq 5$ 时, $F_X(x) = F_X(3) + \int_3^x \frac{3}{8} \mathrm{d}x = \frac{3x}{8} - \frac{7}{8}$

   当 $x > 5$ 时, $F_X(x) = F_X(5) = 1$

综上即有

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

2. 当 $y\le 0$ 时, 有 $x<0$, 因此 $F_Y(x)=0$

当 $y>0$ 时,

$F_Y(y)=\Pr(Y<y)=\Pr(\frac{1}{X}<y)=\Pr(X>\frac{1}{y})=1-F_X(\frac{1}{y})$

因此

$$F_Y(y)=\begin{cases}0, & y<\frac{1}{5}\\ \frac{15}{8}-\frac{3}{8y}, & \frac{1}{5}\le y<\frac{1}{3}\\ \frac{3}{4}, & \frac{1}{3}\le y<1\\ 1-\frac{1}{4y}, & y\ge 1\end{cases} \tag{5}$$

进而有

$$f_Y(y)=\begin{cases}0, & y<\frac{1}{5}\\ \frac{3}{8y^2}, & \frac{1}{5}\le y<\frac{1}{3}\\ 0, & \frac{1}{3}\le y<1\\ \frac{1}{4y^2}, & y\ge 1\end{cases} \tag{6}$$

3. 首先观察到 $Z=\int_0^Z 1\mathrm{d}t=\int_0^\infty \mathbb{I}(Z>t)\mathrm{d}t$

与公式 $\mathbb{E}[Z]=\int_0^\infty zf(z)\mathrm{d}z$

则有

$$
\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}[\int_0^\infty \mathbb{I}(Z>t)\mathrm{d}t] \\
&= \int_0^\infty f(z)\int_0^\infty \mathbb{I}(z>t)\mathrm{d}t\mathrm{d}z \\
&= \int_0^\infty [\int_0^\infty \mathbb{I}(z>t)f(z)\mathrm{d}z]\mathrm{d}t \\
&= \int_0^\infty [\int_0^t \mathbb{I}(z>t)f(z)\mathrm{d}z + \int_t^\infty \mathbb{I}(z>t)f(z)\mathrm{d}z]\mathrm{d}t \\
&= \int_0^\infty [\int_t^\infty f(z)\mathrm{d}z]\mathrm{d}t \\
&= \int_0^\infty \Pr[Z\geq t]\mathrm{d}t \\
&= \int_0^\infty \Pr[Z\geq z]\mathrm{d}z
\end{aligned}
$$

计算随机变量 X 的期望:

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

$$
\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty \Pr[X\geq x]\mathrm{d}x \\
&= \int_0^1 (1-\frac{x}{4})\mathrm{d}x + \int_1^3 (1-\frac{1}{4})\mathrm{d}x \\
&\quad + \int_3^5 (1-\frac{3x}{8}+\frac{7}{8})\mathrm{d}x + \int_5^\infty (1-1)\mathrm{d}x \\
&= \frac{7}{8}+\frac{3}{2}+\frac{3}{4} \\
&= \frac{25}{8}
\end{aligned}
$$

计算随机变量 Y 的期望:

$$\mathbb{E}[Y] = \int_0^\infty yf_Y(y)\mathrm{d}y = \int_{\frac{1}{5}}^{\frac{1}{3}} \frac{3}{8y^2}\cdot y\mathrm{d}y + \int_1^\infty \frac{1}{4y^2}\cdot y\mathrm{d}y = -\frac{3\ln(3)}{8} + \frac{3\ln(5)}{8} + \infty$$

$$\mathbb{E}[Y] = \int_{\frac{1}{5}}^{\frac{1}{3}} (1 - \frac{15}{8} + \frac{3}{8y})\mathrm{d}y + \int_{\frac{1}{3}}^{1} (1 - \frac{3}{4})\mathrm{d}y + \int_{1}^{\infty} \frac{1}{4y}\mathrm{d}y$$
$$= -\frac{3\ln(3)}{8} - \frac{7}{60} + \frac{3\ln(5)}{8} + \frac{1}{6} + \infty$$

均为不收敛.

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

$$\bar{x}_m = \frac{1}{m}\sum_{i=1}^{m} x_i \ , \tag{7}$$

和方差

$$\bar{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x}_m)^2 \tag{8}$$

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

$$\bar{x}_m^* = \frac{1}{m}\sum_{i=1}^{m} x_i^* \ , \tag{9}$$

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. 对于期望有

$$\mathbb{E}[\bar{x}_m] = \mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} x_i] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[x_i] = \frac{1}{m}\sum_{i=1}^{m}\mu = \mu \tag{10}$$

对于第二个式子有

$$
\begin{aligned}
\mathbb{E}[\bar{\sigma}_m^2] &= \frac{1}{m-1}\mathbb{E}[\sum_{i=1}^{m}(x_i-\bar{x}_m)^2] \\
&= \frac{1}{m-1}\mathbb{E}[\sum_{i=1}^{m}x_i^2 - 2\bar{x}_m\sum_{i=1}^{m}x_i + \sum_{i=1}^{m}\bar{x}_m^2] \\
&= \frac{1}{m-1}\mathbb{E}[\sum_{i=1}^{m}x_i^2 - m\bar{x}_m^2] \\
&= \frac{1}{m-1}(\sum_{i=1}^{m}\mathbb{E}[x_i^2] - m\mathbb{E}[\bar{x}_m^2]) \\
&= \frac{1}{m-1}(\sum_{i=1}^{m}(\mathbb{E}[x_i^2]-\mathbb{E}[x_i]^2) - m(\mathbb{E}[\bar{x}_m^2]-\mathbb{E}[\bar{x}_m]^2)) \\
&= \frac{1}{m-1}(\sum_{i=1}^{m}\operatorname{Var}[x_i] - m\operatorname{Var}[\bar{x}_m]) \\
&= \frac{1}{m-1}(m\cdot\sigma^2 - m\cdot(\frac{1}{m^2}\cdot m\sigma^2)) \\
&= \sigma^2
\end{aligned} \tag{11}
$$

2. 计算方差得

$$
\operatorname{Var}[\bar{x}_m] = \frac{1}{m^2}\cdot m\sigma^2 = \frac{1}{m}\sigma^2 \tag{12}
$$

3. 对于任意一个自助法得到的样本 x_i^* 有

$$
\mathbb{E}[x_i^*|x_1,\cdots,x_m] = \frac{1}{m}\sum_{i=1}^{m}x_i = \bar{x}_m \tag{13}
$$

$$\begin{aligned}\mathrm{Var}[x_i^*|x_1,\cdots,x_m] &= \mathbb{E}[(x_i-\mathbb{E}[x_i^*|x_1,\cdots,x_m])^2|x_1,\cdots,x_m] \\ &= \mathbb{E}[(x_i-\bar{x}_m)^2|x_1,\cdots,x_m] \\ &= \frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x}_m)^2 \\ &= \frac{m-1}{m}\bar{\sigma}_m^2\end{aligned} \tag{14}$$

因此

$$\mathbb{E}[\bar{x}_m^*|x_1,\cdots,x_m] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[x_i^*|x_1,\cdots,x_m] = \bar{x}_m \tag{15}$$

$$\mathrm{Var}[\bar{x}_m^*|x_1,\cdots,x_m] = \frac{1}{m^2}\cdot\sum_{i=1}^{m}\mathrm{Var}[x_i^*|x_1,\cdots,x_m] = \frac{m-1}{m^2}\bar{\sigma}_m^2 \tag{16}$$

4. 对于任意一个自助法得到的样本 x_i^* 有期望

$$\mathbb{E}[x_i^*] = \sum_{i=1}^{m}\frac{1}{m}\mathbb{E}[x_i] = \mu \tag{17}$$

和方差

$$\begin{aligned}\mathrm{Var}[x_i^*] &= \mathbb{E}[x_i^{*2}] - \mathbb{E}[x_i^*]^2 \\ &= \sum_{i=1}^{m}\frac{1}{m}\mathbb{E}[x_i^2] - \mu^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}(\mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2) \\ &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{Var}[x_i] \\ &= \sigma^2\end{aligned} \tag{18}$$

因此

$$\mathbb{E}[\bar{x}_m^*] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[x_i^*] = \mu \tag{19}$$

$$\mathrm{Var}[\bar{x}_m^*] = \frac{1}{m^2}\cdot\sum_{i=1}^{m}\mathrm{Var}[x_i^*] = \frac{1}{m}\sigma^2 \tag{20}$$

5. 交叉验证法, 特别是其中的留一法, 同一个样例不会被抽取多次, 和实际上用 D 训练出的模型较为相似, 所以在大数据集上相对准确, 但是计算开销较大.

   自助法, 虽然期望和方差仍然维持与原数据集相同, 但是会重复抽取相同的样本, 改变了初始数据集的分布, 会引入估计偏差, 一般用于数据集较小的时候.

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 3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. 计算得

Table 2: P-R 曲线

点	1	2	3	4	5	6	7	8	9
P	1.0	1.0	1.0	0.67	0.75	0.6	0.67	0.57	0.5
R	0.0	0.25	0.5	0.5	0.75	0.75	1.0	1.0	1.0

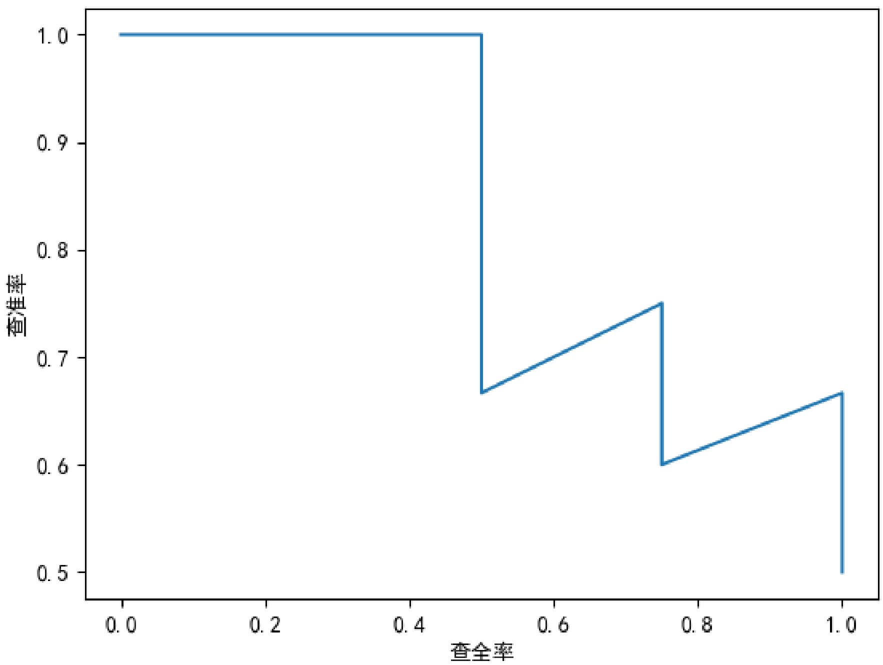

Figure 1: P-R 曲线

2. 计算得

Table 3: P-R 曲线

点	1	2	3	4	5	6	7	8	9
TPR	0.0	0.25	0.5	0.5	0.75	0.75	1.0	1.0	1.0
FPR	0.0	0.0	0.0	0.25	0.25	0.5	0.5	0.75	1.0

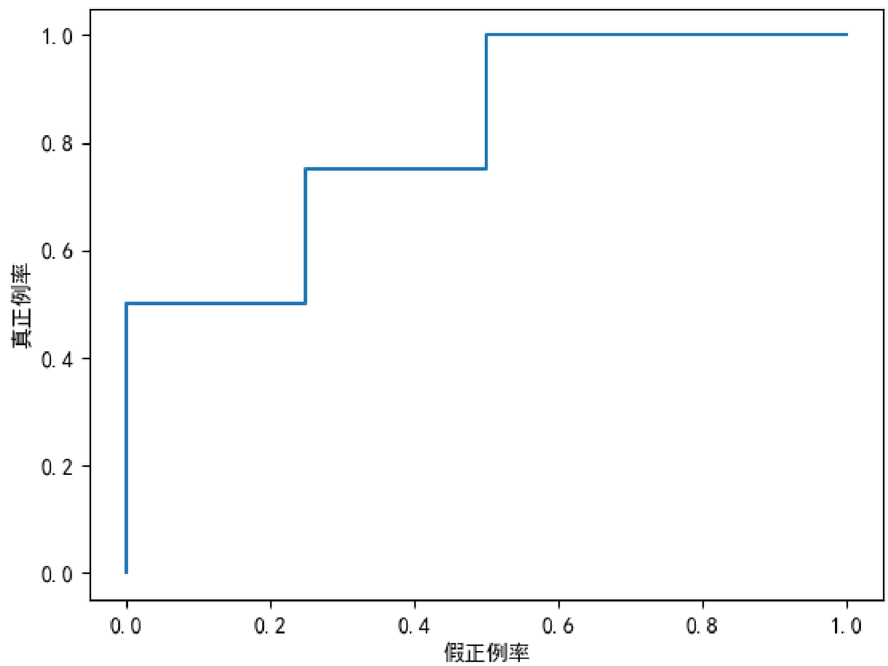

Figure 2: ROC 曲线

AUC 面积为 $S = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i) \cdot (y_i + y_{i+1}) = 0.8125$