Introduction to Big Data Science

Liu Leqi, 12011327

1 Exercise 1

It was done on website.

2 Exercise 2

Since $|x - y| \le |x| + |y|$, we have

$$\begin{aligned}
\sum_{i=1}^{2n-1} |x_{(i)} - c| &= |x_{(n)} - c| + \sum_{i=1}^{n-1} (|x_{(i)} - c| + |x_{(2n-i)} - c|) \\
&\ge |x_{(n)} - c| + \sum_{i=1}^{n-1} |x_{(i)} - x_{(2n-i)}| \\
&\ge \sum_{i=1}^{n-1} |x_{(2n-i)} - x_{(i)}|
\end{aligned} \tag{1}$$

Let $c = x_{(n)}$, then we have

$$\begin{aligned}
\sum_{i=1}^{2n-1} |x_{(i)} - c| &= \sum_{i=1}^{2n-1} |x_{(i)} - x_{(n)}| \\
&= \left[\sum_{i=1}^{n-1} (x_{(n)} - x_{(i)})\right] + (x_{(n)} - x_{(n)}) + \left[\sum_{i=1}^{n-1} (x_{(n+i)} - x_{(n)})\right] \\
&= \sum_{i=1}^{n-1} |x_{(2n-i)} - x_{(i)}|
\end{aligned} \tag{2}$$

So it follows that

$$\min_c \sum_{i=1}^{2n-1} |x_{(i)} - c| \le \sum_{i=1}^{n-1} |x_{(2n-i)} - x_{(i)}| \tag{3}$$

Combining equation (1) and (3), we have

$$\min_c \sum_{i=1}^{2n-1} |x_{(i)} - c| = \sum_{i=1}^{n-1} |x_{(2n-i)} - x_{(i)}| \tag{4}$$

Therefore, the minimum is

$$x_{(n)} = arg \min_c \sum_{i=1}^{2n-1} |x_{(i)} - c|$$

3 Exercise 3

1. E
2. Since for continuous random variables, the probability at a point equals zero.

$$\mathbb{P}(x = 1 | w = 2) = 0$$

3. When w=2,
$$p(1) = 1 - \frac{1}{2} = \frac{1}{2}$$

4 Exercise 4

1.
$$\begin{aligned} E_{px}[E(Y|X)] &= \int_{\mathscr{X}} E(Y|X = x)p_x(x)dx \\ &= \int_{\mathscr{X}} (\int_{\mathscr{Y}} yp_{y|x}(y|x)dy)p_x(x)dx \\ &= \int_{\mathscr{Y}} ydy \int_{\mathscr{X}} p_{y|x}(y|x)p_x(x)dx \\ &= \int_{\mathscr{Y}} yp_y(y)dy \\ &= E_{p_y}Y \end{aligned}$$

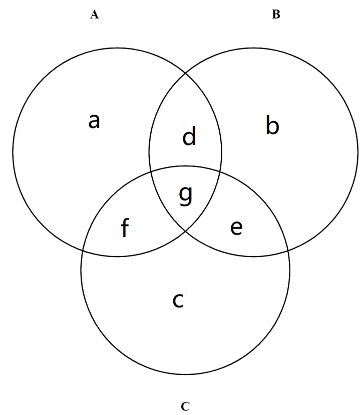
2. If X and Y are independent, then $p_{xy}(x, y) = p_x(x)p_y(y)$, which means
$$E(Y|X = x) = \int_{\mathscr{Y}} yp(y|X = x)dy = \frac{\int_{\mathscr{Y}} yp(x, y)dy}{p_x(x)} = \frac{\int_{\mathscr{Y}} yp_x(x)p_y(y)dy}{p_x(x)} = \int_{\mathscr{Y}} yp_y(y)dy = E(Y)$$

5 Exercise 5

1. Prove positivity:
Since $(A \cap B) \subset A,\ A \subset (A \cup B)$, we have $0 \leq |A \cap B| \leq |A| \leq |A \cup B|$, i.e.
$$\frac{|A \cap B|}{|A \cup B|} \leq 1 \implies 1 - \frac{|A \cap B|}{|A \cup B|} \geq 0 \implies J_\delta(A, B) \geq 0$$
For equality condition, let $J_\delta(A, B) = 0$. Then we have
$$1 - \frac{|A \cap B|}{|A \cup B|} = 0 \implies |A \cap B| = |A \cup B| \implies A = B$$
Conversely, let $A = B$. Then we have
$$|A \cap B| = |A \cup B| \implies 1 - \frac{|A \cap B|}{|A \cup B|} = 0 \implies J_\delta(A, B) = 0$$
Therefore, we can conclude that $J_\delta(A, B) \geq 0$ and the equality holds if and only if $A = B$.

2. Prove symmetry:
By the commutative law of insertion and union, $A \cap B = B \cap A,\ A \cup B = B \cup A$, we have
$$\frac{|A \cap B|}{|A \cup B|} = \frac{|B \cap A|}{|B \cup A|}$$
It follows that
$$J_\delta(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|B \cap A|}{|B \cup A|} = J_\delta(B, A)$$

3. Prove triangle inequality:
Considering the Venn's diagram for three sets A, B and C. We define $a = |(A/B) \cup (A/C)|,\ b = |(B/A) \cup (B/C)|,\ c = |(C/B) \cup (C/A)|,\ g = |A \cap B \cap C|,\ f = |(A \cap C)/B|,\ d = |(A \cap B)/C|,\ e = |(B \cap C)/A|$ By the definition, we have



$$J_\delta(A,B) = 1 - \frac{d+g}{a+f+b+e+d+g} = \frac{a+f+b+e}{a+f+b+e+d+g} \tag{5}$$

$$J_\delta(A,C) = 1 - \frac{f+g}{a+d+c+e+f+g} = \frac{a+d+c+e}{a+d+c+e+f+g} \tag{6}$$

$$J_\delta(B,C) = 1 - \frac{e+g}{c+f+b+d+e+g} = \frac{c+f+b+d}{c+f+b+d+e+g} \tag{7}$$

So, what we need to prove is that

$$\frac{a+b+e+f}{a+b+d+e+f+g} + \frac{b+c+d+f}{b+c+d+e+f+g} - \frac{a+c+d+e}{a+c+d+e+f+g} \geq 0 \tag{8}$$

By simplifying the formula on the left, we have

$$\frac{\begin{aligned} &a^2b + ab^2 + a^2c + 2abc + b^2c + ac^2 + bc^2 + a^2d + 2abd + b^2d + 2acd + 2bcd \\ &+ ad^2 + bd^2 + 2abe + b^2e + 2ace + 2bce + c^2e + ade + 2bde + cde + be^2 + ce^2 \\ &+ a^2f + 4abf + 2b^2f + 4acf + 4bcf + c^2f + 4adf + 5bdf + 3cdf + 2d^2f + 3aef + 5bef \\ &+ 4cef + 4def + 2e^2f + 3af^2 + 4bf^2 + 3cf^2 + 4df^2 + 4ef^2 + 2f^3 + 2abg + 2b^2g + 2acg \\ &+2bcg + adg + 3bdg + 3beg + ceg + 3afg + 6bfg + 3cfg + 4dfg + 4efg + 4f^2g + 2bg^2 + 2fg^2 \end{aligned}}{(a+b+d+e+f+g)(a+c+d+e+f+g)(b+c+d+e+f+g)} \tag{9}$$

Since all the coefficients $a, b, c, d, e, f, g \geq 0$, we can conclude that the formula above is not negative, i.e., the inequality always holds.