Introduction to Big Data Analysis

Homework 1 Reference Answer

2. (15 pts)

By the formula $|x-y| \le |x| + |y|$, we get

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

(The equality holds if and only if c is the median of the given <u>ordered</u> data set.
Notice that $\sum_{i=1}^{n-1} |x_{(i)} - x_{(2n-i)}|$ is a constant for any given data.)
Take $c = x_{(n)}$ (median of the data set), then we have

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

It follows that

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

Combine this with the previous result $\sum_{i=1}^{2n-1} |x_{(i)} - c| \ge \sum_{i=1}^{n-1} |x_{(2n-i)} - x_{(i)}|$, we finally get that

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

and the minimum is taken when $c = x_{(n)}$, i.e.

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

3. (5+10+10 pts)

1. (E)
2. $\mathbb{P}(x = 1|w = 2) = 0$. (There is no probability mass.)
3. when $w = 2$, then

$$p(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1 - x/2, & \text{if } 0 \le x \le 2, \\ 0, & \text{if } 2 < x. \end{cases}$$

   It gives that $p(1) = 1 - 1/2 = 1/2$.

4. (10+10 pts)

(1)

$$\begin{aligned} E_{p_x}[E(Y|X)] &= \int_{\mathcal{X}} E(Y|X = x)p_x(x)dx \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{yp(x, y)}{p_x(x)} \cdot p_x(x)dydx \\ &= \int_{\mathcal{Y}} y \int_{\mathcal{X}} p(x, y)dxdy \\ &= \int_{\mathcal{Y}} yp_y(y)dy = E_{p_y}Y. \end{aligned}$$

(2) If X and Y are independent, then $p(x, y) = p_x(x)p_y(y)$, therefore

$$E(Y|X = x) = \int_{\mathcal{Y}} \frac{yp(x, y)}{p_x(x)} dy = \int_{\mathcal{Y}} yp_y(y)dy = E(Y).$$

5. (10+10+20 pts)

(1) From $(A \cap B) \subset (A \cup B)$, then $|A \cap B| \le |A \cup B|$, therefore $J_\delta(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \ge 0$.

If $J_\delta(A, B) = 0$, then $|A \cap B| = |A \cup B|$, which holds if and only if $A = B$.

(2) Since $A \cap B = B \cap A$, $A \cup B = B \cup A$, we have $J_\delta(A, B) = J_\delta(B, A)$.

(3) First claim that $|A \cap C| \cdot |B \cup C| + |A \cup C| \cdot |B \cap C| \le |C| \cdot (|A| + |B|)$.

Note that

$$\begin{aligned} |A \cap C| \cdot |B \cup C| &= |A \cap C| \cdot (|B| + |C| - |B \cap C|) \\ &= |A \cap C| \cdot (|B| - |B \cap C|) + |C| \cdot |A \cap C| \\ &\le |C| \cdot (|B| - |B \cap C| + |A \cap C|), \end{aligned}$$

by swapping A and B,

$$|A \cup C| \cdot |B \cap C| \leq |C| \cdot (|A| - |A \cap C| + |B \cap C|).$$

Adding up the above two inequality, we obtain

$$|A \cap C| \cdot |B \cup C| + |A \cup C| \cdot |B \cap C| \leq |C| \cdot (|A| + |B|). \tag{1}$$

By setting $A = B$, we get

$$|A \cap C| \cdot |A \cup C| \leq |A| \cdot |C|. \tag{2}$$

To prove $J_\delta(A, B) \leq J_\delta(A, C) + J_\delta(B, C)$, it suffices to show

$$\frac{|A \cap C|}{|A \cup C|} + \frac{|B \cap C|}{|B \cup C|} \leq 1 + \frac{|A \cap B|}{|A \cup B|} = \frac{|A| + |B|}{|A \cup B|}.$$

By applying the inequalities (1) and (2), we have

$$\begin{aligned}
\frac{|A \cap C|}{|A \cup C|} + \frac{|B \cap C|}{|B \cup C|} &= \frac{|A \cap C| \cdot |B \cup C| + |A \cup C| \cdot |B \cap C|}{|A \cup C| \cdot |B \cup C|} \\
&\leq \frac{|C| \cdot (|A| + |B|)}{|A \cup C| \cdot |B \cup C|} \\
&\leq \frac{|C| \cdot (|A| + |B|)}{|(A \cup C) \cap (B \cup C)| \cdot |A \cup B \cup C|} \\
&\leq \frac{|C|}{|(A \cap B) \cup C|} \cdot \frac{|A| + |B|}{|A \cup B|} \\
&\leq \frac{|A| + |B|}{|A \cup B|}.
\end{aligned}$$