# Intro to Big Data Science: Assignment 1 Reference Answer

TAN 12212523@mail.sustech.edu.cn

April 16, 2025

## Exercise 1

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

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

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

(The equality holds if and only if $c$ is the median of the given ordered 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

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

$$= \left[\sum_{i=1}^{n-1} (x_{(n)} - x_{(i)})\right] + (x_{(n)} - x_{(n)}) + \left[\sum_{i=1}^{n-1} (x_{(n+i)} - x_{(n)})\right]$$

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

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.

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

## Exercise 2

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 $p(1) = 1 - 1/2 = 1/2$.

# Exercise 3

1. 

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

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

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

3. The statement can be yielded from

$$\begin{aligned} E[(Y-c)^2|X=x] &= E[(Y^2 - 2cY + c^2)|X=x] \\ &= c^2 - 2E[Y|X=x]c + E[Y^2|X=x], \quad \forall c \in \mathbb{R} \end{aligned}$$

directly.

# Exercise 4

**Proof:**

1. **Positivity**:
Since $|A \triangle B| \geq 0$, it follows directly that $R_\delta(A,B) \geq 0$. Then we need to prove
First, if $A = B$, then $|A \setminus B| = |B \setminus A| = 0$, so $R_\delta(A,B) = 0$.
Conversely, if $R_\delta(A,B) = 0$, then $|A \setminus B| = |B \setminus A| = 0$ since $|A \setminus B| \geq 0$, $|B \setminus A| \geq 0$. Then $A = B$. (The symmetric difference $A \triangle B$ is the set of elements which are in either $A$ or $B$ but not in their intersection, $|A \triangle B| = 0$ implies $A = B$.)

2. **Symmetry**: By the definition of symmetric difference,

$$R_\delta(A,B) = \frac{|A \setminus B|}{|\Omega|} + \frac{|B \setminus A|}{|\Omega|} = \frac{|B \setminus A|}{|\Omega|} + \frac{|A \setminus B|}{|\Omega|} = R_\delta(B,A).$$

3. **Triangle inequality**:
Let $A, B, C \subset \Omega$. We need to show that $R_\delta(A,B) \leq R_\delta(A,C) + R_\delta(C,B)$.
Consider the symmetric differences

$$\begin{aligned} A \triangle B &= (A \setminus B) \cup (B \setminus A), \\ A \triangle C &= (A \setminus C) \cup (C \setminus A), \\ B \triangle C &= (B \setminus C) \cup (C \setminus B). \end{aligned}$$

Notice that

$$\begin{aligned} A \setminus B &= (A \setminus C) \cup (A \cap C \setminus B), \\ B \setminus A &= (B \setminus C) \cup (B \cap C \setminus A). \end{aligned}$$

Therefore

$$\begin{aligned} A \triangle B &= (A \setminus C) \cup (A \cap C \setminus B) \cup (B \setminus C) \cup (B \cap C \setminus A) \\ &\subseteq (A \setminus C) \cup (C \setminus B) \cup (B \setminus C) \cup (C \setminus A) \\ &= (A \triangle C) \cup (B \triangle C), \end{aligned}$$

which implies

$$|A \triangle B| \le |A \triangle C| + |C \triangle B|.$$

Dividing both sides by $|\Omega|$, then we have

$$R_\delta(A,B) = \frac{|A \triangle B|}{|\Omega|} \le \frac{|A \triangle C| + |C \triangle B|}{|\Omega|} = R_\delta(A,C) + R_\delta(C,B).$$

Therefore, the triangle inequality holds for the rand distance $R_\delta$.

$\mathbf{R}_\delta$ is a distance metric.

## remark: Jaccard Distance Properties

(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 $A \cap B = B \cap A$ and $A \cup B = B \cup A$, 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 $B$,

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

Adding up the above two inequality, we obtain

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

By setting $A = B$, we get

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

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

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

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

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