

Intro to Big Data Science: Assignment 1

Due Date: Mar 15, 2024

Exercise 1

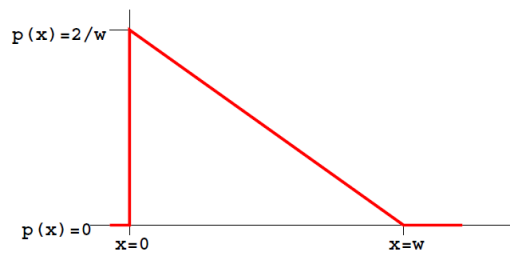
Given the ordered data $\{x_{(i)}\}_{i=1}^{2n-1}$ with increasing order. Show that the median of the data set is equal to the minimizer of the following L^1 minimization problem:

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

Exercise 2

Consider the probability density function (PDF) shown in the following figure and equations:

$$p(x) = \begin{cases} 0, & \text{if } x < 0, \\ \frac{2}{w} - \frac{2x}{w^2}, & \text{if } 0 \leq x \leq w, \\ 0, & \text{if } w < x. \end{cases}$$



1. Which of the following expression is true? (Only one truth.)

$$(A) \ E[X] = \int_{-\infty}^{\infty} (\frac{2}{w} - \frac{2x}{w^2}) dx;$$

$$(B) \ E[X] = \int_{-\infty}^{\infty} x(\frac{2}{w} - \frac{2x}{w^2}) dx;$$

$$(C) \ E[X] = \int_{-\infty}^{\infty} w(\frac{2}{w} - \frac{2x}{w^2}) dx;$$

$$(D) \ E[X] = \int_0^w (\frac{2}{w} - \frac{2x}{w^2}) dx;$$

$$(E) \ E[X] = \int_0^w x(\frac{2}{w} - \frac{2x}{w^2}) dx;$$

$$(F) \ E[X] = \int_0^w w(\frac{2}{w} - \frac{2x}{w^2}) dx;$$

2. What is $\mathbb{P}(x = 1 | w = 2)$?

3. When $w = 2$, what is $p(1)$?

Exercise 3

Let X and Y be two continuous random variables. The conditional expectation of Y on $X = x$ is defined as the expectation of Y with respect to the conditional probability density $p(Y|X)$:

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

where $p_x(x)$ is the marginal probability density of Y . Show the following properties of the conditional expectation:


1. $E_{p_y} Y = E_{p_x}[E(Y|X)]$, where E_{p_y} means taking the expectation with respect to the marginal probability density p_y .

Remark: This formula is sometimes called the tower rule.

2. If X and Y are independent, then $E(Y|X = x) = E(Y)$.
3. The minimizer of the following minimization problem with respect to some constant $c \in \mathbb{R}$

$$\operatorname{argmin}_c E[(Y - c)^2 | X = x]$$

is attained at $c^* = E(Y|X = x)$.

 **Exercise 4** Hereby, I would first say sorry for my mistake made in class: “the rand distance does not satisfy the triangle inequality” is wrong. Here in this problem you can show it.

1. The symmetric distance (or rand distance) between two sets $A \subset \Omega$ and $B \subset \Omega$ is defined as $R_{\delta}(A, B) = \frac{|A \setminus B| + |B \setminus A|}{|\Omega|} = \frac{|A \Delta B|}{|\Omega|}$, where $|S|$ stands for the number of elements in the set S . Show that the rand distance R_{δ} is actually a distance, i.e., it satisfies the three properties:

- a) Positivity: $R_\delta(A, B) \geq 0$, and “=” if and only if $A = B$;
 - b) Symmetry: $R_\delta(A, B) = R_\delta(B, A)$;
 - c) Triangle inequality: $R_\delta(A, B) \leq R_\delta(A, C) + R_\delta(B, C)$.
2. (Optional) The Jaccard distance between two sets A and B is defined as $J_\delta(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \Delta B|}{|A \cup B|}$. Show that the Jaccard distance J_δ is also a distance satisfying the three above properties.