SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Assignment 1

Due Date: Mar 11, 2025

✏ Exercise 1

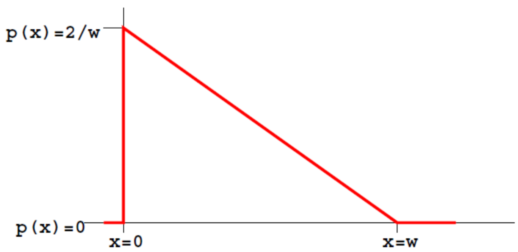
Given the ordered data $\{x_{(i)}\}_{i=1}^{2n-1}$ with increasing order. Show that the median of the data set is equal to the minimizer of the following L^1 minimization problem:

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

✏ Exercise 2

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

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



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

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

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

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

(D) $E[X] = \int_{0}^{w}(\frac{2}{w} - \frac{2x}{w^2})\mathrm{d}x$;

(E) $E[X] = \int_{0}^{w} x(\frac{2}{w} - \frac{2x}{w^2})\mathrm{d}x$;

(F) $E[X] = \int_{0}^{w} w(\frac{2}{w} - \frac{2x}{w^2})\mathrm{d}x$;

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 $p(Y|X)$:

$$\mathrm{E}(Y|X=x) = \int_{\mathcal{Y}} yp(y|X=x)\mathrm{d}y = \frac{\int_{\mathcal{Y}} yp(x,y)\mathrm{d}y}{p_x(x)},$$

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

1. $\mathrm{E}_{p_y} Y = \mathrm{E}_{p_x}[\mathrm{E}(Y|X)]$, 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 $\mathrm{E}(Y|X = x) = \mathrm{E}(Y)$.
3. The minimizer of the following minimization problem with respect to some constant $c \in \mathbb{R}$
   $$\arg\min_c \mathrm{E}[(Y-c)^2|X=x]$$
   is attained at $c^* = \mathrm{E}(Y|X = x)$.

✏ **Exercise 4** In this exercise, we would like to invite you get a comprehensive understanding of the concept of distance. The symmetric distance (or rand distance) between two sets $A \subset \Omega$ and $B \subset \Omega$ is defined as $R_\delta(A,B) = \frac{|A\backslash B|+|B\backslash A|}{|\Omega|} = \frac{|A\triangle B|}{|\Omega|}$, 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:

1. Positivity: $R_\delta(A,B) \geqslant 0$, and "=" if and only if $A = B$;
2. Symmetry: $R_\delta(A,B) = R_\delta(B,A)$;

3. Triangle inequality: $R_\delta(A,B) \leqslant R_\delta(A,C) + R_\delta(B,C)$.

Remark: For students who are interested in the this concept, we invite you to consider why the Jaccard distance between two sets A and B ($J_\delta(A,B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \triangle B|}{|A \bigcup B|}$) also satisfies the three above properties.