SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Assignment 1

Due Date: Mar 8, 2022

✏ Exercise 1

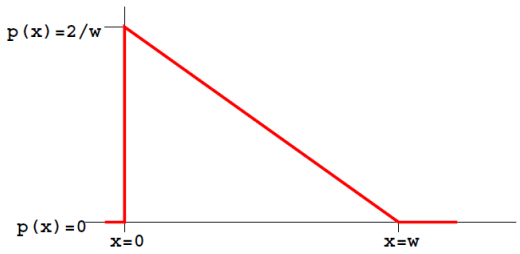
Given the ordered data $\{x_{(i)}\}_{i=1}^{2n-1}$ with increasing order. Show that the median of the data set is equal to the minimizer of the following L^1 minimization problem:

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

✏ 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 $p(Y|X)$:

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

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**

The Jaccard distance between two sets A and B is defined as $J_\delta(A, B) = 1 - \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \triangle B|}{|A \bigcup B|}$, where $|S|$ stands for the number of elements in the set S. Show that the Jaccard distance J_δ is actually a distance, i.e., it satisfies the three properties:

1. Positivity: $J_\delta(A, B) \geqslant 0$, and "=" if and only if $A = B$;
2. Symmetry: $J_\delta(A, B) = J_\delta(B, A)$;
3. Triangle inequality: $J_\delta(A, B) \leqslant J_\delta(A, C) + J_\delta(B, C)$.

✏ **Exercise 5** Suppose you have 10,000 data points $\{(x_k, y_k) : k = 1, 2, \ldots, 10000\}$. Your dataset has one input and one output. The k-th data point is generated by the following recipe:

$$x_k = k/10000, y_k \sim N(0, 2^2).$$

So that y_k is all noise: drawn from a Gaussian with mean 0 and variance $\sigma^2 = 4$. Note that its value is independent of all other y values. You are considering two learning algorithms:

- Algorithm NN: 1-nearest neighbor.
- Algorithm Zero: Always predict zero.

Define the mean squared error by $E_{P(X,Y)}(Y - f(X)^2)$ which can be approximated by its empirical version $\frac{1}{n}\sum_{k=1}^{n}(y_k - f(x_k))^2$ with $(x_k, y_k) \overset{i.i.d.}{\sim} P(X, Y)$.

1. What is the **expected mean squared training error**, i.e., $\mathrm{E}\left(\frac{1}{n}\sum_{k=1}^{n}(y_k - f(x_k))^2\right)$ where the expectation is taken over all samples (x_k, y_k), for Algorithm NN?
2. What is the **expected mean squared training error** for Algorithm Zero?
3. Recall the leave-one-out cross validation estimator is defined over the training set $\{(x_k, y_k) : k = 1, 2, \ldots, 10000, k \neq i\}$ for each sample (x_i, y_i). What is the **expected mean squared leave-one-out cross-validation error** for Algorithm NN?
4. What is the **expected mean squared leave-one-out cross-validation error** for Algorithm Zero?