# Intro to Big Data Science — Spring 2023-2024

**Name:** _____ **ID No.:** _____

**Quiz1: Concepts in Data Science and Preprocessing**

**To receive credit, this worksheet MUST be handed in at the end of the class.**

1. What are the key features of "BIG" data? (4 big "V")

2. You are running a company, and you want to develop learning algorithms to address each of two problems.

   - Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
   - Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

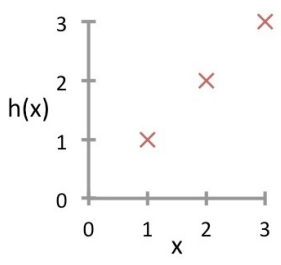
   Should you treat these as classification or as regression problems?

   (A) Treat both as classification problems.
   (B) Treat both as regression problems.
   (C) Treat problem 1 as a classification problem, problem 2 as a regression problem.
   (D) Treat problem 1 as a regression problem, problem 2 as a classification problem.

3. Of the following examples, which would you address using an unsupervised learning algorithm? (Select all that apply)

   (A) Given email labeled as spam/not spam, learn a spam filter.
   (B) Given a set of news articles found on the web, group them into set of articles about the same story.
   (C) Given a database of customer data, automatically discover market segments and group customers into different market segments.
   (D) Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

4. Suppose we have a training set with $m = 3$ samples, plotted below. Our hypothesis representation is $h_\theta(x) = \theta_1 x$, with parameter $\theta_1$. The loss function is $J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$. What is $J(0)$?



5. Which of the following statistics could be applied to missing value completion? (Select all that apply)

   (A) Mean

   (B) Variance.

   (C) Standard deviation

   (D) Median

   (E) Mode

   (F) Zero

   p

6. True or false:

   1) Dummy variable is used to deal with the missing values in continuous variable

   2) Both Manhattan distance and Jaccard distance satisfy the three properties: positive definiteness, symmetry, and triangle inequality.

7. For a two-class problem, compare the 1-nearest-neighbor method vs. Bayes classifier (classify the point to the most probable class, i.e., the class with greater probability), which method has a larger classification error? And why?