Intro to Big Data Science — Spring 2021-2022

Name: 刘乐奇 ID No.: 12011327

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")

   Volume、Variety、Value、Velocity

D

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.

A

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

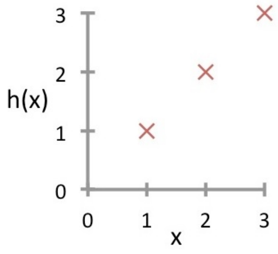
   (A) Given email labeled as spam/not spam, learn a spam filter.

   ~~(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 θ_1. The loss function is $J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$. What is $J(0)$?



$$J(0) = \frac{1}{6}\sum_{i=1}^{3}(0 \cdot x^{(i)} - y^{(i)})^2$$

$$= \frac{1}{6}\sum_{i=1}^{3}(y^{(i)})^2 = \frac{49}{3}$$

AEF

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

BD

6. Which of the following statement is false?

   (A) When the sample size is large and the number of missing values is small, the best way is deleting the data with missing values ✓

   (B) For the data of numeric type, we could fill in the missing values with the means or medians of the corresponding columns

   (C) For the data of non-numeric type, we could fill in the missing values with the modes of the corresponding columns ✓

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

B

7. Which of the following statement is false?

   (A) The assumption for the 3σ rule is that the data follow the normal distribution approximately ✓

   (B) 3σ rule treats the samples which deviate from the mean by three times standard deviation as outliers

   (C) 3σ rule is effective for multiple dimensional data ✓

   (D) Box plot consists of non-outlier minimum, lower quartile, median, upper quartile, and non-outlier maximum ✓

8. 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?

1-nearest-neighbor method has a larger classification error.

Bayes error $= 1 - P_{c^*}(x)$ where $c^* = \arg\max_c P_c(x)$

1NN error $= \sum_{c=1}^{C} P_c(x)(1 - P_c(x)) \rightarrow 1 - \sum_{c=1}^{C} P_c^2(x)$

$\leq 1 - P_{c^*}^2(x)$

$\leq 2(1 - P_{c^*}(x)) = 2$(Bayes Error)