BYU IDAHO

02 Prepare : RAT - Data Exploration and Clustering

CSE 450 – Readiness Assurance Test

Instructions

Indicate your selection of the correct answer on your accompanying answer sheet. Make sure to mark them in some way on this question sheet in addition to your answer sheet, so you will have your answers available to discuss with your team.

1. What is a potential problem with using the mean for a feature that has a bimodal distribution?

    A. It is difficult to compute

    B. It will be drawn to heavily to one of the peaks

    C. It may be that very few instances actually have this value

    D. It will be skewed by the large presence of outliers

2. When does the k-means algorithm terminate?

    A. When the cluster assignments of the points do not change

    B. After k steps

    C. After a predetermined number of steps

    D. When the number of clusters does not change

3. Which of the following is a true with regard to range normalization (Min-Max scaling)?

    A. It ensures that the standard deviation of every attribute is 0

    B. It can have problems when outliers are present

    C. It can be very computationally expensive if the original scale is very large

    D. It does not work very well when there are multiple attributes

4. How does the k-means algorithm reach the optimal clustering assignment?

    A. By using all possible values for k

    B. By continuing to iterate until improvement stops

    C. By iteratively running the process with different initial values

    D. It does not guarantee an optimal clustering

5. What is the effect of standardizing the data (i.e., z-score normalization)?

    A. It allows attributes of different units to be mixed mathematically

    B. It ensures that all the data in the dataset is numeric in nature

    C. It keeps only standard data points, effectively removing outliers

    D. It ensures that each target value (or class) will have equal representation in the dataset

6. What changes must be made to k-means so it can be run in more than 2 dimensions?

    A. The number of clusters must be increased

    B. The number of points must be increased

    C. The number of centers must be increased

    D. No changes are required

7. What is one of the main advantages to using binning on a feature?
    - A. It allows us to handle low cardinalities.
    - B. It allows us to transform the data into a normal distribution.
    - C. It allows us to ignore missing values.
    - D. It allows us to transform a continuous feature into a categorical feature.
8. What type of machine learning algorithm is k-means?
    - A. Unsupervised
    - B. Classification
    - C. Regression
    - D. Supervised
9. The purpose of plotting a dendrogram in Hierarchical Clustering is to:
    - A. Help us to avoid having to manually decide the number of clusters to use.
    - B. Help us determine the number of clusters to use
    - C. Provide an extra dimension to our data to help delineate outliers.
    - D. Help us to calculate the interquartile range without having to use a box plot.
10. What is the biggest risk when using imputation to handle missing values?
    - A. It can result in too many rows being deleted.
    - B. It can result in too many features being deleted.
    - C. It can introduce bias into the data.
    - D. There are no dangers when using imputation. It is always the preferred method of handling missing values.