CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

Quiz 03 Solution

Name (Print):_____ Section:_____

1. What is the primary purpose of a pivot table?
   a.) Visualize data trends
   b.) **Summarize and organize data**
   c.) Perform statistical tests
   d.) Create predictive models

2. Which of the following is an example of a summarization technique in data mining?
   a.) **Calculating the total sales data by region**
   b.) Predicting next month's sales using regression
   c.) Creating a scatter plot of sales vs. time
   d.) Calculating the Minkowski distance

3. Which of the following is an example of a predictive method?
   a.) Clustering customer data into segments
   b.) **Forecasting the likelihood of customer churn**
   c.) Summarizing sales data by region
   d.) Calculating average sales over time

4. Descriptive methods are primarily used to
   a.) **Explain patterns and relationships in data**
   b.) Predict future trends
   c.) Test hypotheses

5. You are analyzing customer purchase data to identify common purchasing patterns. Which method would you use?
   a.) Predictive
   b.) **Descriptive**
   c.) Both

6. Which similarity metric is based on the absolute differences between data points?

   Manhattan

7. Which of the following is true about the L2 (Euclidean) distance?

   a.) It is less sensitive to outliers than L1 distance
   b.) It calculates the absolute difference between dimensions
   c.) It can only be used for clustering categorical data
   d.) **It is equivalent to the square root of the sum of squared differences**

8. The Minkowski metric is:

   a.) A special case of the Manhattan distance
   b.) **A generalized distance metric for L1 and L2 with a variable parameter p**
   c.) Used only in hierarchical clustering
   d.) A special case of the Euclidean distance

9. Which of the following is a characteristic of hierarchical clustering?

   a.) It requires the number of clusters to be specified in advance
   b.) **It builds a tree-like structure to represent data relationships**
   c.) It assigns data points to clusters based on their centroid
   d.) It is computationally faster than K-means for large datasets

10. What is a limitation of K-means clustering?

    a.) It cannot handle large datasets
    b.) It does not require specifying the number of clusters
    c.) **It requires the number of clusters to be specified in advance**
    d.) It only works for hierarchical data structures

11. In hierarchical clustering, the structure representing nested clusters is called:

    a.) Centroid
    b.) Tree diagram
    c.) **Dendrogram**

12. Compare Hierarchical clustering and K-means clustering. What are the advantages and disadvantages of each? Briefly describe in your own words.

    (see table in the lecture notes)