Dhirubhai Ambani Institute of
Information and Communication Technology

End-Semester Exam

IT496: Introduction to Data Mining

Date: 13th December 2022 | Timings: 9:00 AM – 10:30 AM

READ THE INSTRUCTIONS CAREFULLY

- ALL questions are compulsory.
- For each question, write your answer with **brief** and **justified** explanations in 1-2 sentences in your answer sheets.
- Do not waste your time by overly explaining your answers.

10 questions | **40** total points | **90** minutes | Points are marked beside each question.

1. Suppose that for a data set - there are m points and K clusters. Half the points and clusters are in "more dense" regions, the other half the points and clusters are in "less dense" regions, and the two regions are well-separated from each other.

   For the given data set, which of the following should occur to minimize the squared error when finding K clusters. Justify your answer for each of the following cases.

   1.1. Centroids should be equally distributed between more dense and less dense regions. **[2 pts.]**

   1.2. More centroids should be allocated to the less dense region. **[2 pts.]**

   1.3. More centroids should be allocated to the denser region. **[2 pts.]**

   Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

2. Consider the mean of a cluster of objects from a binary transaction data set.

   2.1. What are the minimum and maximum values of the components of the mean? **[1 pt.]**

   2.2. What is the interpretation of components of the cluster mean? **[2 pts.]**

   2.3. Which components most accurately characterize the objects in the cluster? **[3 pts.]**

3. Would the cosine measure be the appropriate similarity measure to use with K-means clustering for time series data? Why or why not? If not, what similarity measure would be more appropriate? **[3 pts.]**

4. The total Sum of Squared Error (SSE) is the sum of the SSE for each separate attribute.

4.1. What does it mean if the SSE for one variable is low for all clusters? **[1 pt.]**

4.2. Low for just one cluster? **[1 pt.]**

4.3. High for all clusters? **[1 pt.]**

4.4. High for just one cluster? **[1 pt.]**

4.5. How could you use the per-variable SSE information to improve your clustering? **[1 pt.]**

5. The classic *Olivetti* faces dataset contains 400 grayscale 64 x 64-pixel images of faces. 40 different people were photographed (10 times each). Now, create the pipeline of operations to create clusters of similar faces. Just mention the operations' names in order and why you are using each one of them. **[5 pts.]**

Hint: (i) Flattening – to flatten the images into a 1D vector of size 4096; and so on.

6. Suppose you perform PCA on a 1,000-dimensional dataset, setting the *explained variance ratio* to 95%. Discuss the cases below.

6.1. What can be the minimum number of dimensions, and when? **[1 pt.]**

6.2. What can be the maximum number of dimensions, and when? **[1 pt.]**

7. Suppose you apply PCA (a dimensionality reduction algorithm) to your dataset.

7.1. How can you evaluate its performance? Justify your answer. **[2 pts.]**

7.2. Does it make sense to *chain* two PCA algorithms? Justify your answer. **[2 pts.]**

8. If you have trained five different models on the same training data, and they all achieve 95% precision, is there any chance you can combine these models to get better results? If so, how? If not, why? **[2 pts.]**

9. Consider the Decision Tree algorithm -

9.1. If it takes one hour to train a Decision Tree on a training set containing 1 million instances, roughly how much time will it take to train another Decision Tree on a training set containing 10 million instances? **[1 pt.]**

9.2. If your training set contains 100,000 instances, will setting **`presort=True`** speed up training? **[1 pt.]**

9.3. If a decision tree overfits the training set, can you suggest some way-outs? **[2 pts.]**

9.4. Is a node's Gini impurity generally lower or greater than its parent's? Is it generally lower/greater, or always lower/greater? **[1 pt.]**

10. Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training, while the remaining half is used for testing.

10.1. Suppose there are an equal number of positive and negative records in the data, and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data? **[0.5 pt.]**

10.2. Repeat the previous analysis assuming that the classifier predicts each test record to be a positive class with a probability of 0.8 and a negative class with a probability of 0.2. **[0.5 pt.]**

10.3. Suppose two-thirds of the data belong to the positive class, and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive? **[0.5 pt.]**

10.4. Repeat the previous analysis assuming that the classifier predicts each test record to be a positive class with a probability of 2/3 and a negative class with a probability of 1/3. **[0.5 pt.]**

*** **End of the Paper** ***