ID: 1174066

CPSC 552 –Midterm Exam... – Spring 2024

Problem #1: What is the sum of all Ø components in the GMM model

a) 1 b) 100 c) 0 d) any number

Problem #2: If a dataset is loaded into a dataframe called df. Then the statement print(df.head()) prints:

a) Statistics about the data such as mean, min max values and variance.
b) Names of columns and first few rows and column values.
c) Sum of each column d) Name of each column only.

Problem #3: Find the Euclidean distance between the following two vectors:

$$v1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad v2 = \begin{bmatrix} 4 \\ 2 \\ 7 \end{bmatrix}$$

a) 24 b) 5 c) 25 d) 6

Problem #4: If a dataset is loaded into a dataframe called df. Then the statement print(df.describe()) prints:

a) Statistics about the data such as mean, std, min max values.
b) Names of columns and first few rows and column values.
c) Sum of each column and its name.
d) Name of each column only.

Problem #5: If you wanted to remove null and duplicate values from a dataframe called df, then the code to accomplish this is:

a) df.dropna() b) df.dropnulls() c) df.clean() d) df.removenulldup()

Problem #6: In assignment 2, for the **Iris** dataset, we split the data into training and test parts. The number of total data items i.e., **total rows, size of train set, size of test set** was:

a) 500, 300, 200 b) 100, 50, 50 c) 300,200, 100 d) 150, 100, 50

Problem #7: The Naïve Baye's algorithm operates on the principle that the features in a given data item are: a) correlated b) independent c) have 0 mean d) have std=1

Problem #8: The three classes in the Iris dataset are named:

a) setosa, versicolor, virginica b) selenium, calcium, valium
c) small, médium, large d) moore, penrose, newton

Problem #9: If Naïve Bayes algorithm is applied to the Iris dataset, then for each feature, we will need to compute the mean and variance per category. **How many categories of flowers are there, and how many features each flower has**?

a) 4, 4 b) 3, 4 c) 5, 4 d) 150, 3.

Problem 10: Suppose you have to apply Gaussian Mixture Model to the Iris dataset. How many Gaussian distributions will you create?

a) 2 b) 5 c) 3 d) 4

Problem #11: The mean of each distribution in GMM.is initialized to:

a) First row of data b) Last row of data c) Middle row of data d) random rows in data

Problem #12: Suppose the data given to you is weights of persons with half being males and the other half females. The mean weight for males is 140, and the mean weight for females is 120. The std deviation for males is 20, and for females 20. If a person has a weight of 130 lbs, how will it be classified by the GMM model:

a) male b) female c) either male and female d) unknown category

Problem #13: What is the inner product of the following two vectors:

$$v1 = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \quad v2 = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$$

a) 0 b) 8 c) 9 d) 3

Problem #14: Suppose the data given to you is weights of persons with half being males and the other half females. The mean weight for males is 140, and the mean weight for females is 120. The std deviation for males is 20, and for females 20. If a person has a weight of 131 lbs, how will it be classified by the GMM model:

a) male b) female c) either male and female d) unknown category

Problem #15: Suppose you are given data with five features i.e., height, weight, age, blood pressure, and glucose level. You are given data for 500 persons with some having diabetes, some prediabetes, and the rest no diabetes. If you were to create a GMM model for this dataset, **what will be size of covariance matrix, and how many Gaussian distributions will you model**

a) 3, 5 b) 5, 5 c) 5, 3 d) 5, 500

Problem #16: In assignment 4, we analyzed the Abalone dataset using the KNN algorithm. We had dropped one of the data columns. What was the name of the data column that was dropped?

a) Diameter b) Length c) Sex d) Shucked weight

Problem #17: In the KNN algorithm, what is the lowest number of neighbors you can have.

a) 1 b) 2 c) 3 d) 5

Problem #18: Using the KNN algorithm, if the 5 nearest neighbors to the unknown Abalone are: neighbor 1 : 10 rings with dist=12, neighbor 2 :12 rings with dist=9, neighbor 3: 13 rings with dist=12.5, neighbor 4 :11 rings with dist=11.5, and neighbor 5: 10 rings with dist 10.6, then the predicted age of the Abalone using the mean approach is:

a) 11.2 b) 10 c) 12 d) 11

Problem #19: For the previous problem, the predicted age of the Abalone using the mode approach is:

a) 11.2 b) 10 c) 12.2 d) 11

Problem #20: If k=3, then the predicted age of the Abalone using the mean approach is:

a) 11.2 b) 10 c) 12.2 d) 11

Problem #21: Which one of the following vectors is **<u>not</u>** a basis vector (hint: inner product of any two basis vectors is 0)?

1. $\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 1 \end{bmatrix}$ 2. $\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 1 \end{bmatrix}$ 3. $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$ 4. $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

a) 1 b) 2 c) 3 d) 4

Problem #22: Which of the following statements is false

a) The Eigen vectors are mutually orthogonal to each other
b) SVD decomposes a matrix into three matrices
c) It is better to create and use Eigen vectors corresponding to larger Eigen values
d) SVD is a popular technique for dimensionality reduction of data

Problem #23: If we apply SVD to compress an image that has 100x80 pixels, what will be the size of $\sum$ Matrix?

a) 100x80 b) 80x100 c) 80x80 d) 100x100

Problem 24: If we apply SVD to compress an image that has 100x80 pixels, and use only two Eigen values to compress the image. The size of data in the compressed image will be:

a) 362 b) 4000 c) 200 d) 562

Problem #25: Compute the Eigen values for the following matrix are (hint: det($\lambda I - A) = 0$):

$$A = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}$$

a) 2, 4 b) 1, 3 c) 2,5 d) 1,5

Problem #26: Suppose you wanted to implement PCA for Face recognition and reduce the dimensionality of 100x100 pixel face image to 40. If you have 30000 pictures in your dataset, then the size of the covariance matrix will be:

a) 10000x30000 b) 10000x10000 c) 30000x30000 d) 30000x10000

Problem #27:

Compute the SVD of $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$?:

The *U* matrix will be:

a) $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$ c) $\begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$ d) $\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$

Problem #28: Compute the SVD of $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$?:

The $\sum$ matrix will be:

a) $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$ d) $\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$

Problem #29: :

Which of the following statements is false

a) Pseudo Inverse is used for the inverse of a square matrix
b) SVD is used in computing the Pseudo inverse
c) SVD is a popular technique for data reduction
d) PCA is a popular technique for dimensionality reduction of data

Problem #30: Which assignment numbers in CPSC 552 cover the GMM and the SVD:

a) 2, 4 b) 3, 6 c) 2,5 d) 4,5