CSC3005

DATA ANALYTICS SAMPLE LAB TEST

Sample Lab Test - Trimester 3, Academic Year 2020/21

xx, xx/xx/202x

xx:xxam – xx:xx am

(1 hours 30 minutes)

Instructions to students:

1. This sample lab test comprises **THREE (3)** Questions printed on **THREE (3)** pages including the cover page.
2. You must answer **ALL** the questions.
3. **Submission Instruction**
    a. **All source code with jupyter notebook with .ipynb file must be submitted electronically to xSITE drop box under CSC3005 Data Analytics.**
    b. **Name the .ipynb file as your name_SIT_student_number.ipynb**

Question 1: PCA, Choosing the data dimension (30 marks)

Download the then sampletest1.csv dataset from Xsite dropbox.

Analyse the dataset and perform principle component analysis (PCA) to determine the dimension reduction required. Following step should be performed:

A. Load the dataset and display the data
B. Data Visualization on the dataset
C. PCA Eigen Decomposition
D. Plot the eigen variance ratio and determine the number of principle component (eigenvector) to use. Hint : The total variance energy of the PCs being chosen should be minimum 65% or more.
E. Dimension Reduction using the chosen number of PCs
F. Plot the new dimension reduced data using the chosen PCs in the eigenspace and compared to the original dataset with the same number of data feature as the chosen PCs.

Question 2: Classifier on given data labels (40 marks)

Using the dimension reduced data, performed the classification performance on the following two classifiers namely

I. K-Nearest Neighbour (KNN)
II. Decision Tree Classifier

Following objective should be achieved:

A. Create training and testing dataset from the Principle Component Data data based on 70,30 split
B. Test the performance of KNN and decision tree classifier to determine the best value of K and max depth respectively. Plot of number of neighbours versus Accuracy and Max depth versus accuracy with the training and testing dataset for both algorithms respectively should be performed
C. Decision boundary Visualization of KNN and Decision Tree against the original dataset should be plot.
D. Conclude the performance of both algorithms based on accuracy in (B) and Decision boundary visualization in (C)

Question 3: A Mystery (30 marks)

---END OF PAPER---