COMP-5411 FA & FB
Fall 2019
Assignment 2

This is an individual assignment. Cheating and plagiarism (copying from friends/internet) in any part of the assignment will be dealt with according to the university policy.

Due Date: Monday, November 25, 11:59 pm
Late submission is NOT allowed.

1) Write a program to implement the *k-medoids* algorithm. The program will take two inputs: a data file containing multiple continuous features and a value for *k*. Your program will load the data file, calculate the dissimilarity matrix and then apply the k-medoids algorithm. The clustering process will continue until cluster assignments do not change or maximum 100 iterations have been completed. The program will also calculate average silhouette width for the final clustering solution. The program will add a new column containing cluster ids (1, 2, 3, …, k) to the original input file and save it to clusters_k.csv, where k is the number of clusters. The program will also show the average silhouette width in the console. You are allowed to use a library for distance calculation. However, ***you cannot use any library for clustering or silhouette width calculation***.

2) Write a program to implement the Naïve Bayes Algorithm (m-estimate version) for a binary classification problem. The program will take two inputs: a training dataset and a test dataset. The value of *m* will be the number of possible values of a feature. All features will be categorical in the dataset. The last column of the datasets will be the class variable. Rename the two class values as: positive and negative. Your program will first load the training dataset, calculate all required probabilities, and then predict the class of each instance in the test set. The program will add a new column containing predicted classes to the test dataset and save it as predictions.csv. The program will also show the accuracy, sensitivity and specificity on the console. **You *cannot* use any library for naïve Bayes algorithm or performance calculation.**

3) Write a program to implement the sequential forward selection algorithm to identify important features. The program will take one input: a dataset where the last column is the class variable. The program will load the dataset, and then use wrapper approach with sequential forward selection strategy to find a set of important features. You can use any supervised learning method for measuring the performance (accuracy) in the wrapper approach. You will use stratified 5-fold cross validation for measuring accuracy. Your program will keep adding the features as long as there is some improvement in the classification accuracy. The output of the program will be the set of important features on the console. You are allowed to use a library for the supervised learning algorithm of your choice. However, ***you cannot use libraries for feature selection or 5-fold cross validation.*** You must write a function on your own to calculate accuracy using stratified 5-fold cross validation.

Deliverables on mycourselink:

1) Clustering: a zipped folder containing the following:
    a. Source code
    b. The dataset that you have used to validate your program. The dataset must have minimum 3 numeric features.
    c. Use 2 different values of k and upload clusters_k.csv files
2) Naïve Bayes: a zipped folder containing the following:
    a. Source code
    b. Training and test datasets that you have used to validate your program. The dataset must have minimum 3 categorical features.
    c. Predictions.csv file
3) Feature selection: a zipped folder containing the following:
    a. Source code
    b. The dataset you have used to validate your program. The dataset must have minimum 20 features. The data type of the features does not matter.

Important notes:

i) Read the assignment carefully and make sure you understand which libraries you can or cannot use.
ii) Your source code should have enough comments for readability. It is your responsibility to make sure that your code is understandable by others. Uncommented code will significantly reduce the probability of getting marks in different individual components of the assignment.
iii) The code should execute properly with your uploaded input files. You will lose a significant part of your marks, if the code does not run without errors.