COMP-4311-WA
Winter 2020
Assignment 2 & 3

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: Friday, March 20, 10:00 pm
Late submission is NOT allowed.

1) Write a program to implement the *k-means* algorithm. The program will take two inputs: a data file containing multiple (at least 3) continuous features and a value for *k*. Your program will load the data file and then apply the k-means 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 one input: a dataset with multiple categorical features (at least 3) where the last column is the class variable. All features will be categorical in the dataset. Rename the two class values as: positive and negative. You will use stratified 5-fold cross validation for measuring accuracy, sensitivity and specificity. The program will add a new column containing predicted classes to the original 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 cross validation or performance calculation.**

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. The dataset you have used to conduct your experiments. It must have minimum 3 categorical features.
   c. Predictions.csv file

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 the evaluator.