CSE 5243
Instructor: Jason Van Hulse
Homework 5 (Clustering)

Due Date: Tuesday, 4/2/2019 11:59pm (Submit the via Carmen). *There will be a 15% penalty for all homework handed in up to 24 hours late; homework will not be accepted more than 24 hours late]*

Objective: In this lab, you will focus on *clustering*. This lab is divided into 2 parts. In the first part, you will implement the k-Means clustering algorithm and test your program. In the second part, you will use any off-the-shelf clustering algorithm to cluster the Wine Dataset.

Dataset 1: *TwoDimHard* - Two numeric independent variables; 4 true, slightly overlapping clusters; 400 examples
Dataset 2: *Wine* - (https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv. Use the *winequality-red.csv* dataset.) Remove the class variable from the dataset. When doing clustering, do NOT use the quality attribute (the dependent variable). This attribute will be used for external validation of the clusters as discussed below.

Part 1 - Implement a k-Means clustering algorithm (Python or R) for Dataset 1 ONLY

The program should accept as a parameter the number of clusters k, specified by the user. With the k-Means algorithm, implement the standard Euclidean distance measure.

The output of the program should consist of two columns - (1) the row ID, and (2) the cluster that each record belongs to, as determined by the clustering method.

Note: It is expected for you to code this algorithm from scratch, and not to use existing functions.The only built in *mathematical* or *statistical* functions you should use are mean, median, standard deviation, minimum and maximum.

In addition to turning in your program, you should create a brief report which describes your program and how it works. Discuss the design decisions that you made. You do *not* need to turn in the output dataset, rather these will be obtained by running your code.

A. Given that you know the true clusters, compute the true cluster SSE, the overall SSE and the between-cluster sum of squares SSB for each dataset.
B. Run your k-Means algorithm (assuming k = 4)
    1. Compute the SSE for each cluster (and the overall SSE) as well as the between-cluster sum of squares (SSB)
    2. Create scatterplots for Dataset 1, overlaying the true cluster with the cluster produced by k-Means (or you can have two side by side scatterplots, one showing the true cluster membership, the other showing the clusters assigned by k-Means).

3. Create a cross tabulation matrix comparing the actual and assigned clusters

C. Change the number of clusters to k = 3. Run your k-Means program and compute each cluster SSE, the overall SSE and the SSB, the scatterplot and cross tabulation matrix (as in part B 1- 3). Analyze these results compared to part B above. Answer the question on whether changing the number of clusters changes the results, and if so, for better or worse?

Part 2 - Run off-the-shelf clustering algorithms for the Wine dataset ONLY

Using off the shelf algorithms available in Python or R, perform clustering analysis on the Wine dataset. You do NOT need to write code to implement any clustering algorithms. You should use at least 3 different clustering methods - e.g., k-Means, DBSCAN, Agglomorative Hierarchical.

The report should include the following components:

1. Discuss the clustering methods that you used and any parameter settings that you chose
2. Present the results of your cluster analysis. What conclusions can you draw from this analysis? You can use the *quality* attribute as an external validation measure.

What you need to turn in:

1. Code, either in Python or R
2. Readme - contains all the important information about the directory, including how to run the program and how to view the resulting output.
3. Report - Make sure you include both the description of your k-Means algorithm as well as your clustering analysis of the Wine dataset as described above.

Grading Rubric

1. **k-Means clustering Code** (40%): The grader should be able to run your k-Means clustering code, which reads in the dataset, performs any preprocessing, and executes the k-Means algorithm. The code does not need to be make interactive (for example, asking the user to provide a value for k). Instead, just hard-code a default value. The output of the program should be as specified above - i.e., the row id and the cluster assignment.
2. **Readability of your report** (30%): Is it well organized and does the presentation flow in a logical manner; are there many grammar and spelling mistakes; do the charts/graphs relate to the text, etc…
3. **Clustering analysis** (30%): Do the clustering you used and the parameter choices make sense for the data? Did you answer all of the questions outlined above? Did you justify your choices and discuss different options?

How to hand in your work:

Please do this work in either Python or R. All the related files (code and/or report) except for the data will be tarred in a *.zip file or *.tgz file, and submitted via Carmen. Use this naming convention: “Project5_Surnames_DotNumber.zip" or “Project5_Surnames_DotNumber.tgz.” The submitted file should be less than 5MB.