CSE 5243
Instructor: Jason Van Hulse
Homework 5

Due Date: Monday, 11/6/2017 5:30pm. Submit your material in Carmen.

This lab is divided into 2 parts. In the first part, you will implement the k-means clustering algorithm and test your program on two different datasets. In the second part, you will use any off-the-shelf clustering algorithm to cluster these same datasets.

Dataset 1: *TwoDimHard* - Two numeric independent variables; 4 true, slightly overlapping clusters; 400 examples
Dataset 2: *Wine* - Remove the class variable from the dataset if you are using the data from homework #4. 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.

You are highly encouraged to work on this entire homework as part of a team (max of 2 people per team).

Part 1 (75% of grade)

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.

In addition to turning in the program, you should create a report which includes the following:

A. Describe your program and how it works. Discuss the design decisions that you made.

For Dataset 1:

B. 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.
C. Run k-means for each dataset (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
D. 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 C, parts 1 - 3). Analyze these results compared to part C above. Answer the question on whether changing the number of clusters changes the results, and if so, for better or worse?

For the Wine dataset:

E. Experiment with different numbers of clusters and compare your results. What conclusions can you draw from your analysis?
F. Provide detailed analysis of the results. What trends did you observe? Provide graphs and/or statistics to back up and support your observations. Do you have a preferred clustering?
G. Use the **quality** attribute for external validation. Compare your clustering results to this attribute.

Part 2 (25% of grade)

Using any off-the-shelf k-Means clustering method (e.g., in R, Matlab, Python, Weka, etc), run cluster analysis on the two datasets. You do NOT need to write code to implement any additional clustering algorithm.

The report should include the following components:

1. Discuss the clustering method that you used and any parameter settings that you chose.

2. Present the results of your cluster analysis, using similar metrics that were already discussed in Part 1.
3. Compare the results of this clustering method to the k-means implementation from Part 1.

What you need to turn in:

1. Code (from Part 1 only)
2. Makefile - *(note: no need to create Makefile as a PDF document)*
3. Readme - contains all the important information about the directory, including how to run the program and how to view the resulting output.
4. Report - The report should be well-written and be a maximum of 9 pages in length (14 pages for teams of 2). Hand in a hard copy of your report in class. Please proof-read and remove spelling and grammar errors and typos. *Writing and presentation will be part of your grade for this assignment.*

You do *not* need to turn in the output datasets, rather these will be obtained by running your code.

Note: It is expected for you to code these 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.

How to hand in your work:

Please choose one of the programming languages from: JAVA, Python, R. All the related files except for the data will be tarred in a **single** *.zip file or *.tgz file, and submitted via Carmen. Please use this naming convention: ”Project5_Surname_DotNumber.zip" or “Project5_Surname_DotNumber.tgz.” The submitted file should be less than 5MB.

On Linux System (Mac OS, Ubuntu, RedHat, etc.)

[Source Code, BashScript, Readme.txt, Report] should be submitted. Do not submit raw dataset.

The program should be able to run on a standard Linux system. Readme.txt tells me where to put input data (raw dataset) and where to find output data and how to interpret the output.

When I type command "bash BashScript", the output would be generated.

On Windows system

[Source Code, Readme.txt, Report] should be submitted. Do not submit raw dataset.

The program should be able to run on a Win-7 system. Readme.txt tells me where to put input data (raw dataset) and where to find output data and how to interpret the output. Readme also tells me how to compile and run the program.

If you use JAVA, please make sure your program can get compiled and run on Eclipse.