# MSc Data Mining and Machine Learning (2019)

# Lab 2 – Clustering

|  |  |  |
| --- | --- | --- |
| Instructor: |  | Dr Peter Jancovic |
| Written by: |  | James J Nkhata |
|  |  |  |
| Experiment Date: |  | 07/02/2019 |
| Report Delivery Date: |  | 21/02/2019 |

# PART 1: CLUSTERING

## Purpose

This laboratory (lab) activity aimed to discover the structure of a given dataset (Lab2Data) using Clustering. A text description of the nature of the identified data was given after observing the results.

To apply Clustering, Agglomerative clustering was used to obtain k-initial centroids which were later, used to apply K-means clustering to the dataset.

Lab2Data consisted of 1,000 points in 5-dimensional space with each point’s data held within a row of the dataset.

## Procedure

### Compiled the provided C files (.c) to executable files (.exe) using Microsoft Visual Studio Command prompt. The files provided for the Lab were agglom.c and k-means.c:

#### At the Microsoft Visual Studio command prompt typing ‘cl agglom.c <ENTER>’ generated agglom.exe.

#### At the command prompt typing ‘cl k-means.c <ENTER>’ to generated k-means.exe.

### Applied agglomerative clustering to the dataset to obtain a set of K-initial centroids (K representing the parameter passed on for the number of centroids):

#### At the command prompt typing ‘agglom Lab2Data agglom\_cent1 1 <ENTER>’ applied the agglomerative clustering to the Lab2Data dataset until “1” centroid was found, and the results were saved into the file agglom\_cent1 (centFile).

#### Repeated the process in step 1.2.2 for centroid values “2” up to “10” while altering the agglom\_cent file accordingly. Example ‘agglom Lab2Data agglom\_cent2 2 <ENTER>’ for 2 centroids.

Table 1: format for running agglomerative clustering on data set using different initial centroids (Lab2Data)

|  |  |  |  |
| --- | --- | --- | --- |
| agglom | Lab2Data | agglom\_cent1 | 1 |
|  |  |  |  |
| agglomerative executable | dataset | centFile to save centroid data | number of centroids  (numCent) |

### Applied 10 iterations of K-means clustering on the dataset Lab2Data using each of the centroid data obtained from step 1.2.2 “agglom\_cent” files (agglomerative clustering’s centFile):

#### At the command prompt typing ‘k-means Lab2Data agglom\_cent1 k-means-op1 10’ applied k-means clustering to the Lab2Data using the centFiles (“agglom\_cent1” to “agglom\_cent10”) for 10 iterations (numIter) each. The results were saved as “k-means-op1” (opFile).

#### Repeated the process in step 1.2.3 for “agglom\_cent2” to “agglom\_cent10” while saving the results to files “k-means-op2” up to “k-means-op10” respectively.

#### 

Table 2: Format for running 10 iterations of k-means clustering on data set (Lab2Data)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| k-means | Lab2Data | agglom\_cent1 | k-means-op1 | 10 |
|  |  |  |  |  |
| k-means executable | data set | Centroid data  (centFile) | saved distortions and new centroids (opFile) | number of iterations (numIter) |

## Experimental data

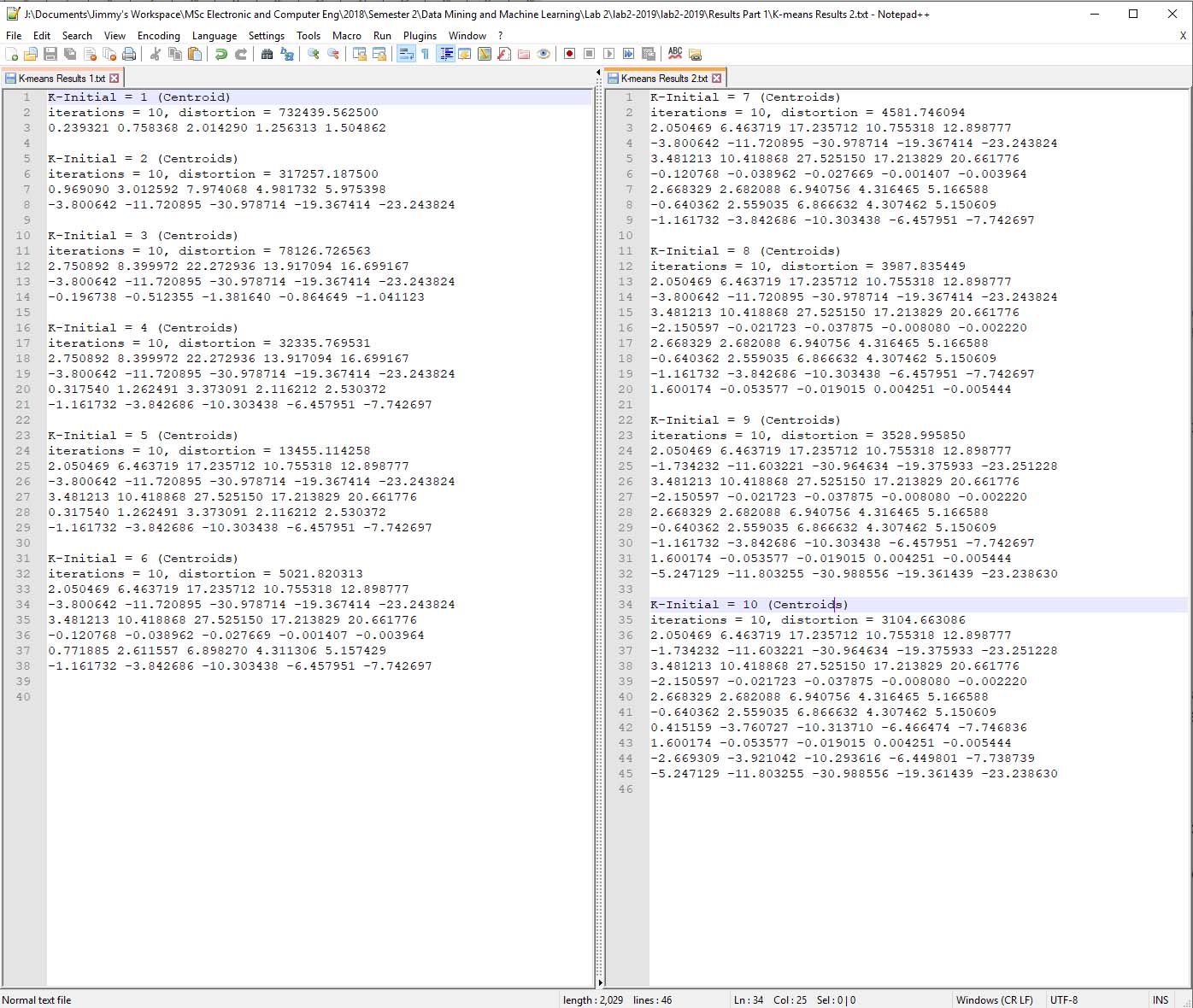


Figure 1: 10th iteration results for agglom and k-means applied for K (centroids 1 - 10)

Table 3: K-Intial values to distortions (K-means)

|  |  |
| --- | --- |
| K-Initial (Centroid number) | Distortion |
|  |  |
| 1 | 732439.5625 |
| 2 | 317257.1875 |
| 3 | 78126.726563 |
| 4 | 32335.769531 |
| 5 | 13455.114258 |
| 6 | 5021.820313 |
| 7 | 4581.746094 |
| 8 | 3987.835449 |
| 9 | 3528.99585 |
| 10 | 3104.663086 |

Figure 2: distortion as a function of K (centroids) graph

## Questions

When the k-means algorithm was applied to the saved centFiles (example agglom\_cent1) ‘k-means Lab2Data agglom\_cent1 k-means-op1 10’, the resulting saved opFiles (k-means-op1) returned the recalculated centroid coordinates and distortion for that iteration. The numbers at the bottom of the produced list were the centroid coordinates and distortion value of the 10th iteration.

## Conclusion

From the plotted graph of the results (Figure 2) obtained from k-means clustering of Lab2Data, it was observed that distortion for K (initial centroid) values “1” going to “5” were widely distributed. However, for K values of “6” up to “10” the distortion values were smaller and started to converge. The distortion for K = “1” was the highest and decreased as the number of centroids were increased with K=”10” having the lowest distortion value.

It was evident that from the results, the K value “10” initial centroids gave the lowest distortion. This led to the assumption that the dataset “Lab2Data” contained data from multiple sources that could be grouped into 6 or more clusters.

### References

Jancovic, P (2019) **Data Mining and Machine Learning** [Online]. Available from: <https://canvas.bham.ac.uk/courses/34771> [Accessed 30 January 2019].