

CSCE 5380  
Lab Assignment 2  
Date Issued: 09/21  
Date Due: 12/02  
Total Points: 100

---

This assignment is based on analyzing clustering techniques and you will be using the same data mining tool - *Weka* as in the previous lab assignment. The objective of this assignment is to become familiar with various clustering algorithms and evaluate or compare their performance using *Weka*. The tasks you need to do are listed below:

Task 1:

In this task, you would be comparing the performance of two clustering algorithms: Simple K-means and DBScan based on different parameters.

Problem 1: (20 points)

Select the “segment-test.arff” dataset in Weka and run the Simple K-means algorithm on it using 3, 6, 9, and 15 clusters with 2 different distance functions: *EuclideanDistance* and *ManhattanDistance* respectively with a 44% percentage split (in total you would need to run it 8 times – 4 clusters with each distance function). You can keep the default values of the remaining parameters like maximum number of iterations, random seed, etc.

1. Is there any correlation between the number of clusters and the sum of the squared errors (SSE) for the Euclidean distance function? If there is, why?
2. Is there any correlation between the number of clusters and the sum of within cluster distances for the Manhattan distance function? If there is, why?

Problem 2: (20 points)

Select the “segment-test.arff” dataset in Weka and run the DBScan algorithm on it using 3 values of *epsilon*: 0.3, 0.6 and 1.0 for 3 different values of *minPoints*: 4, 8, and 12 (in total you would need to run it 9 times – 3 for each) with a 44% percentage split. You can keep the default values for the remaining parameters (database type and distance function).

1. Is there any correlation between the *epsilon* values (keeping the *minPoints* constant) and the number of clustered and un-clustered instances (noise)?
2. As we increase both, the *epsilon* and *minPoints* values, what trend do you observe in the number of un-clustered instances (noise)? Why do you think this occurs?

## Task 2:

This task is based on loading a few external datasets into Weka and analyzing them using different clustering algorithms. The datasets are available on the instructor's webpage.

### Problem 3: (30 points)

In this problem you need to [download the "t4.8k.txt"](#) dataset and load it into Weka. Note that you cannot directly analyze this dataset in Weka. You will have to convert it into an appropriate format (either .csv or .arff) before running any algorithm on it.

After loading the dataset into Weka, run the Simple K-means algorithm on it.

1. How many clusters do you get?
2. Visualize the clusters (graph) and take a screenshot of your results to turn-in along with your answers.
3. Try running DBScan on it. You may use 0.05 as the value for epsilon and 6 as the value for MinPoints in DBScan. Do you get the same number of clusters with Simple K-means and DBScan? Why? Take a screenshot of your DBScan results to turn-in.

### Problem 4: (30 points)

In this problem you need to [download the "synthetic control.txt"](#) dataset and load it into Weka. Note that you cannot directly analyze this dataset in Weka. You will have to convert it into an appropriate format (either .csv or .arff) before running any algorithm on it.

After loading the dataset into Weka, run the 8 algorithms in the Weka clustering suite on the dataset (all algorithms except *FilteredCluster*)

1. How many clusters did you get for each one of the algorithms?
2. Were you able to run all of them? There would be a few of them that "don't seem to work". Which ones and why do you think they don't work?

### Submission instructions:

Please turn in a hard copy containing the answers to the four problems and screenshots before class on the due date.