**CSE 572: Data Mining Spring 2017**

**Assignment 4 / Mini Project 2**

**GROUP MEMEBRS:**

ADITI SONIK 1211009433

KANDHAN SEKAR 1209331327

PRIYADHARSHINI GUNARAJ 1211075356

**Problem 1**

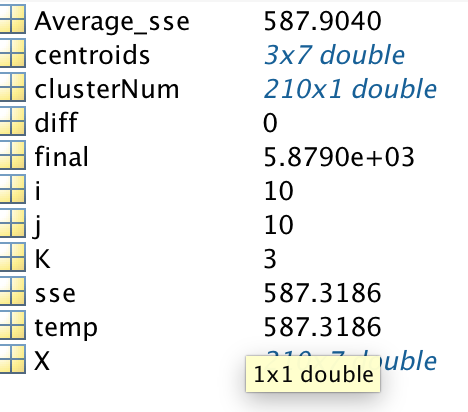
Implemented the K-MEANS clustering algorithm on the ‘seeds’ dataset using Euclidean Distance Metric with user defined ‘K’ value.

First, the function ‘initialCentroids’ assigns ‘K’ random centroids using ‘randperm’ function to generate random points from the records in the dataset. Next, the function ‘closestCentroids’ finds the nearest cluster of each data point using Euclidean Distance. Now, the Sum of Squared Error is computed by the function ‘computeSSE’. Finally, to get the new centroids for each cluster, ‘computeNewCentroids’ is used which computes new centroid to be the mean sum of data points of that cluster.

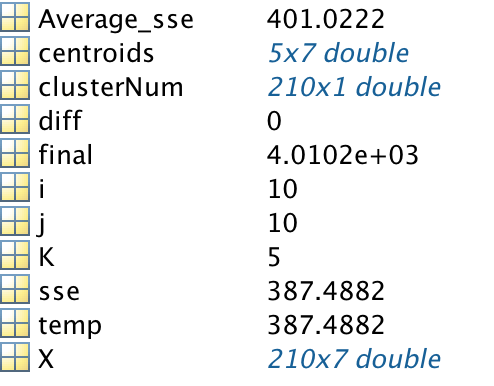
The algorithm runs for 100 iterations calculating new centroids and nearest clusters of all data points and stops. The algorithm also stops when the SSE for 2 successive iterations in <= 0.001.

The Average SSE value when,

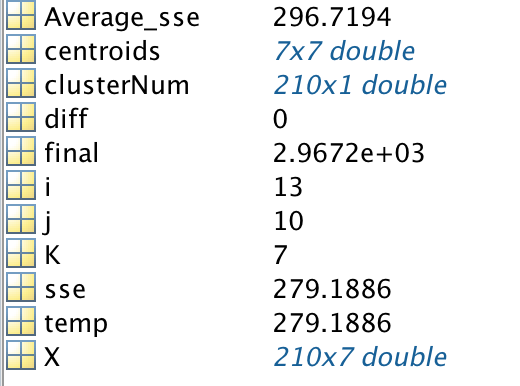
1. “K=3” is “**587.9040**”



1. “k=5” is between “**385**” and “**410**”



1. “k=7” is between “**280”** and “**310**”



**Problem 2**

Implemented the active learning strategies (i)Random Sampling (ii)Uncertainty Based Sampling and found the accuracy/performance of each of them in ‘MindReading’ and ‘MMI’ datasets.

The first step is to train the LR classifier with the training matrices. Then, “ACCURACY” is got by ‘test\_LR\_classifier’ using test matrix. Now ‘k’ (k=10) values of unlabeled matrix is moved to training matrix and starts from first step again. This is done for ‘N’ (N=50) times for each of the 3 matrices of both datasets.

There are 2 active learning strategies mentioned before to choose the ‘k’ values. (i)Random Sampling is done by randomly choosing ‘k’ values from the data samples of the unlabeled matrix using ‘randperm’ function [‘random’ function called from ‘main’ does this]. (ii)Uncertainty Based Sampling is done by finding entropy of the unlabeled matrix using its trained weights got by ‘train\_LR\_classifier’ and getting the ‘k’ highest entropy values by sorting [‘uncertainity\_based’ function called from ‘main’ does this].

‘findAccuracy’ function calculates the ACCURACY of test set using ‘test\_LR\_classifier’ from trained weights got by ‘train\_LR\_classifier’. ‘findEntropy’ function is used to get the ENTROPY of the unlabeled matrix, where entropy e = - Σ pi log (pi) and pi is probability got from ‘test\_LR\_classifier’ of unlabeled matrix.

The Accuracy we obtained are,

1. ‘MindReading’:

### “Random Sampling”: starts from “**43.2385”** increases and reaches “**68.9389**”.

### “Uncertainty Based Sampling”: starts from “**43.2385”** increases and reaches “**72.7553**”.

1. ‘MMI’:

### “Random Sampling”: starts from “**60.0187**” increases and reaches “**93.9122**”.

### “Uncertainty Based Sampling”: starts from “**60.0187**” increases and reaches “**95.5929**”.

From the above results, it’s clear that Uncertainty Based Sampling gives better accuracy than Random Sampling. The screenshot below shows the two figures where ‘Figure1’ is for ‘MindReading’ dataset and ‘Figure2’ is for ‘MMI’ dataset. In both figures, ‘blue’ curve represents Random Sampling and ‘red’ curve represents Uncertainty Based Sampling.

