CSE 572: Data Mining
Spring 2017
Assignment 4 / Mini Project 2

Team members:-

1. Saurabh Jagdhane (1209572595)
2. Himanshu Tyagi (1209283279)
3. Rachita Gupta (1209247724)

Problem 1

To solve this problem, we define our own kmeans function. In this function we start with randomly initializing k centroids using randsample(N,k). This function selects k random data points from N data points. Then we use *pdist2* function. Using this function we can calculate distance between all the data points and the selected centroids. Using the *min* function, for every data sample we find the cluster to which it belongs.
Then centroids are re-initialized based on the average of the k clusters formed, this is done using *mean* function. For every iterations sum of squared errors is calculated by summing the squared distance of each data sample in a cluster from its centroid. This reinitialization is done 100 times or difference of 2 SSE's is greater than 0.001.

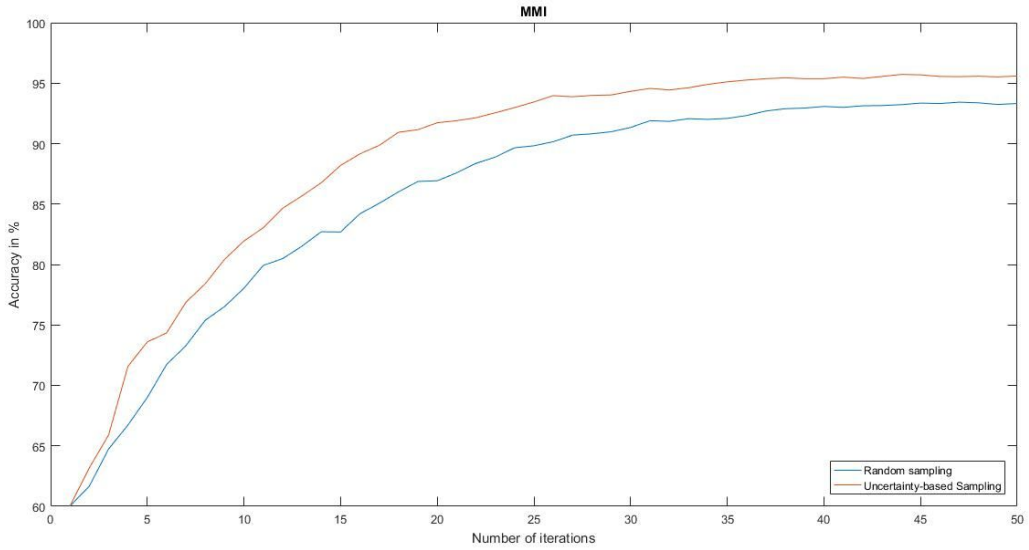
This is done for three values of K= 3,5,7 and last SSE is found out in each case. And average for each K is reported below and it may vary each time.

```
Random initialization of centroids: 10 times
Average SSE for k=3 is 588.0503
Average SSE for k=5 is 409.6588
Average SSE for k=7 is 307.3583
```

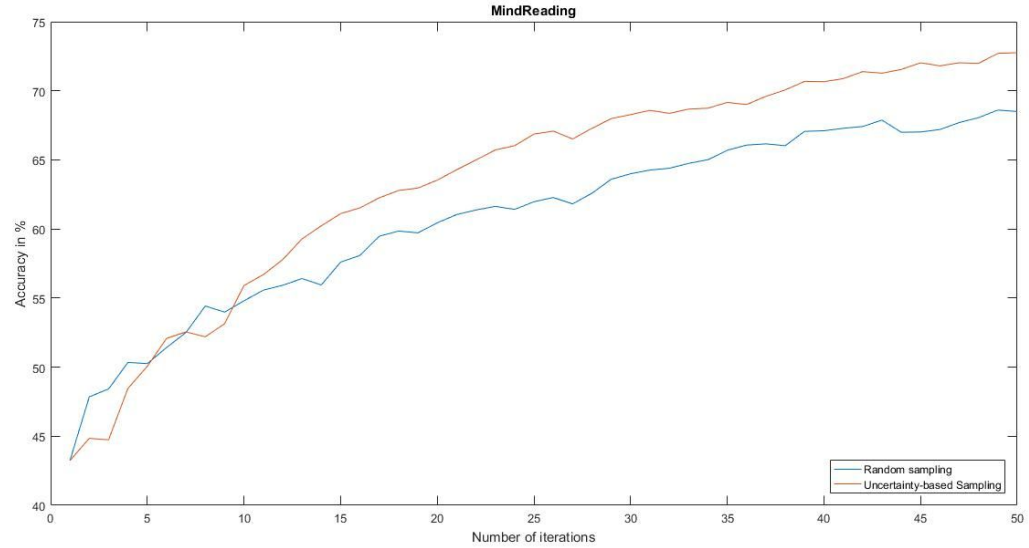
Problem 2

This problem can be solved using many Active learning strategies but two of them are used in here. 1) Random Sampling, 2) Uncertainty based sampling. Two user defined functions does the job of random sampling (myActiveLearningRandom.m) and Uncertainty based sampling (myActiveLearningUncertaintyBased.m). This user defined function replicates the pseudocode for N=50 iterations and k=10 batch size. This process is done for 2 datasets: MMI and MindReading having 3 initial training sets. The average accuracy (accuracy_vector) obtained from these 3 sets is used to plot the results for both the strategies. Each graph is a plot of accuracy_vector against the number of iterations. Results are as follows:

1. Performance of Random Sampling and Uncertainty-based Sampling [**MMI**] :



2. Performance of Random Sampling and Uncertainty-based Sampling [**MindReading**] :



Conclusion:

An active learning algorithm is considered to be better than another if the accuracy growth is at faster rate. Therefore, for both the given datasets it is observed that as the number of iterations goes on increasing Uncertainty based sampling strategy can perform better compared to Random Sampling.

Functions used:

1. randsample(): Returns a vector sampled uniformly at random without replacement.
2. pdist2(): Pairwise distance between 2-sets of observation.
3. min(): Minimum distance for determining the cluster to which sample belongs.
4. sumsqr(): Sum of squared elements of matrix.
5. mean(): Average of an array elements.
6. reshape(): To generate iteration vector with values 1 to 50.
7. datasample(): Gives certain number of random samples and its row (index) number from data with or without replacement as specified option.
8. max(): Returns the maximum element and its index from row or column vector.
9. sort(): Used to sort the matrix and returns indices of sorted elements along with sorted matrix in ascending or descending order as Name,Value specified.
10. removerows(): Processes matrix by removing rows with the specified indices.