CSCI479/679 Introduction to Data Mining

Assignment #3: kMmeans Clustering

Deadline: Tuesday, October, 17th at 11:59PM

The objectives of this project are:

1. Understanding the steps involved in the kMeans clustering.
2. Learning the limitations of kMeans clustering on non-globular shape clusters.
3. Investigating how random initialization of kMeans affects the quality of the clustering.

1 Description

In this assignment you will implement the kMeans algorithm. You will write an R or Python script called kmeans.R/.py that implements the kMeans clustering algorithm. You should your implementation on several datasets (twoCircles.txt, twoEllipses.txt, fourCircles, t4.8k.txt, and iris.txt). One of the datasets is the well known iris.txt dataset that has 3 clusters. Your goal is to discover the three clusters, corresponding to the three species of Iris in the dataset. For the other datasets, plot the points to see how many clusters (k) you should ask for.

For initialization, use random k points as the initial clusters' centers. For convergence testing, you can compare the difference between the sum of the squared euclidean distances between the old means and the new means is less than a threshold. If we define δ to be the sum as follows:

$$\delta = \sum_{i=1}^{K} ||\mu_i^{t+1} - \mu_i^t||^2$$

Then, if $\delta \leq \epsilon$, then you may stop. You may assume that $\epsilon = 0.001$. Another stopping condition is that the number of iterations should be less than a pre-defined number, you may set the maximum number of iterations to 20. In summary, you can stop iterating when meeting any of the criteria above.

1.1 Output:

Your program output should consist of the following information:

1. The final mean and size for each cluster.
2. Final cluster assignment of all the points, which cluster each point is assigned to. For example, if we have 7 points and two clusters with, $C_1 = \{p_1, p_4, p_5\}$ and $C_2 = \{p_2, p_3, p_6, p_7\}$, the output should look like, $1, 2, 2, 1, 1, 2, 2$.
3. Number of iterations, the final δ.

2 Hints:

1. Follow Algorithm 13.1 in the book.
2. To choose k random centers, use the sample function from in R.

3 What to turn in:

1. A report (in pdf) that includes the output (image, mean and size of each cluster, number of iterations, δ) on each of the datasets. You should also submit your code the Python, Matlab, Octave, or R code.

Submission: You should submit your code along with the report to the black board. The file should be named in the following format, useridAssig3.zip, useridAssig3.tar, or .tgz. Late submission will get a 10% penalty for every late day.

Good Luck!