Assignment 4 Unsupervised learning

Due Nov 24st in class

K-means Clustering

In this assignment you will apply k-means clustering to the provided two data sets. The first data set "cluster.csv" contains the data that needs to be clustered. The second data set "random.csv" contains a randomly distributed reference data whose value ranges match those of the "cluster.csv" data.

Note that the data is in the csv format, so when using weka to open the file you need to change the "files of type" option to "CSV data files", which stands for comma separated values.

You need to perform the following experiments.

1. Cluster the cluster.csv data using "simplekmeans" under the cluster tab, with k set to 1, 2, 3, 4, 5 respectively. The cluster mode should be set to using training set. With each clustering run, you can obtain the "within cluster sum of squared error" for that run in the output.
2. For each k value, we run kmeans for **five** times, each time set the ***seed*** value to a different value (please use 1, 2, 3, 4, 5). Among the five runs, choose the one that gives the ***lowest*** "Within cluster sum of squared error" and record the value. (Note that for k=1, no clustering is run, just record the value once will do.) --- **This step should give you a set of values we denote $WCSE_1$, $WCSE_2$, $WCSE_3$,$WCSE_4$,$WCSE_5$ for the cluster.csv data**
3. Plot the values of $WCSE_i$ for i =1, 2, 3, 4, 5. Can you tell the number of clusters from this plot?
4. Repeat steps 1 and 2 for the "random.csv" data and obtain the "within cluster squared error" for k =1, 2, 3, 4, 5 respectively. Let's denote them by $WCSE'_1$, $WCSE'_2$, $WCSE'_3$,$WCSE'_4$,$WCSE'_5$.
5. Plot the values of $WCSE_i/WCSE'_i$, for i =1, 2, 3, 4, 5. From this plot, how many clusters do you think this data contains?

HAC clustering

Create by hand the clustering dendrogram for the following samples of ten points in one dimension.

Sample = (-1.8, -1.7, -0.3, 0.1, 0.2, 0.4, 1.6, 1.7, 1.9, 2.0)

a. Using single link.
b. Using complete link

Note that below are example dendrograms for the following three points (1, 2, 4) with single and complete link