RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Laboratory Week 4

Aims of this lab

- Learn how run the Kmeans and EM clustering algorithms and interpret the results.
- Learn that the interpretation of clustering is not necessarily straight forward and requires some judgment.

---

1. You will need to have access to the WEKA package.
2. The data files for this lab can be found at
   `/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data`
3. Load the file `arff/mystery-data5.arff`.
   (a) Go to the Visualize screen. How many clusters are in the data?
   (b) Go to the Cluster screen and select SimpleKMeans.
   (c) Run the algorithm for K=2 and analyse the output.
   (d) Visualise the clusters by a left click on the file in the result list.
   (e) Does this appear to be a good clustering result?
   (f) Repeat the runs for 5 different initial centres. This done by changing the seed (try 10 and 11)and describe the effect of changing the initial centres..
   (g) Repeat the runs and visualizations with K = 3,4,5,10,20.
   (h) How can you tell when you have right number of clusters?
   (i) Does the "Within cluster sum of squared errors" give any indication about the number of clusters?
4. Restart Weka and reload the file `arff/mystery-data5.arff`.
   (a) Go to the Cluster screen and select EM.
   (b) Run the algorithm with K=2,3,4,5 and compare the clusters to the ones found by Kmeans. Are they the same? Would you expect them to be the same?
   (c) Run the EM with N (numClusters) set to -1. Visualise the clusters.
   (d) Does this appear to be a good clustering result?
   (e) How does it compare with the K-Means results?

(f) The clusters generated by EM can be quite sensitive to the values of *minStdDev*. Explore the effect of different values for this parameters [To start with, try order of magnitude changes, ie `minStdDev 1.0E-6 --> 1.0E-5 --> .... --> .01 --> .1 -->1 -->10`] Summarize your observations.

(g) The clusters generated by EM can also be quite sensitive to the values of *minLogLikelihoodImprovementIterating, minLogLikelihoodImprovementCV*. Explore the effect of different values for these parameters. Summarise your observations.

5. Load the file `arff/student-data1-small.arff`.

   (a) Run the EM algorithm with default parameters on this file. Visualize the output. What do you find?

   (b) Adjust the parameters to get right number of clusters.

   (c) Give English language descriptions of the clusters.

6. Load the file `arff/student-data1.arff` and run the EM algorithm. What do you notice about the execution speed on this file? Can anything be done?

7. Load the file `arff/UCI/iris.arff`. and remove the class attribute. Run EM with the default parameters. How many clusters are generated? It is known that there are three classes in this data. Why aren't there 3 clusters?