TDT4300 Datavarehus og datagruvedrift
Spring 2014
Assignment 5: Cluster Validation

1 K-Means and Cluster Evaluation

You are provided two datasets in arff format. Use at least these two datasets (*iris.arff* and *segment-challenge.arff*) and show the effects of a variation of k.

You are further provided with some code that reads datasets from the arff format into a plain double array representation and normalizes them by unit length. Note that I use vector and datapoint as equivalent. This double array is used for the similarity computations later on. Each cluster stores the indices of the data points assigned to it.

You should implement:

- *KMeans.recalculateClusters()* to perform classic k-means cluster calculation
- *Euclidian.distance(double[] vector1, double[] vector2)* to compute the euclidean distance between two data points
- *Cluster.SSE()* to compute the sum of the squared error of a cluster (also show how this method can be used to compute an overall SSE for a given clustering in *KMeans.getSSE()*)
- *Cluster.SSE()* - implement the separation based measure introduced in the lecture (also *KMeans.SSB()* for the whole clustering)
- *KMeans.getAverageSilhouetteValue()* to compute the average Silhouette value of all points

Notes on Implementation In the implementation you're given, each cluster stores information about what datapoints are associated with it in terms of indices. This means it only stores a list of indices in the overall *data[][]* variable which is a field of *KMeans*. Use *KMeansClustererMain* as a starting point for working with the code.

2 Evaluation

Run the clustering with a varying k (from $k = 2$ to $k = 8$). Keep record of the SSE. Produce a plot of k versus SSE (probably Excel is the easiest way to get this done, other solutions include Matlab or R).

Describe how this can be used to determine an appropriate level of k, and show what k should be in this case (sammenlikne med "knekk" som ble vist i forelesningen). Also show the development of SSB and the average silhouette value. Describe the relation between SSB, SSE, and the silhouette value. What do the measures show and when are they best applicable?

Notes

Your submission in its learning is your report in form of a **pdf** file and the source code of the four methods you should implement. In the report you should provide the graphs for both data sets and results for the evaluation measures.