CSI445/660 – Data Mining Assignment #1: K-Means Clustering

Due: 03/15/05 at the start of class
One full grade will be deducted per week (including partial weeks) delay

Marks: 25% of final grade

Clustering is one of the most used data mining tasks with the K-Means algorithm being arguably the most popular and is available in most (if not all) data mining suites/packages. However, its run-time performance is \( O(kmnI) \) for any one application and the basic implementation in these packages can lead to misleading results. In this assignment you will implement the basic algorithm to learn some of the subtleties of the algorithm’s behavior as well as several common techniques to speed-up the algorithm.

Question 1): Basic implementation (25 points)

Implement the K-Means algorithm (any language will be fine except assembler) and provide the pseudo code for your algorithm. We shall refer to this implementation as K-Means-Batch. Apply the algorithm on the three continuous data sets listed on the web page (note data sets to be used depend on surname). For each data set use Euclidean distances, \( k=4 \) and terminate the algorithm under the condition \( z_i^t = z_i^{t-1} \forall i \).

Report:

1) A plot of the vector quantization error against iteration number.
2) The average and variance of the number of iterations until convergence for five random restarts of the algorithm and the vector quantization error.


A common variation of K-Means-Batch is to recalculate the centroids after each data point assignment, this is known as on-line/incremental K-Means. Change your K-Means code to implement this change. For each data set use Euclidean distances, \( k=4 \) and terminate the algorithm under the condition \( z_i^t = z_i^{t-1} \forall i \). Report:

3) A plot of the vector quantization error against iteration number.
4) The average and variance of the number of iterations until convergence for five random restarts of the algorithm and the vector quantization error.
5) Did incremental K-Means improve performance? If so why?

Question 3): Determining the value of K (15 points)

Each possible set of centroids for a particular value of \( K \) is considered to be a model and all models for a particular value of \( K \) is considered a model class. Determining the best value of \( K \) is then a specific version of the general problem of selecting the best model class. One of the most common method of choosing the best model class is the Bayesian information criterion (BIC) which states the best model class is the value of \( k \) that maximizes: \(-\log(P(Data | BestModel)) – (No. Parameters in Model)\times \log(Number of Data Points)/2\). We can see that this expression trades off model fit and model complexity.

Report:

6) How you shall calculate \( P(Data | Model) \)
7) Plot for just one data set the BIC value against the values of \( k=1 \) through to 10.

Question 4): Clustering Under Constraints (35 points)

In typical applications of clustering we have little background knowledge. Consider the case where we have background knowledge in the form that: a) Two instances must be in the same cluster, b) Two
instances must not be in the same cluster. This situation may arise where we have partially labeled data but the majority of data is unlabeled. We shall investigate the performance of these constraints on the algorithm convergence and performance at minimizing the vector quantization error. Examples of constraints are given on the web sites. For each data set use Euclidean distances, k=2,4,6 and terminate the algorithm under the condition $z^t_i = z^{t-1}_i, \forall i$. Report:

8) A plot of the vector quantization error against iteration number for each value of k
9) The average and variance of the number of iterations until convergence for five random restarts of the algorithm and the vector quantization error for each value of k.
10) Suggest a principled method of generating the must-link and cannot-link constraints.