K-Mean Clustering

Ravi Kumar
14MCMI12
SCIS, University of Hyderabad

March 12, 2015

1 Introduction

Clustering is the process of partitioning a group of data points into a small number of clusters. For instance, the items in a supermarket are clustered in categories (butter, cheese and milk are grouped in dairy products). Of course this is a qualitative kind of partitioning. A quantitative approach would be to measure certain features of the products, say percentage of milk and others, and products with high percentage of milk would be grouped together. In general, we have n data points $x_i i = 1...n$ that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions $\mu_i, i = 1...k$ of the clusters that minimize the distance from the data points to the cluster. K-means clustering solves

$$arg_c min \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|^2$$

where c_i is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance $\|x - \mu_i\|^2$. This problem is not trivial (in fact it is NP-hard), so the K-means algorithm only hopes to find the global minimum, possibly getting stuck in a different solution.

k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters
2. Attribute the closest cluster to each data point
3. Set the position of each cluster to the mean of all data points belonging to that cluster
4. Repeat steps 2-3 until convergence

The algorithm eventually converges to a point, although it is not necessarily the minimum of the sum of squares. That is because the problem is non-convex and the algorithm is just a heuristic, converging to a local minimum. The algorithm stops when the assignments do not change from one iteration to the next.

2 Modules

2.1 Preprocessing

The data is read from a file which is in csv format, then the labels are removed and the data is divided into two. Training(80%) and Testing(20%).

2.2 Initialization of centres

The initial value of centroid is assigned by taking random number of data points from dataset.

2.3 Distance Matrix

This modules returns a matrix which is a distance matrix which contains the euclidean distance between data points and the centroid.

2.4 Group

This modules actual contains the classification of data points. The group matrix is having number of rows as cluster number and column numbers as data points. If a data point is present in a cluster its value is kept as 1 else 0.

2.5 k-means

This is main k-means algorithm. This modules iterate over a period of time to calculate the centroid and classify the data using the other modules.

2.6 New centroid

This modules calculates the new centroid value from the distance matrix and group(classification) matrix.

2.7 Error

The error is measured as the ration between number of unsuccessful classification to the total number of instance given.

3 Contribution

My contribution to this Assignment is the implementation of new centroid function and group function.

4 Experimental Result

This code is running comparatively slow with C code since it is written on high level language (R), which uses very high level constructs. Hence the code is tested on smaller data sets which provides 70% error rate. The error rate is also depend upon the type of data set that is involved. The given data set was around uniformly distributed across different classes.

5 Conclusion

This is to conclude that the k-means algorithm has been coded and tested on 1,2 data sets giving average results on performance also. This code can be improved further.