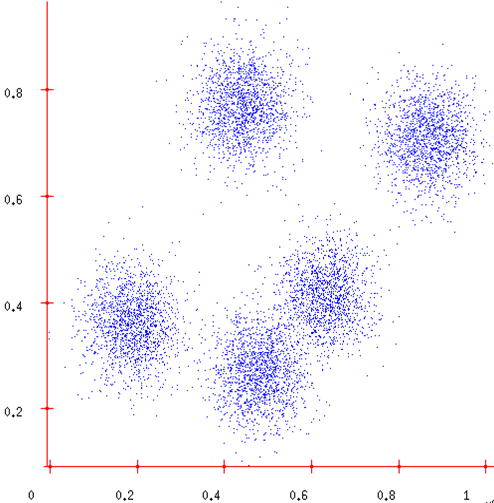
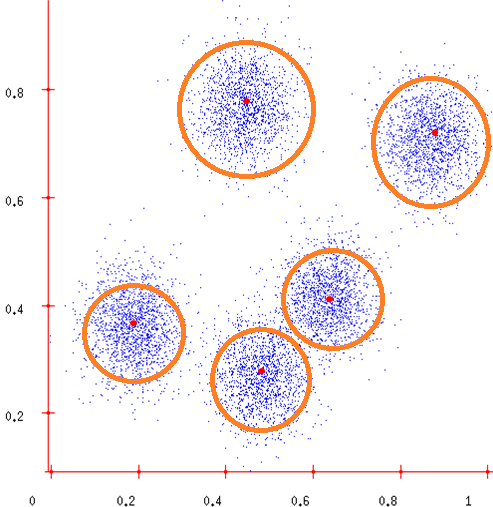
Review for Data Clustering

1. What is clustering ？

Clustering is a way of grouping together data samples that are similar in some way-according to some criteria that we pick. So it’s a method of data exploration, a way of looking for patterns of structure in the data that are of interest. Clustering can be considered the most important unsupervisedlearning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

We can show this with a simple graphical example:

Before clustering after clustering

In this case we easily identify the 5 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is the simplest clustering, it base on the distance. More generally, we cluster our data in conceptually: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.

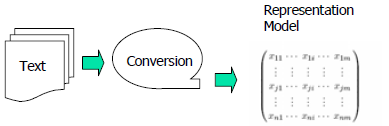
1. Clustering Algorithms—K-means

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

So K-means is a method of [clustering analysis](http://en.wikipedia.org/wiki/Cluster_analysis) which aims to [partition](http://en.wikipedia.org/wiki/Partition_of_a_set) n observations into k clusters in which each observation belongs to the cluster with the nearest [mean](http://en.wikipedia.org/wiki/Mean). This results in a partitioning of the data space into [Voronoi cells](http://en.wikipedia.org/wiki/Voronoi_cell).

* 1. Preparation for K-means

Some preparations need to be done before the K-means algorithm. The data we used to clustering by K-means algorithm might not be structured, such as text, volume, or picture. We need to convert the data and build corresponding mathematical model. Text for example, we need to do something for preprocessing, such as splitting, deleting and Stemming. Most importantly, we need to build a corresponding mathematical model; we can convert our data to math model-Vector Space Model (VSM), as shown in the picture below.



In order to build our VSM model, we should calculate the weighting terms for each keyword to weight the frequency and importance of keyword in our data. The tf, idf parameter is needed in that step, it calculated as follows:



Where tf (dj, ti) is the frequency of term ti in document di, |D| is the total number of documents, and df (ti) is the number of documents in which ti occurs.

* 1. K-means Algorithm

K-means Algorithm is carried out in four steps:

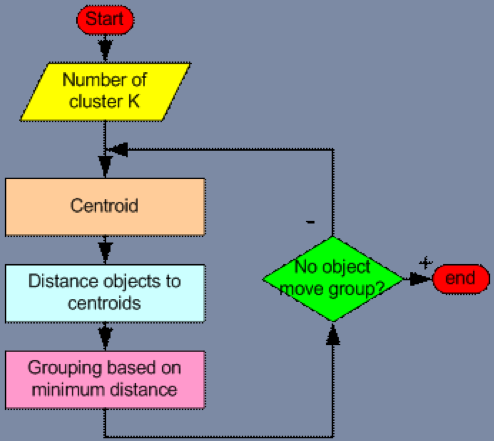
Firstly, given the cluster number k and set seed points, this is the initialization of algorithm;

Secondly, assign each object to the cluster with the nearest seed point;

Thirdly, compute seed points as the centroids of the clusters of current partition (the centroid is center, i.e. mean point of the cluster);

Finally, go back to step 2, stop when no more new assignment.

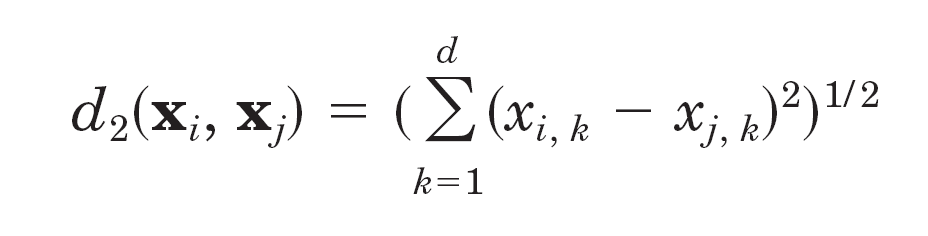
K-means Algorithm flow chart is as follows：



Now let’s think in detail about K-means Algorithm.

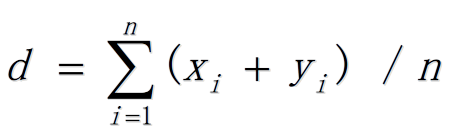
In the first step of K-means Algorithm, there are two points we should take in to consideration: 1, how to determine the K, the number of clusters, in K-means Algorithm, the number of k is not easy to choose unless we have preview of our data, but actually we do not know the detail of data before clustering, so if the value of k is given in advance, we can easy to implement the algorithm, otherwise we need several attempts; 2, how to select the initial seed points. In most cases, we get the initial seed points by randomly, but we can make a little improvement, random select the points that are already exist.

In the second step, in order to assign each object to the cluster with the nearest seed point, we need calculate the distance of two points, algorithm is as follows:

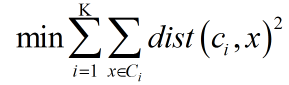


For each of object point, we calculate the distance between it and each of centroid point, and then find the minimum distance, thus the object point is assign to the minimum distance centroid point.

In the third step, we already had clustered our data in to K clusters, at this step we should recomputed seed points as the centroids of the clusters of current partition. In order to calculate the new centroids, we can use the following formula to calculate the central value of each dimension, and then we can get the new efficiently centroids.



In the last step, when no more new assignment, algorithm should be stopped, so we need to find a way to determine when we should stop. So the following formula is used to determine when to stop. We store an old dist of all cluster, and for each iteration of the loop, we recalculate the new dist. When the difference between the old and the new is less than 1, algorithm can be stopped.



1. Comments on the K-means

Each algorithm has its own advantages and limitations; K-means algorithm is no exception.

Strength:

* Relatively efficient: *O*(*tkn*), where *n* is instances, k is clusters, and *t* is iterations. Normally, *k*, *t* << *n*;
* Often terminates at a local optimum. The *global optimum* may be found using techniques such as: simulated annealing or genetic algorithms.

Weakness:

* Applicable only when *mean* is defined;
* Need to specify k*,* the number of clusters, in advance;
* Unable to handle noisy data and outliers;
* Not suitable to discover clusters with non-convex shapes.