CM50246 Machine learning task I: Unsupervised Learning

Name: Su Chen

Student Number: 169410920

Course ID: **CM50246**

Introduction

Clustering algorithm is an important part of machine learning. It is the unsupervised classification of objects into different groups according to their similar features (Jain, Murty and Flynn, 1999). Human brains have good abilities to divide object into groups. For example, biologists set natural world different categories such as plant and animal. This operation is called *clustering*. K-Means clustering is an algorithm whose aim is to divide M points in N dimensions into K clusters, each point will belong to the nearest mean. In K-Means clustering we just assign point to the nearest cluster by calculating the Euclidean distance which is not a good metric. Gaussian Mixture Model can be regarded as a type of clustering algorithm. The Gaussian Mixture Model uses Expectation Maximization and assume every cluster has an independent Gaussian distribution, mean and covariance matrix. At the meantime, Gaussian Mixture Model can be used to make density estimation. This report will discuss two main cluster method: K-means method and Gaussian Mixture Model.

The result of both algorithm will be compared and the performance of both algorithm will be evaluated.

1.K-Means Clustering

Method

The input of K-Means clustering including a matrix of M points and N dimensions, in this report the K-means clustering will proceed in 3 dimensions.

K-means would minimize the sum of distance from every data point to their cluster mean values, which are the centroids of clusters.

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \tag{1}$$

Variable r_{nk} is 1 when data point n is divided in to cluster k, or its value is 0. Variable μ_k is the centroid of every cluster, its value is the mean of data points belong to this cluster. It is difficult to find r_{nk} and μ_k to minimize J. The method is use iteration, fix the centroid μ_k and chose the data points belong to it, then change the centroid to minimize J. When J is the minimum:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{2}$$

Blow is the step of demonstrating the K-Means Clustering algorithm:

1) Choose K initial random centroids μ_k, the random mean can make influence on the final result so that we choose different initial value can chose the best result.
2) Divide every data point to the cluster represented by the nearest centroids.
3) Calculate the new means to be the centroids in the new clusters.
4) Repeat the second and third step until the largest step or the J reaches the minimum value.

Figure 1 is the flow figure of K-means clustering:

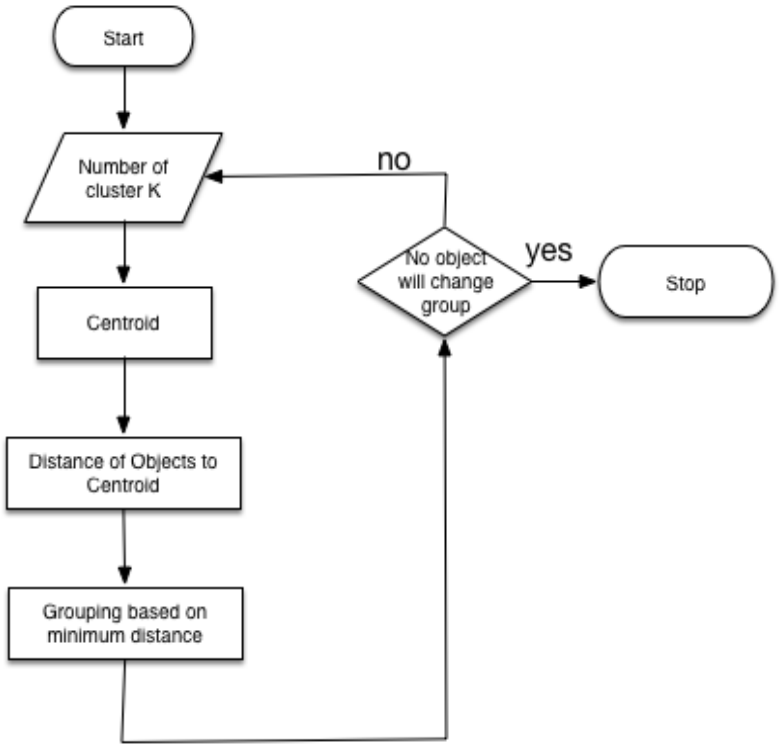

Figure1: The flow figure of k-means clustering.

Result and discussion

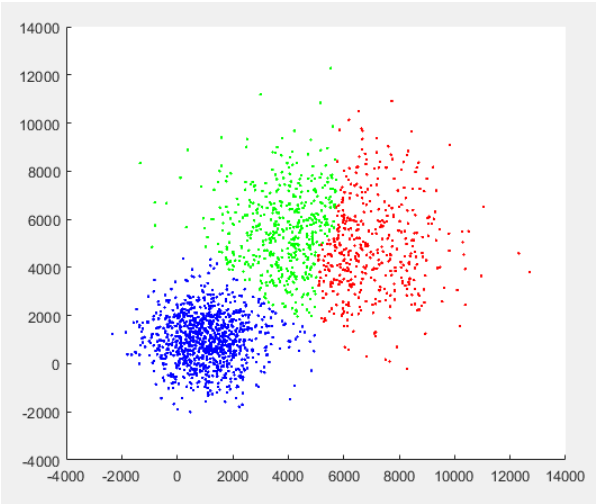

Figure2:The result of k-means clustering, in which we random 3000 data points which value is 0-100, and we divide the data points into 3 clusters. This result changes when we choose different data points, the time of the algorithm is 0.249518 seconds.

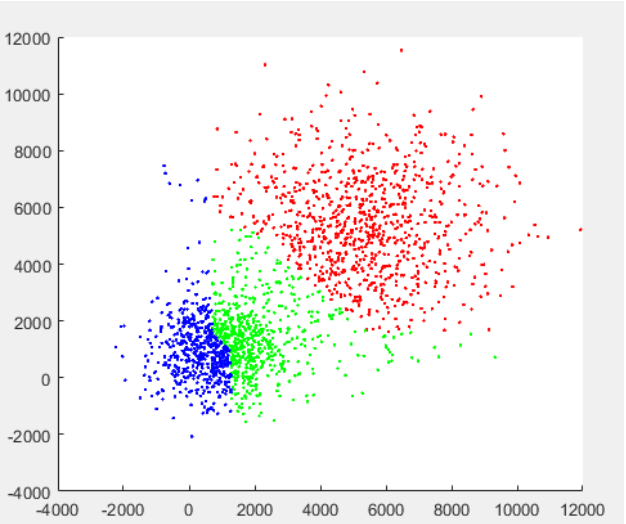

Figure3 : The result of K-means in which the value of iteration is 3.

K-means clustering works well when dataset is suitable for this algorithm, but when k-means clustering use some datasets such as iris flower data set when k=3 the results fails to separate three iris species in the data set.

The advantages of K-means are:

1) The k-means has a faster computation speed.
2) K-means produce tight clusters.

The disadvantages of K-means are:

1) It is difficult to predict K value, in the implement we input k value but it is not always a perfect value for the clustering.
2) K-means does not work well for datasets of different size and different density.
3) K-means algorithm comes out different result with different initial data point, it is hard to choose the best result.

2.Gaussian Mixture Module

Method

We assume that the value of Gaussian Mixture Module follow the Mixture Gaussian Distribution. Every Gaussian Mixture Module consists of K Gaussian Distribution, the Gaussian distributions are called Component. Sum all the component will come out the probability density function:

$$\begin{aligned} p(x) &= \sum_{k=1}^{K} p(k)p(x|k) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned} \tag{3}$$

π_k is the probability of being divided into the cluster.

The log-likehood function of GMM is:

$$\sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right\} \tag{4}$$

Assume the probability of the data point belongs to a component

$$\gamma(i,k) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)} \tag{5}$$

Calculate the parameter of components. The components are standard Gaussian distribute so the parameter can be calculated by the Expectation-maximization

$$\begin{aligned} \mu_k &= \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i \\ \Sigma_k &= \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T \end{aligned} \tag{6}$$

Repeat iteration until convergence.

Result and discussion

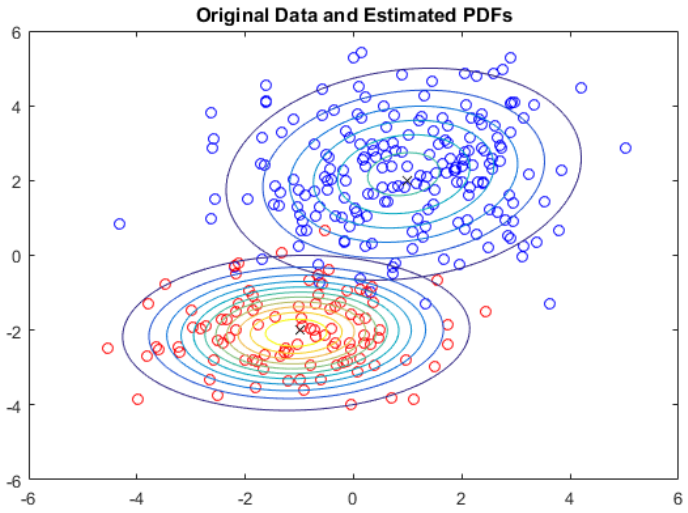

Figure4: The result of GMM, we random 300 data points and divide them into two groups. The number of iteration is 221.

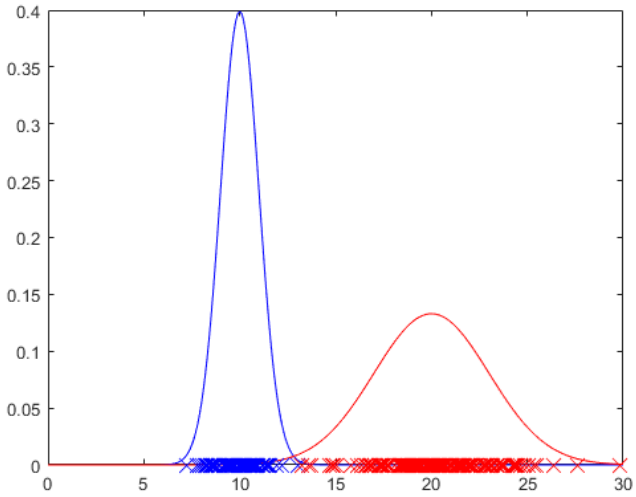

Figure5: The result in 2 dimensions.

The Gaussian mixture model is the fastest algorithm for learning mixture models but still has some disadvantage:

1) Estimation the covariance matrices will be difficult when data set does not have sufficient points per mixture.
2) Gaussian mixture model need held out data to decide how many components to use.

3. Evaluation and Comparison

Both method use iteration and the strategy of iteration is the same, and at the beginning of algorithm they will initialize the data point and implement two steps and repeat the steps. The differences are: in k-means clustering the parameter to be calculated is the position of the centroids, in Gaussian mixture module the parameter to be calculated is the Gaussian distribution of different component. And the method of calculation is different, in k-means clustering we calculate the mean of data points in same cluster, while in Gaussian mixture module we calculate the maximum likelihood.

If the k and dimension in k-means algorithm is fixed , the problem can be solved in time $O(n^{dk+1} \log n)$, time per iteration and in the test case it took 0.249518 seconds. The time of GMM solve in time O(kn) . K-means has a faster speed than GMM. GMM not just divide data point to k initial centres, it gives every point's probability of belonging to a cluster. GMM is more useful in application.

Reference

David J.C. MacKay(2003), Information Theory, Inference, and Learning Algorithms

Kanungo T et al(2002), An Efficient k-Means Clustering Algorithm:
Analysis and Implementation, EEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002

Biernacki C, Celeux G, Govaert G(2000), Assessing a Mixture Model for Clustering
with the Integrated Completed Likelihood, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 7, JULY 2000