**Machine Learning Assignment One**

**Name：XINTAO LU**

One、The Comprehension Of Algorithms

1. EM Algorithm

(1)、Description

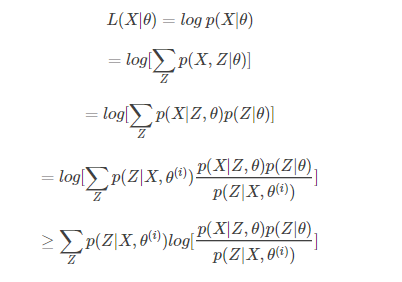
The full name of the EM algorithm is called Expectation Maximization, which is the expectation maximization algorithm. It is a frequently used algorithm in the model with hidden variables, which is aimed at the unsupervised learning problem of hidden variables. It is involved in the following two algorithms called GMM and K-means mentioned below. Briefly speaking, GMM and K-means are two-case implementations of the EM.

(2)、The problem description and solution process Of EM

Problem definition: Define an observable variable X. The observable samples of n obtained for X are {x1, x2, x3, ..., xn}, and find the hidden variables Z, P(Z|X, θ) and P(X|Z, θ) for each sample.

Solving: The optimal solution Z, θ under the loss function L(X|θ).

Simple derivation process:



（3）、Explanation

Among above, the process of maximizing the result is the procedure of solving the optimal solution, it is obvious that each time is the maximum lower limit, and the lower limit is . The step of maximizing the lower limit is called the M（maximization）, and the step of solving is called the E(exception).

1. K-means Algorithm

(1)、Description

The K-means algorithm is a simple implementation of the clustering algorithm, but the ideas contained inside are more innovative. Clustering is a form of unsupervised learning, while supervised learning (such as support vector machines, etc.) is class-labeled, which means the class of the sample has been given. However, the corresponding classification C is not given in the sample of K-means clustering, only given the characteristic of classification X. The purpose of clustering is to find the potential category C in each sample X and put samples X of the same category C together.

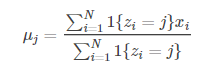
Among the clustering problems, given the training samples , each , without the category label C. The K-means algorithm clusters the samples into K clusters, the specific algorithm is described as follows:

1)、Select K cluster randomly centroid points as , repeat the following process until convergence{

For each sample i, calculate that which classification should belong to.



For each category j, recalculate the centroid of the category.



}

K is the number of clusters given in advance, and zi can take from 1 to K, which represents the closest classification of the samples i and K, and represents the centroid of the same sample. If we want to aggregate all the samples into K classifications, then K points are randomly selected as the centroids of K classifications. The first step calculates the distance to K centroids for each sample, and selects the nearest classification as zi. The second step is to recalculate the centroid for each classification, and iterate the first and second steps until the centroid is constant or the change is small.