# Machine Learning Homework 2

专业：软件工程 姓名：沈金龙 学号：18214806

1. **实验题目**

* Please build a Gaussian mixture model (GMM) to model the data in file TrainingData\_GMM.csv. Note that the data is composed of 4 clusters, and the model should be trained by expectation maximization (EM) algorithm.
* Based on the GMM learned above, assign each training data point into one of 4 different clusters.

1. **实验要求**

* Show how the log-likelihood evolves as the training proceeds.
* The learned mathematical expression for the GMM model after training on the

given dataset.

* Randomly select 500 data points from the given dataset and plot them on a 2-

dimensional coordinate system. Mark the data points coming from the same

cluster (using the results of Problem 2) with the same color.

* Some analyses on the impacts of initialization on the converged values of EM

Algorithm.

* Some analyses on the results you obtained.

1. **实验过程及代码**

* 本实验采用python3.6完成。
* 关于高斯混合模型GMM

高斯混合模型（GMM, Gaussian Mixture Model）是多个高斯模型的线性叠加，高斯混合模型的概率分布可以表示如下：



其中，表示模型的个数，是第个模型的系数，表示出现该模型的概率，是第个高斯模型的概率分布。

注：考虑多个随机变量的情况，即多元高斯分布，因此高斯分布中的参数不再是方差，而是协方差矩阵。

我们的目标是给定一堆没有标签的样本和模型的个数，以此求得混合模型的参数，然后就可以用这个模型来对样本进行聚类。

* 关于期望最大化算法EM

期望最大算法（EM，Expectation-Maximization）是通过不断迭代来求得最佳参数的。在执行该算法之前，需要先给出一个初始化的模型参数。我们让每个模型的为随机值，为单位矩阵，为，即每个模型初始时都是等概率出现的。EM算法可以分为E步和M步。

**E步:**

直观理解就是我们已经知道了样本，那么它是由哪个模型产生的呢？我们这里求的就是：样本来自于第个模型的概率，我们把这个概率称为模型对样本的“责任”，也叫“响应度”，记作，计算公式如下：



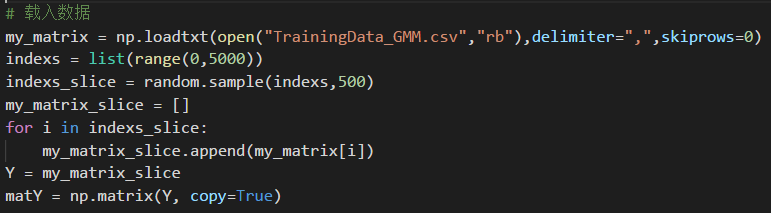
**M步:**

根据样本和当前矩阵重新估计参数，注意这里为列向量：

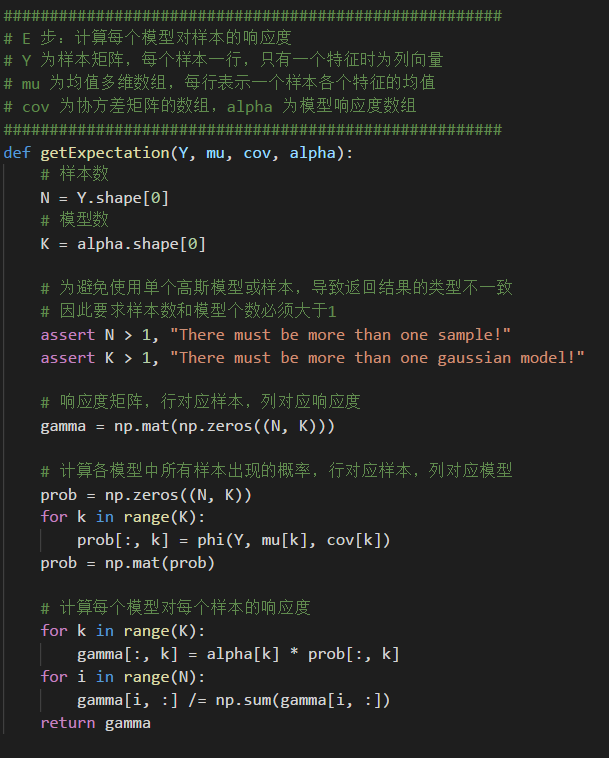


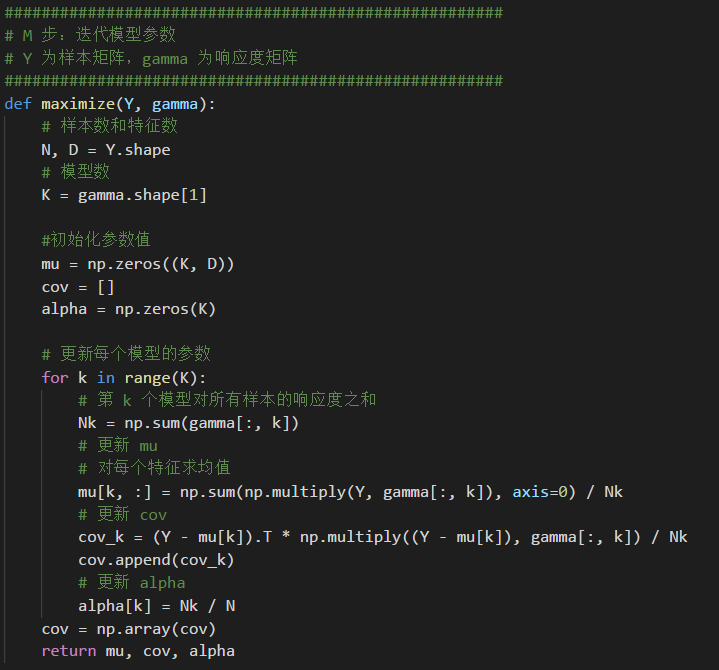
1. **实验结果与分析**

* 随机选择500个点并在二维坐标系中画出来，以同样的颜色标出各自的聚类：

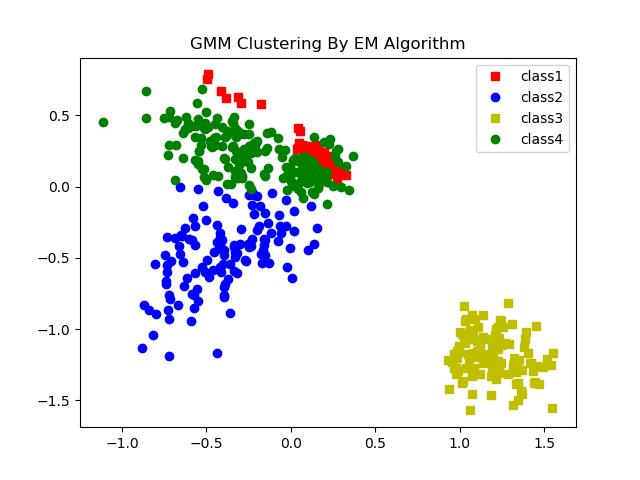


* EM算法的核心代码：





* 聚类结果：



* 基于给定数据得到的相关参数如下所示：

：

[[0.44845318 0.77713784] [0.28080906 0.48586375]

[0.86577615 0.16207885][0.34891429 0.75986815]]

：

[[[ 0.00629149 -0.00577172][-0.00577172 0.00599824]]

[[ 0.0099899 0.00733243][ 0.00733243 0.01512776]]

[[ 0.00319738 -0.00074098][-0.00074098 0.00416401]]

[[ 0.01335553 -0.00431358][-0.00431358 0.0043973 ]]]

：

[0.09262192 0.27225105 0.246 0.38912703]

* 关于EM算法的初始化的讨论：

传统的EM算法对初始值敏感，聚类结果随不同的初始值而波动较大。且随着迭代的次数，越往后收敛速度越慢，同时很容易trap在local解附近。总的来说，EM算法收敛的优劣很大程度上取决于其初始参数。

1. **总结**

本实验详细了解了GMM和EM的相关知识，并完成基于python的实现。在实验中发现，传统的EM算法对初始值敏感，算法收敛的优劣很大程度上取决于其初始参数。据此可能采取的优化措施为：采用一种基于网格的聚类算法来初始化EM算法。