

**2020年春季学期  
计算学部《机器学习》课程**

**K-means聚类方法和混合高斯模型**

|  |  |
| --- | --- |
| 姓名 | 李卓君 |
| 学号 | 1180300210 |
| 班号 | 1803104 |
| 电子邮件 | zhuojunlee724@gmail.com |
| 手机号码 | 18845636557 |

**目录**

[1 实验目的与实验要求 3](#_Toc25993)

[1.1 实验目的 3](#_Toc31750)

[1.2 实验要求 3](#_Toc15413)

[1.3 实验环境 3](#_Toc32353)

[2 实验背景与原理 3](#_Toc18082)

[2.1 实验背景 3](#_Toc15552)

[2.2 实验原理 4](#_Toc27933)

[2.2.1 K-means算法 4](#_Toc3710)

[2.2.2 GMM-EM算法 4](#_Toc2111)

[3 K-Means算法 6](#_Toc21180)

[3.1 生成数据 6](#_Toc6071)

[3.2 具体过程 6](#_Toc14431)

[3.3 GMM-EM算法 8](#_Toc12495)

[3.4 UCI数据集 8](#_Toc31931)

[4 分析以及结论 9](#_Toc27633)

# 实验目的与实验要求

## 实验目的

实现一个k-means算法和混合高斯模型，并且用EM算法估计模型中的参数。

## 实验要求

用高斯分布产生k个高斯分布的数据（不同均值和方差）（其中参数自己设定）。

（1）用k-means聚类，测试效果；

（2）用混合高斯模型和你实现的EM算法估计参数，看看每次迭代后似然值变化情况，考察EM算法是否可以获得正确的结果（与你设定的结果比较）。

## 实验环境

Windows 10, Visual Studio Code, Python 3.8.5

# 实验背景与原理

## 实验背景

在之前的实验中我们进行的都是监督学习，即给定特征和预测值，拟合一个模型来反映二者之间的映射关系。而本次实验我们进行的为非监督学习，即只给定特征，拟合模型找到数据集的内在结构，直白地说，就是给定数据集和聚类数目，通过K-means和GMM-EM算法来将数据集分为类，其中每个类内部聚合程度最高，类与类之间的聚合程度最低。

## 实验原理

为了描述方便，我们从较简单的二维点集的2-Means问题出发。假定有二维点集用矩阵表示，利用算法将该点集中的所有点分为两类。接下来我们着重介绍K-means算法和GMM-EM算法。

### K-means算法

K-means算法的步骤如下：

1. 首先随机地在数据集中选择两个点，称为聚类中心。

2. 分配类别：根据聚类中心将所有的样本分为两个组，分类标准为选择距离样本点最近的聚类中心所在组为我们为该样本点分配的组，距离以欧式距离计，即

为方便计算，将等式两侧都进行平方处理。

3. 移动聚类中心：计算当前得到的两个组分别的中心，

然后将原有的聚类中心设置为新计算得到的中心。

4. 重复(2)(3)直到结果收敛。

这当中值得注意的是第一步初始化步骤，在之前的实验中，我们的初始化为随机的，一般将参数的各维度都置0，通常情况下也不会出错，至多出现迭代次数过多的问题，但在该实验中不确定的元素太多，如果初值选择不合理可能出现聚类结果与设想出现较大偏差。

课堂上介绍了一种方法用来规避这种可能，多次随机化选择，对于每次选择出的k个聚类中心和用欧式距离计算得到的k个聚类，计算其欧式距离之和，选取得数最小的一组聚类中心作为K-means算法迭代的初始值。

### GMM-EM算法

首先给出n维高斯分布的密度函数如下：

再给出混合高斯分布的定义：

可以看出，混合高斯分布由k个混合成分构成，每个混合成分对应一个高斯分布，每个高斯分布前都有一个混合系数，从贝叶斯学派的角度，其意义是先验分布，即。现在假设对于样本集满足高斯混合分布，实际上在后续实验步骤中我们的样本集也是按照高斯分布产生的。根据贝叶斯定理，的后验分布为

这样一来，对于每一个样本我们都可以给定一个标记:

如果给定样本集X可以采用极大似然估计，再进行对数运算，我们有：

为使其最大化，分别对，求导为0有：

对于混合系数，由于其有限制条件，所以使用拉格朗日乘数法，有

对求导并令导数为0有：

等式两边分别乘上并进行加和，得到：

由于，所以有

则有

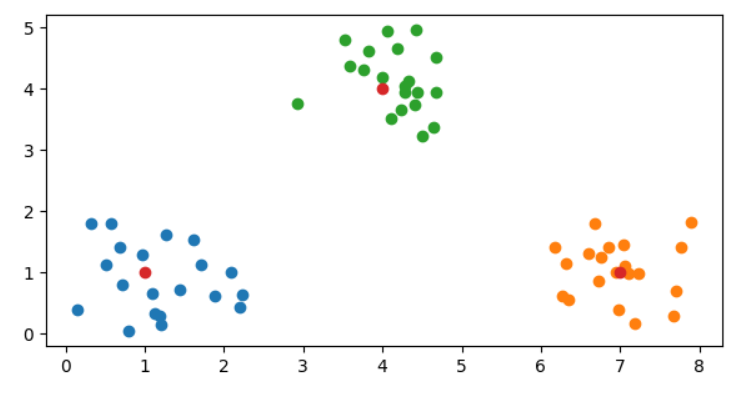
以上就是高斯混合模型(GMM)的原理分析，在GMM问题的求解过程中常用的方法是EM算法：首先根据当前参数计算每个样本属于每个混合高斯成分的后验概率，再根据之前对参数求导得到的公式对参数进行更新。

# K-Means算法

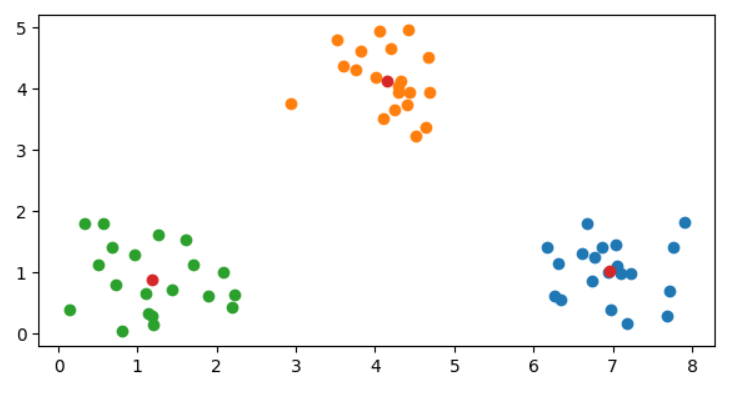
## 生成数据

数据选取二维点，便于作图表现，一共设置三种类型的高斯分布，每一聚类的均值分别是，协方差矩阵均为，每一类的数目均为20个。

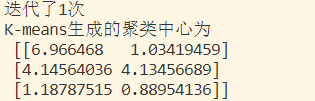
## 具体过程

首先选取初始的聚类中心，随机选择100次，每次利用欧几里得距离计算花费，最终选取最小的一组聚类中心作为最佳的初始点对。接下来通过迭代找出聚类中心，并根据最终的聚类中心划分聚类，实验结果以及控制台输出如下：

1根据高斯分布生成的数据和聚类中心

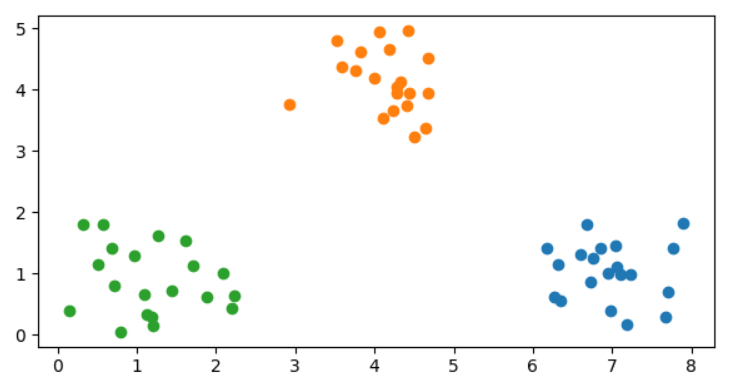


2根据K-Means算法计算的聚类中心及划分聚类



3控制台输出迭代次数以及聚类中心

## GMM-EM算法

因为GMM-EM算法不需要聚类中心，所以结果以及控制台输出如下：

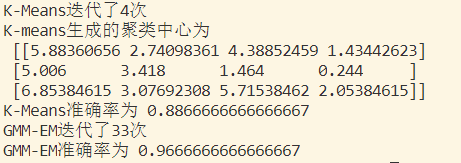
4.利用GMM-EM算法划分的聚类



5.控制台输出如图

## UCI数据集

选用UCI上较为流行的iris数据集，聚类点的特征有4个，聚类类型K共3种，每种有50个聚类点共150种，设置初始协方差矩阵为实验结果如下：



6.控制台输出如图

可以看出，对于该数据集，GMM-EM的算法效果较好。

# 分析以及结论

由于K-Means算法使用了平方差公式，实际上在进行聚类时对于每一类是使用一个圆形范围圈定每一类，具有一定的局限性，而GMM-EM算法利用贝叶斯后验分布公式，直观上使用椭圆来圈定每一类，应用性更加广泛，也可以解释在iris数据集的拟合效果较好。