

**2020年春季学期  
计算学部《机器学习》课程**

**Lab 1实验报告**

|  |  |
| --- | --- |
| 姓名 | 王丁子睿 |
| 学号 | 1183710211 |
| 班号 | 1803104 |
| 电子邮件 | 1183710211@stu.hit.edu.cn |
| 手机号码 | 19845178018 |

**目录**

[1 问题概述 3](#_Toc55219699)

[2 数据生成 3](#_Toc55219700)

[3 问题求解 3](#_Toc55219701)

[3.1 K-means 3](#_Toc55219702)

[3.2 高斯混合模型的EM解法 4](#_Toc55219703)

[4 应用 5](#_Toc55219704)

# 问题概述

给定一个点集，求一个的划分，使得对于任意，中的元素在期望的方面尽量接近/相似。

# 数据生成

利用高斯分布来生成训练和测试数据。

在本次实验中，共利用四个二维高斯分布来生成数据，每个高斯分布的训练、验证、测试数据量之比为60:40:40。

各高斯分布具体参数如下：

下文中如不特殊说明，得到的结果均出自该数据集。

# 问题求解

## K-means

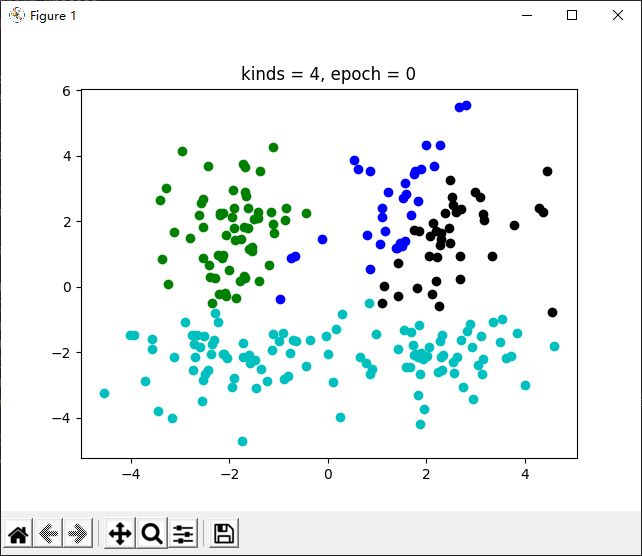
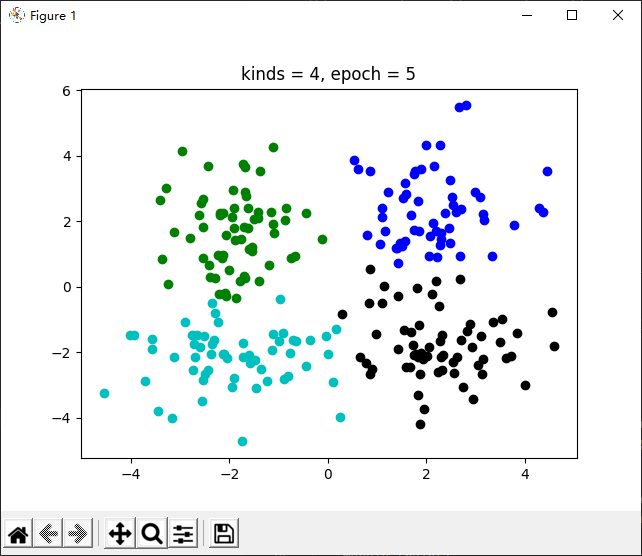
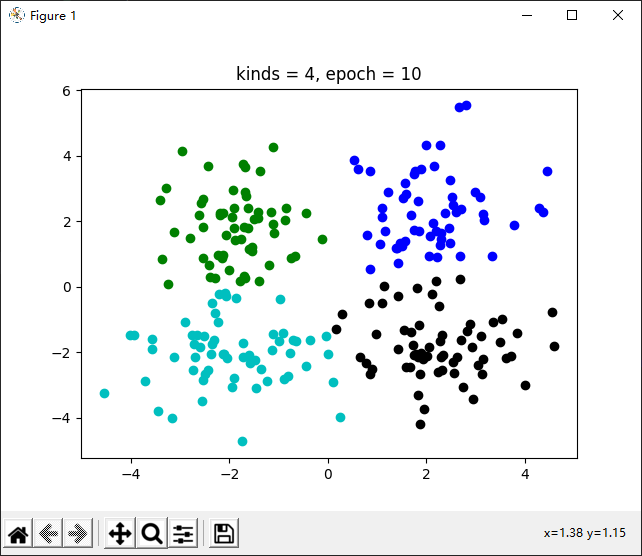
该方法基于欧氏距离，将欧氏距离最接近的一系列点划分为一个聚类。

在具体使用时，首先随机选取空间内K个点作为起始的聚类中心，每个点按照其最近的聚类中心划分为相应的聚类。

然后每次迭代重新求出每个聚类的中心（即所有该聚类点的坐标的平均），按照新的中心重新划分聚类。

不断重复上述过程，直到所有点所属的聚类不再变化，或者达到规定的迭代次数。

下取迭代次数分别为0、5、10，得到的结果如下图所示：

迭代次数大于10后，得到的结果基本稳定，故不再展示。

## 高斯混合模型的EM解法

高斯混合模型的形式为

其中是样本中类中的数据所占的比率，为第类中的高斯分布的概率分布函数，即

则每个点属于某个高斯分布的概率（又称响应度）为

由于上述模型中参数的形式过于复杂，利用梯度下降求解会非常不便，故采用EM算法。

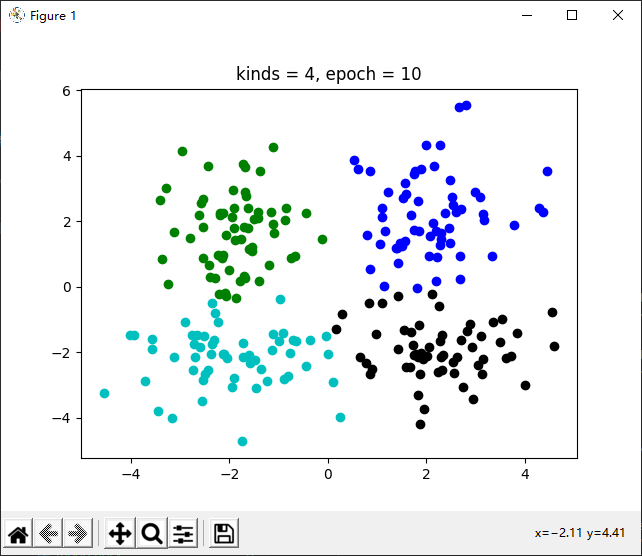
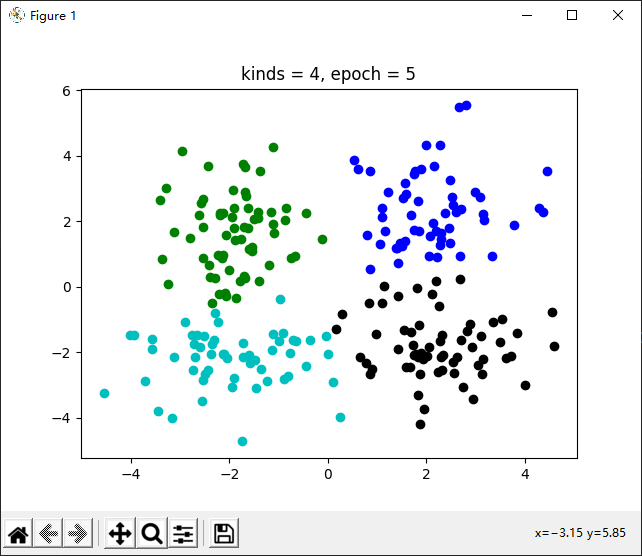
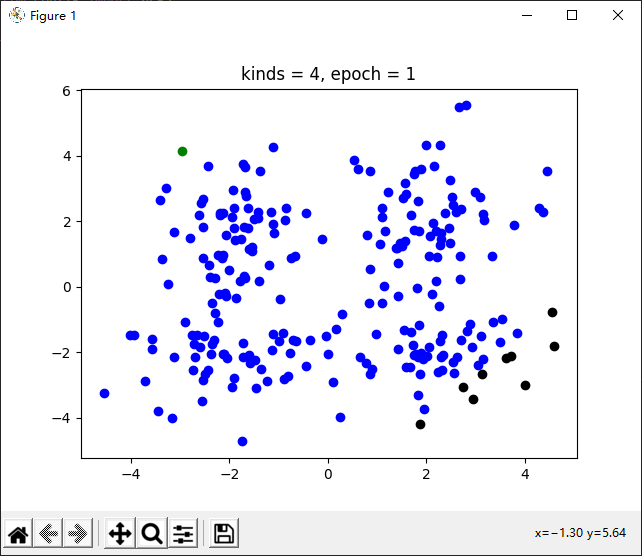
具体来说，我们首先随机选取个点作为初始高斯分布的均值点，按照欧氏距离最近求得属于各个聚类的点，再根据这些点求出初始高斯分布的方差，同时据此得到。

算法分为如下两个步骤：

E步——根据当前的和，更新响应度矩阵。

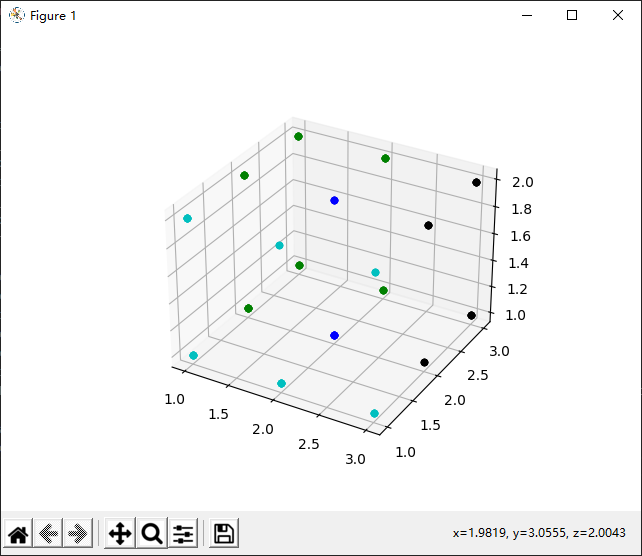
M步——根据当前的和估计的，更新和，满足

取迭代次数为1、5、10，得到的结果如下图所示：



# 应用

取UCI的数据集[MONK's Problems Data Set](http://archive.ics.uci.edu/ml/datasets/MONK%27s+Problems)进行测试，其坐标取值均为整数。得到的结果如下图所示：



可以看到整体上数据被分成了四个聚类，但覆盖的范围并不相近，可能是无法显示所有维度导致的。