# 模式识别实验报告

## 实验一 K-均值聚类

学院：机电工程

姓名： 王一径

学号：16S008203

1. **实验内容**
2. 使用C或Matlab编程实现K-均值聚类算法：要求独立完成算法编程，禁止调用已有函数库或工具箱中的函数；
3. 使用仿真数据测试算法的正确性：将下列19个样本聚成2个聚类：



1. MNIST数据集测试：ClusterSamples中的10000个784维特征手写数字样本聚类为10个类别，根据SampleLabels中的标签统计每个聚类中不同样本的数量。测试不同初始值对聚类结果的影响。
2. **程序代码（python）**

（K-均值算法部分代码）

class kmeans(object):

def \_\_init\_\_(self,data,k):

self.data = np.array(data).astype(float)

self.u = self.data[np.random.randint(0,self.data.shape[0],k),:]

self.label = np.random.randint(0,k,self.data.shape[0])

def train(self):

e = 1

while e > 1e-16:

#E step

i=0

while i < self.data.shape[0]:

temp0 = (self.data[i,:]-self.u)\*\*2

temp1 = np.sqrt(np.sum(temp0, axis=1))

self.label[i] = np.argmin(temp1)

i += 1

#M step

preU = copy.deepcopy(self.u)

j=0

while j < self.u.shape[0]:

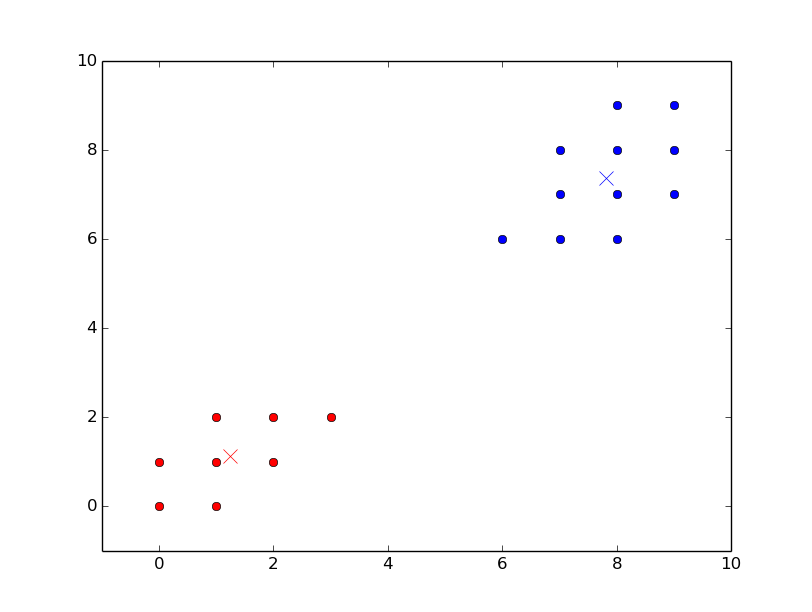
n = self.label[self.label == j].size

self.u[j] = np.sum(self.data[self.label==j,:], axis=0)/n

j += 1

e = np.sum((preU-self.u)\*\*2)

1. **实验结果**
2. 仿真数据实验结果：（可以列出每个聚类中包含的样本，也可以画图显示不同聚类）



1. MNIST数据集实验结果：

**每个聚类中包含不同类别样本数量统计表**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 聚类0 | 26 | 421 | 1 | 27 | 2 | 454 | 2 | 23 | 2 | 2 |
| 聚类1 | 3 | 0 | 1 | 0 | 1101 | 0 | 6 | 3 | 2 | 10 |
| 聚类2 | 20 | 17 | 26 | 31 | 120 | 1 | 9 | 39 | 5 | 726 |
| 聚类3 | 158 | 16 | 16 | 14 | 70 | 4 | 27 | 645 | 7 | 33 |
| 聚类4 | 3 | 1 | 353 | 19 | 29 | 0 | 251 | 0 | 300 | 3 |
| 聚类5 | 285 | 34 | 32 | 23 | 143 | 11 | 35 | 308 | 66 | 4 |
| 聚类6 | 20 | 27 | 25 | 791 | 73 | 13 | 0 | 7 | 0 | 23 |
| 聚类7 | 2 | 1 | 99 | 2 | 69 | 1 | 422 | 0 | 457 | 5 |
| 聚类8 | 565 | 4 | 27 | 15 | 113 | 5 | 22 | 184 | 32 | 12 |
| 聚类9 | 11 | 0 | 286 | 1 | 22 | 5 | 368 | 14 | 306 | 1 |