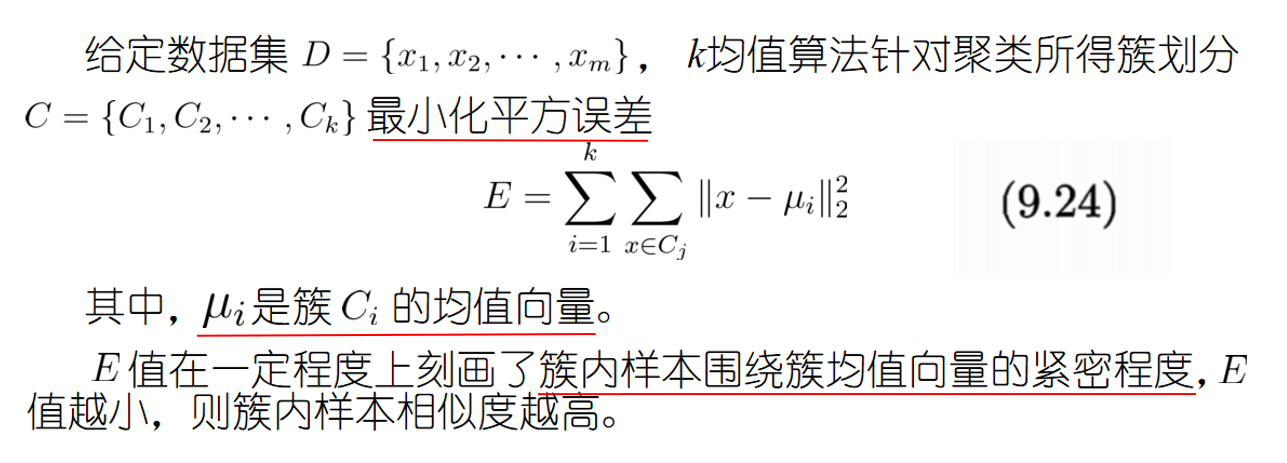
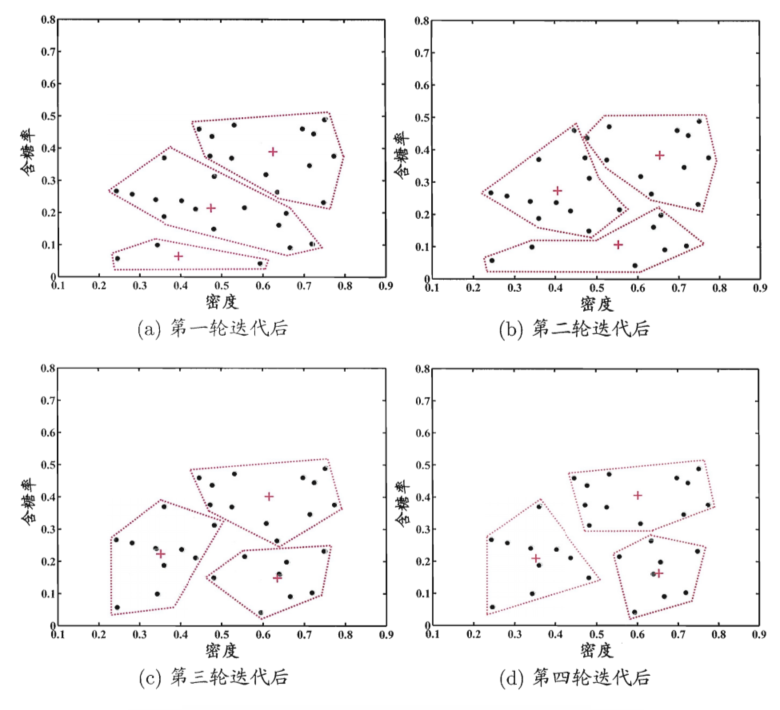
本次实验是针对课本第9章聚类

聚类，即根据相似性原则，将具有较高相似度的数据对象划分至同一类簇，将具有较高相异度的数据对象划分至不同类簇。聚类与分类最大的区别在于，聚类过程为无监督过程，即待处理数据对象没有任何先验知识，而分类过程为有监督过程，即存在有先验知识的训练数据集。

本次实验采用的是K-Means算法，又称K均值算法，是应用最广泛的聚类算法之一。

K-means算法的原理是：对于给定的样本集，按照样本之间的距离大小，将样本集划分为K个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。





这幅图片是课本上的示意图，在迭代5轮之后与第四轮迭代相同，于是算法停止，得到了最终的簇划分

讲一下算法流程

实验9是主成分分析PCA

PCA（主成分分析）是一种常用的降维技术，其大体思想是通过线性变换将原始数据投影到一个新的坐标系中，使得数据在新坐标系中的方差最大化。具体来说，PCA通过找到数据中的主成分（主要方向）来实现降维，从而保留数据中最重要的信息。

在图示中，原始数据是三维的，但是我们分析发现，通过某种线性变换可以找到两个主成分 PC1 和 PC2。这两个主成分构成了一个新的坐标系，使得数据可以在这个新坐标系中被更好地表示。PC1 和 PC2 是原始数据各个维度特征的线性组合，其选择是为了最大程度地保留原始数据的方差。将数据投射到 PC1 和 PC2 构成的平面上，实现了对原始三维数据的降维，同时尽可能地保留了原始数据的信息。

总的来说，PCA 的目标是通过寻找数据中的主成分，将数据投影到一个低维的子空间中，从而实现数据的降维，同时保留尽可能多的信息。

接下来简要看一下PCA的算法流程

算法的输入：首先是样本集合D，和所要降维到的低维空间维度数d撇

PCA算法的过程通常包括以下几个步骤：

1. 样本数据中心化

首先，对所有样本进行中心化处理，即从每个样本中减去数据的均值。这样做的目的是消除不同特征之间的单位和量级差异，使数据的均值为0。

2. 计算协方差矩阵

接着，计算样本的协方差矩阵。协方差矩阵用于度量不同特征之间的相关性



3. 特征值分解

对协方差矩阵进行特征值分解。这一步骤的目的是找到数据中方差最大的方向，即数据中最多变的方向。特征值分解会给出特征值和对应的特征向量，其中特征值表示了数据在特征向量上的投影值的方差。

4. 选择主成分

从特征值分解中取出最大的 d撇个特征值对应的特征向量。这些特征向量代表了数据中的主成分，它们是降维后数据的基。

最后是：构造投影矩阵

将选出的特征向量组合成一个投影矩阵。这个矩阵将用于将原始数据映射到低维空间。

数据降维

最后，使用投影矩阵将原始数据集投影到选定的主成分上，从而得到降维后的数据。