**1.实验目的**

完成 Kmeans 聚类算法以及二分 kmeans 的实现，并利用该算法对一些数据集进行预测，以此掌握基本的 Kmeans 算法以及二分 kmeans 算法，并期望能够利用该模型对数据集做出合理的预测。

首先完成了 Kmeans 基本聚类算法的实现，在此基础上针对 Kmeans 初始化 k 个中心点时可能出现由于位置太极端导致某些点在迭代过程中消失的问题，实现了二分kmeans 算法。同时，意识到二分 Kmeans 因为采取每次二分的方式，因此迭代次数会比较大。因此，针对这个问题，再次改进了算法，实现了 Kmeans++算法，同样是初始化 K个数据点，但 Kmeans++的基本原则是使得初始化中心点间的距离尽可能远，从而避免出现 Kmeans 问题。 最后，针对以上 Kmeans 算法以及改进算法中心点的个数 K 是手动选择可能不是最优解，给出了两种可视化方法：聚类散点图，以及手肘法效果图，以判断不同的 K 的聚类效果的好坏。

**2.普通K-means**

Kmeans 聚类算法：聚类算法属于无监督学习，将一个庞大的数据集中具有相似特征的 数据汇聚在一类，称为一个簇。簇内的数据点距离越近越好，簇外的数据点相聚越远越好。 Kmeans 就是 K 均值，即构造 K 个不同的簇，且每一个簇的中心点均由该簇内点的平均值构成。

Kmeans 分类核心是欧氏距离，公式为：𝑑 = √∑𝑛 𝑖=1 (𝑥𝑖 − 𝑦𝑖) 2。Kmeans 通过计算数据点之间的欧氏距离来判断数据点的远近，并以此判断该数据点属于哪一个类簇。Kmeans 首先随机地初始化 k 个数据点，然后分别计算每个样本点到中心点的距离，并将该样本点归到最近的中心点的簇中。其次，在所有簇划分完毕后，计算每一个簇中样本点的平均值，以此形成新的中心点并替代原来该簇中的中心点。重复以上过程，直到每一个样本点的归类均不再改变为止。