3 算法设计

3.1样本生成

根据题目要求，本次实验没有直接提供样本集，要求我们利用图形特征，在双月中两个区域随机取得1000个数据点作为无标志训练样本作为样本集。此外，图形的具体参数并未直接给出，我们尝试后，选择了较为合适的半径与宽度，分别为10，6。为了便于之后的数据处理，我们选择用numpy库进行随机样本的生成，即先生成事先给定的圆心与半径，再通过数学计算，最后得到了区域中的随机数据点。至此，就可以顺利生成题目中要求的样本集。

3.2 k均值聚类算法

本次实验的主要内容是利用k均值聚类算法对“双月数据集”进行聚类处理和分析。因此，在该部分中，我们先对k均值聚类算法进行讲解。

第一步，选择聚类数，初始化聚类中心位置，一般可以取随机样本点作为初始聚类中心。当然，在此之前可以进行数据的预处理，例如标准化、异常点过滤等。

第二步，对每一个样本点，将其分配到距离最近的聚类中心，这里的距离可以自行定义。

第三步，对每一个聚类，利用类中的样本点重新计算该类的中心。

重复第二步与第三步，当每一个样本点的类与上一次相比不发生变化时聚类结束。也可以选择将迭代次数或者损失函数是否已经收敛作为结束标志。

本次实验中所利用的最基础的k均值聚类算法同样基于这些步骤。以样本点所属类不发生变化为结束标志。相关函数的返回值包括聚类信息，即每个样本点对应所属的类和与该类的距离。

3.3 改进算法

考虑各种改进，分析其在本次实验中是否适合使用。

二分k均值：采用中位数代替平均数的聚类方法。适合簇中存在异常点的情况。本次实验样本点均来自代码随机生成，区域在给出区域参数后也是完全确定的，没有必要用中位数代替平均数来聚类。因此我们没有考虑该聚类方法。

重复选取质心：

接下来根据题目要求，进行仿真实验并对各个问题进行结果展示以及深入分析。

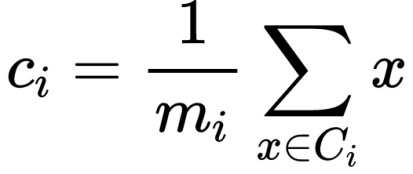
4.1 聚类结果展示

根据第三部分中提到的代码逻辑，我们可以得到最基本的聚类结果，在这一部分中，我们将对8个不同的垂直间隔d，直接以图形的方式展示聚类结果，如下图所示：

可以看到，在这种参数选择的情况下，对于不同的垂直间隔，所得到的聚类结果大致是相同的，即两个区域分别聚出三类，并且分布也较为平均，有一定的规律性。

4.2 定量结果分析

以上是聚类结果直观的展示，但想要深入的分析离不开定量的信息。接下来将根据题目要求对相应的结果进行展示并分析。首先将不同垂直间隔下各聚类中心点和损失函数进行展示：



其中表示第i个簇，是中的点，是第i个簇的均值，是第i个簇的样本总数。

