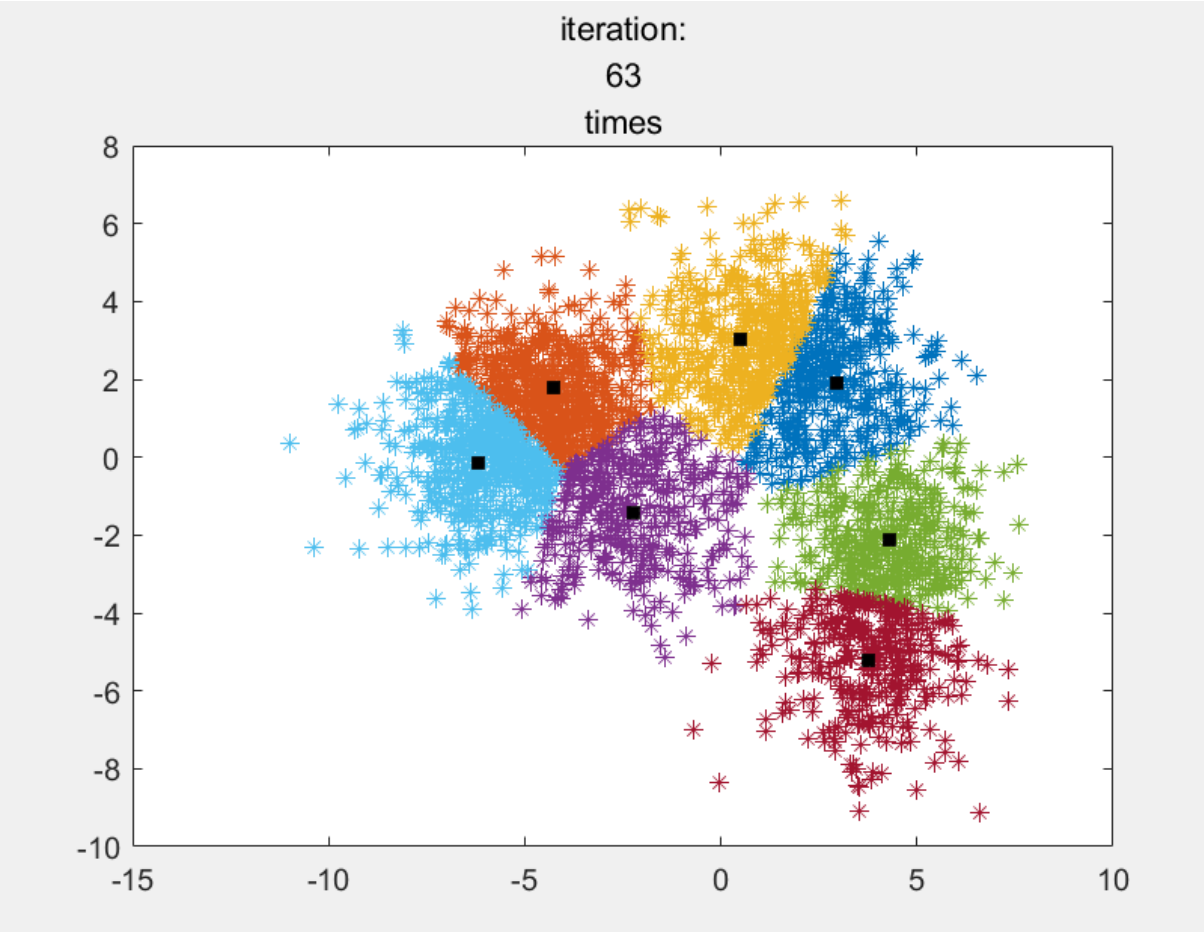
1. k-means 聚类一定会收敛吗？为什么？

：k-means聚类法是先选取k个Centroid。通过欧几里德距离算法，使所有数据点属于最近的cluster。然后算各个聚类的均值找新的Centroid。反复这些过程中，新的centroid值差别特别小，而取最小值。不能保证全局收敛，但是能达到局部最优值。

2. 完成函数function label = kmeans clustering(data; num)，其中输入变量data为N行m 列，每一行为一个数据点， num 表示聚类数目；输出变量label 为N 行1 列，表示对应的数据点属于哪一类（比如属于第一类的点label 就为1）



:当k=7时的k-means聚类分析结果如上图。

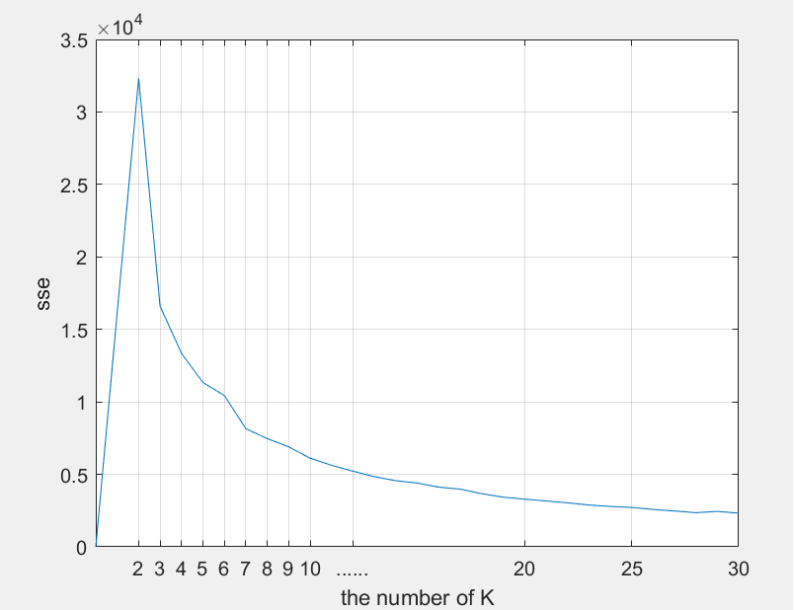
我设计的函数格式是

```
function [label,sse,std_Init,iteration] =
                                kmeans_clustering(data, k,setting)
```

其中，sse：误差平方和 ， std_Init：初始点的标准偏差， iteration：迭代次数，setting：分别函数功能

3. 聚类数目从2类开始逐渐增加，分别进行计算并分析聚类效果，决定最合适的

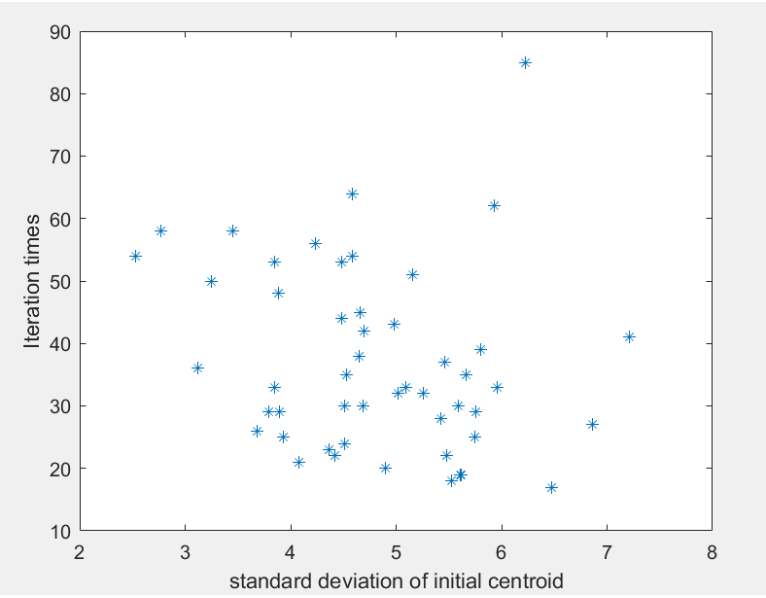
聚类数目并说明理由



$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist(x, C_i)^2$表示每个样本到聚类中心的距离之和。

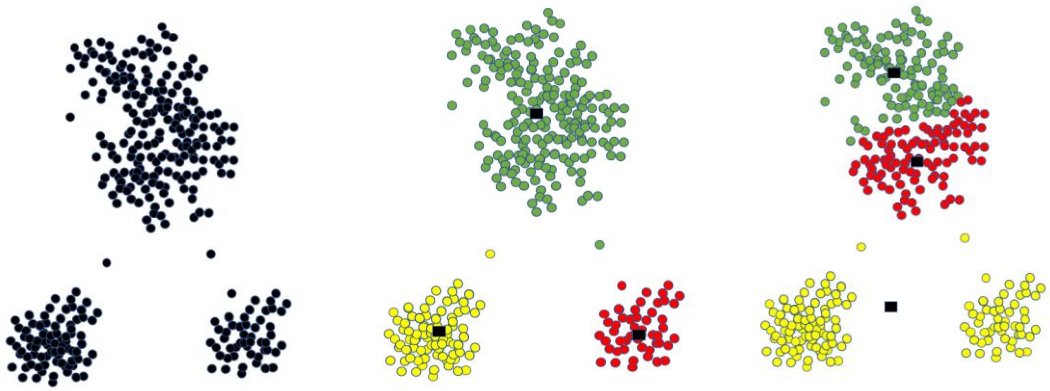
利用Elbow法，随着k数目的增加，找出SSE值收敛前递减程度比减小的某一区间，适当的k值是其区间的一个值。如上图，k=7开始SSE值递减的程度变小，所以我选取k=7为分析中适当的k数目。

4. 选择不同的初始点多次实验，观察初始点的选择对最终结果的影响，并分析为什么会有这种影响。



：如上图是固定k数目情况下观察随着初始点的标准偏差的迭代次数变化。结论不是一

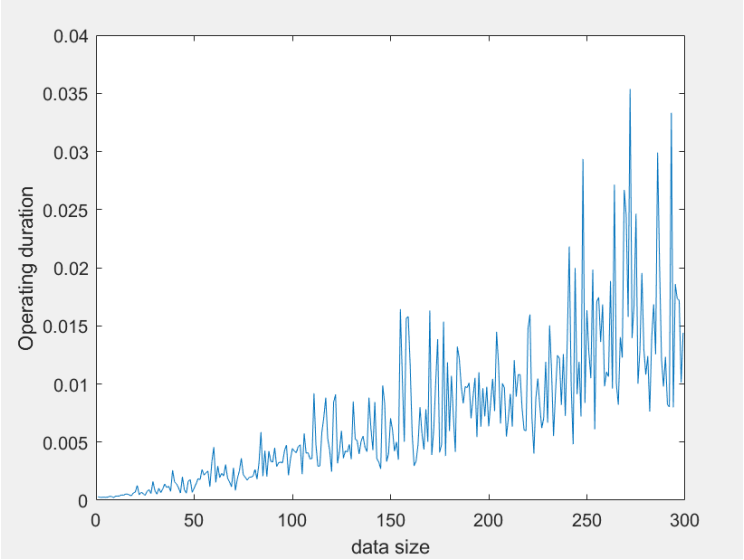
定的，但是可以看不同的初始点影响聚类算法收敛的速度，迭代次数差异比较大的趋势。



:还有一个可能性是聚类的不是理想。比如说，如第一个上图的数据，按初始点的不同，大部分的结果是第二个图一样生成聚类。但是很少有可能会发生第三图一样的聚类发生。

两个结论看出初始点对最终结果的影响不是很大。但是速度方面和聚类结果有时有影响。

5. 选择不同的数据规模进行实验，计算你的程序耗时，观察耗时与数据规模之间的关系，从中你能得到什么结论？



：如上图是，观察耗时与数据规模之间的关系。通过Matlab实验，容易看出，随着数据的规模的增加，耗时逐渐增加。

附录》

hw6.m ： 按照题目分结果。

kmeans_clustering.m ： k均值聚类函数