模式识别第六次作业

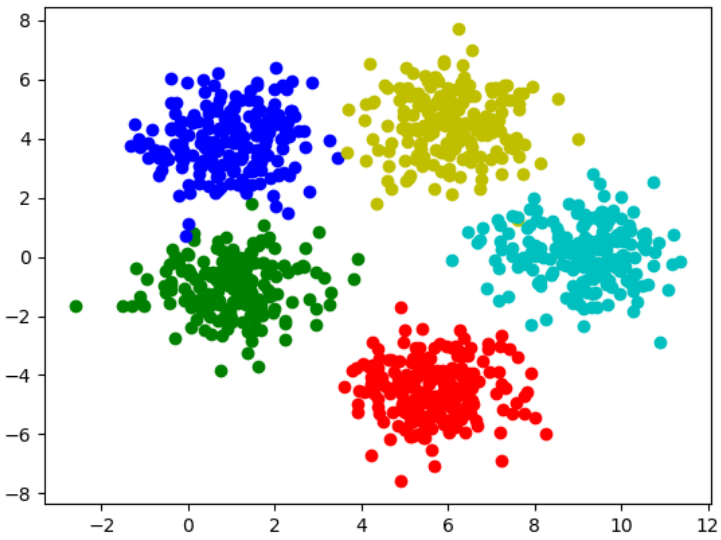
K-means聚类

实验内容

随机生成1000个服从5个不同高斯分布的数据点，并使用K-means聚类算法对这1000个二维空间数据点进行聚类。

实验结果

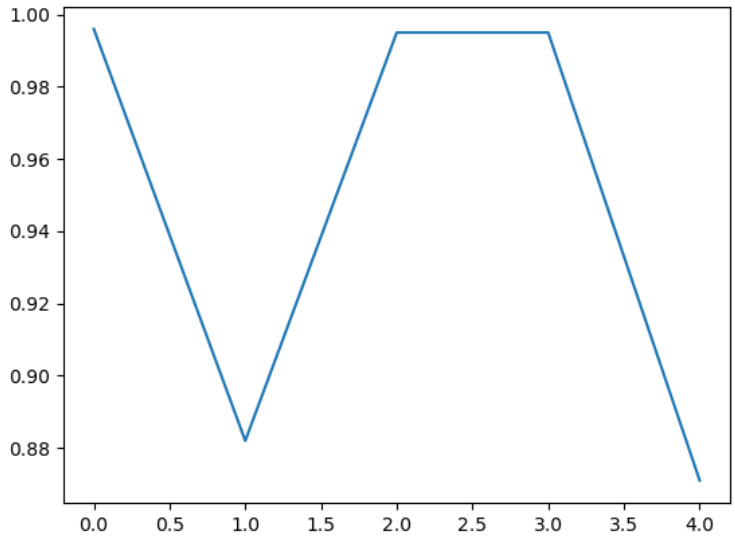
聚类结果示意图如下所示：



聚类中心，精度和误差如下所示：

```
Centers:
[[ 5.51312614 -4.47680837]
 [ 0.94863059 -0.83729642]
 [ 6.11377863  4.40594737]
 [ 9.1077563  -0.02690987]
 [ 0.95394992  4.11491752]]
Error: 0.07927509997437023
Accuracy: 0.992
```

随机选择初始值得到的聚类精度曲线图如下所示：

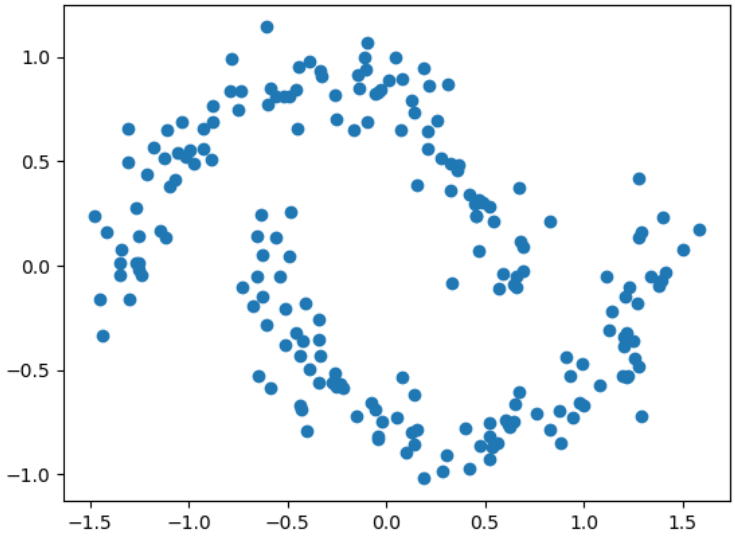


可以看出，在随机5次实验下，聚类精度波动较大，这说明初始值是影响K-means聚类算法性能的重要因素。

谱聚类

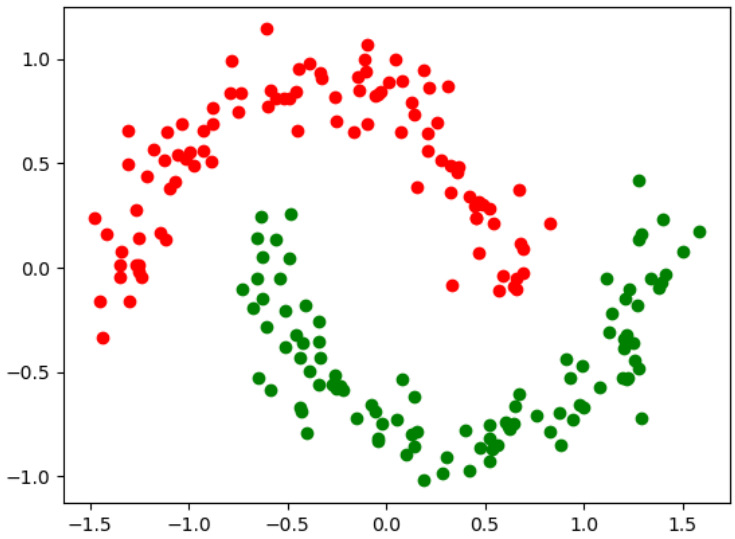
实验内容

给定200个由两个半月形分布生成的数据点（如下图所示），使用谱聚类算法实现 Normalized Spectral Clustering (Ng 算法）。

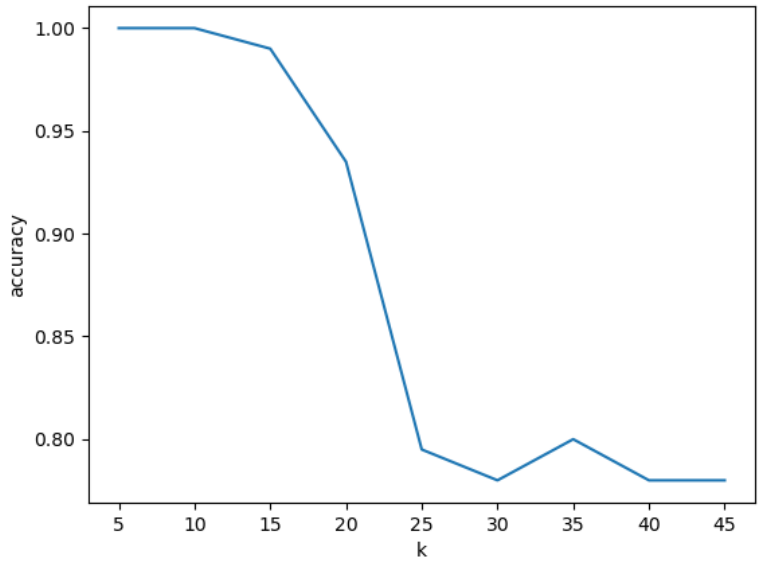


实验结果

谱聚类结果示意图如下所示：

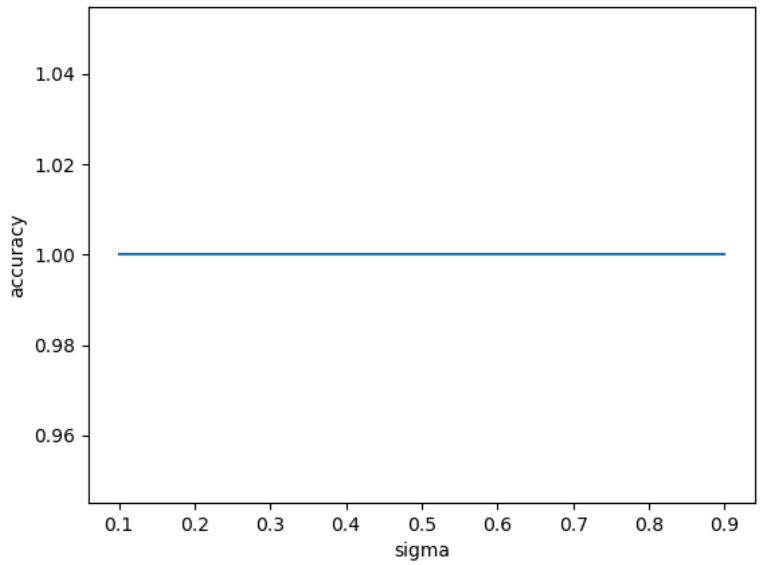


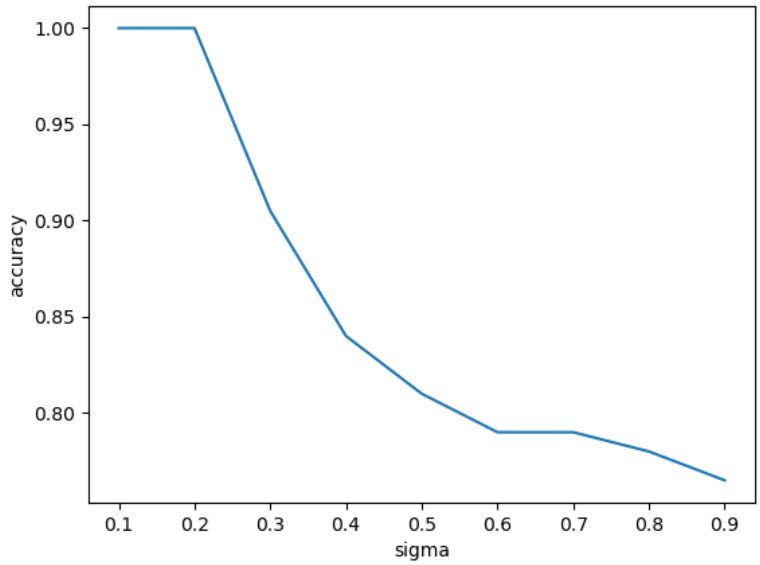
固定σ=1，聚类精度随着k值的变化曲线如下所示：



可以看出，k的取值较小时谱聚类的性能较好，这是因为此时对数据点计算得到的k近邻邻居都是位于同一个半月牙分布，从而后续得到的相似度矩阵较为准确的刻画了数据点之间的关系。如果k的取值较大，数据点的k近邻邻居则可能位于不同的半月牙分布，从而降低了算法的聚类精度。

分别固定k=5，k=50，聚类精度随着σ值的变化曲线如下所示：





可以看出k较小时，σ的取值几乎不影响聚类性能，但在k较大时，太大的σ取值会强化两个半月牙形状之间的连边，导致聚类精度的下降。