系统工程导论作业六

聚类分析

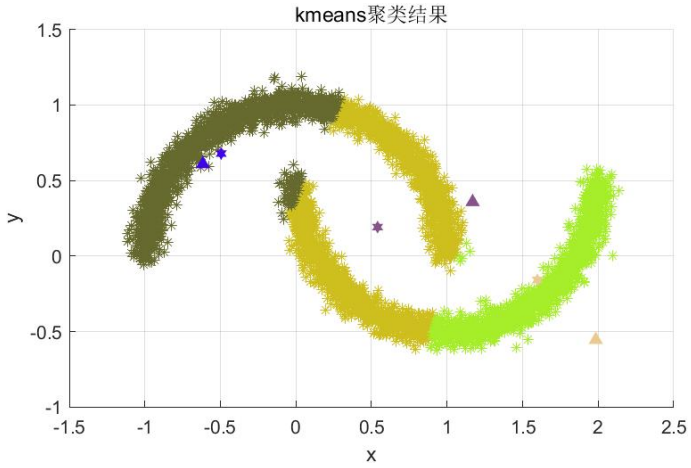
刘若涵 自 05 2020011126

1 kmeans 聚类

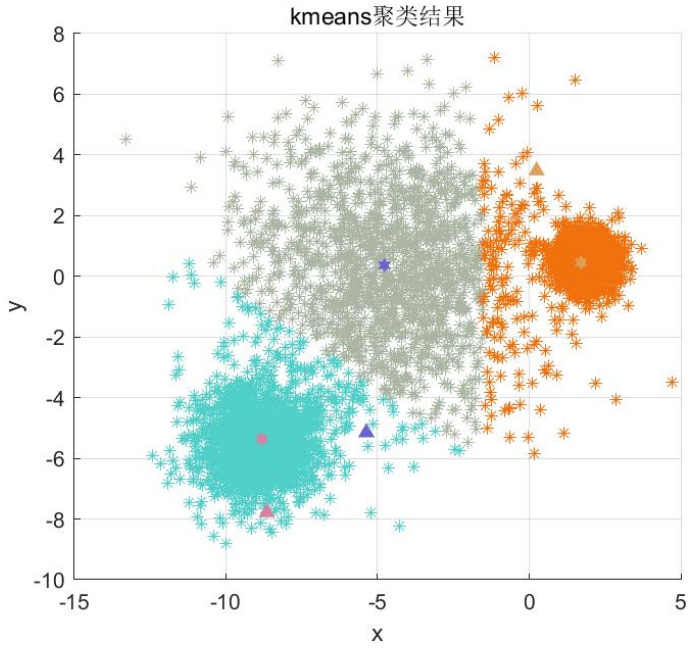
以下所有聚类结果图片中，不同聚类的点染为不同的随机颜色。“▲”为随机产生的初始聚类中心，“✡”为最终聚类中心，同颜色的“▲”和“✡”为同一聚类的初始和最终中心。

1.1 聚类结果

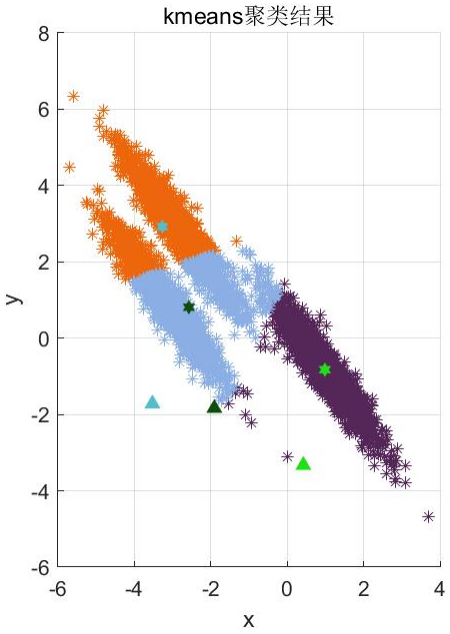
k = 3 时，data1 的 kmeas 聚类结果：



k = 3 时，data2 的 kmeas 聚类结果：

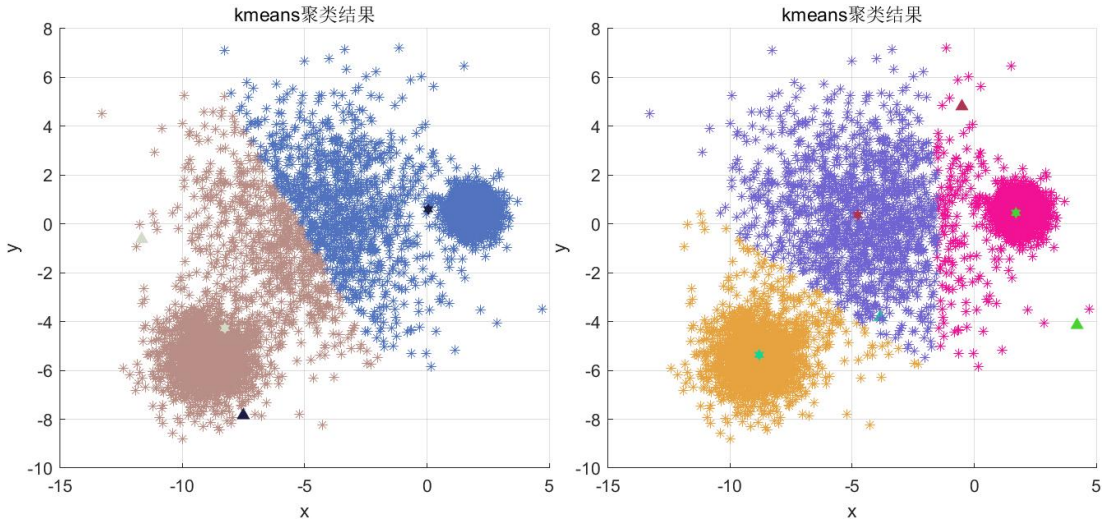


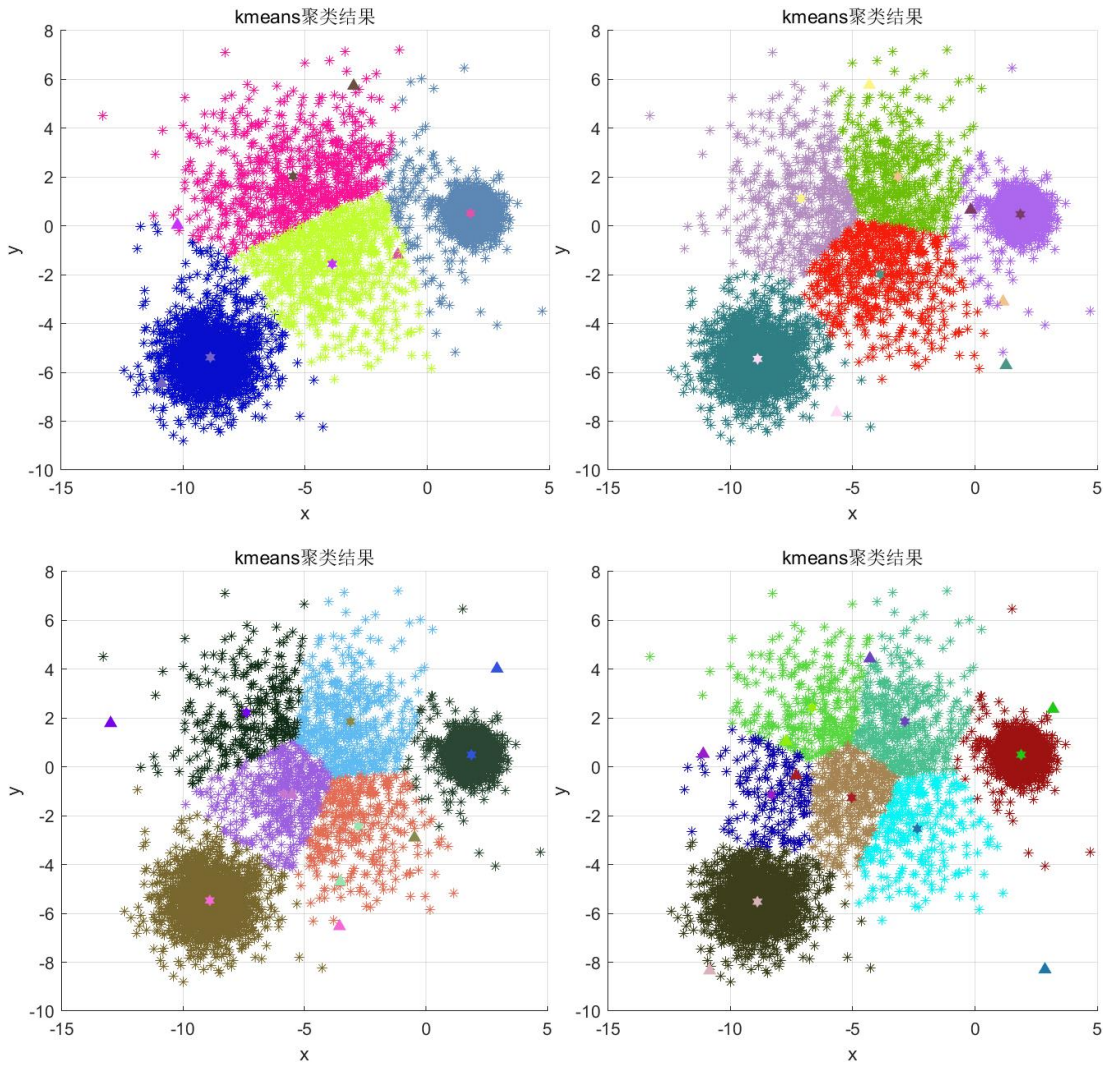
k = 3 时，data3 的 kmeas 聚类结果：



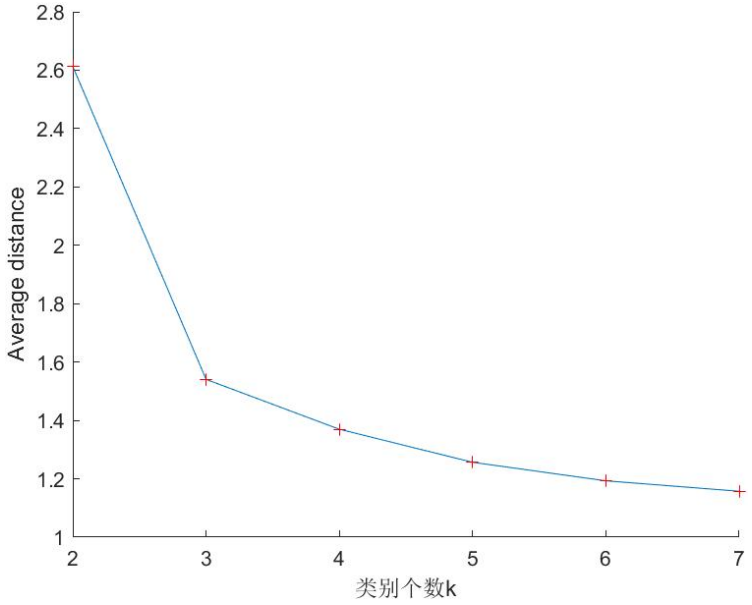
1.2 聚类数目 k 对结果的影响

k = 2 ~ 7 时，data2 的 kmeas 聚类结果如下所示。





所有数据点到其对应聚类中心的距离的平均值（Average distance）关于聚类数目 k 的曲线如下图所示。

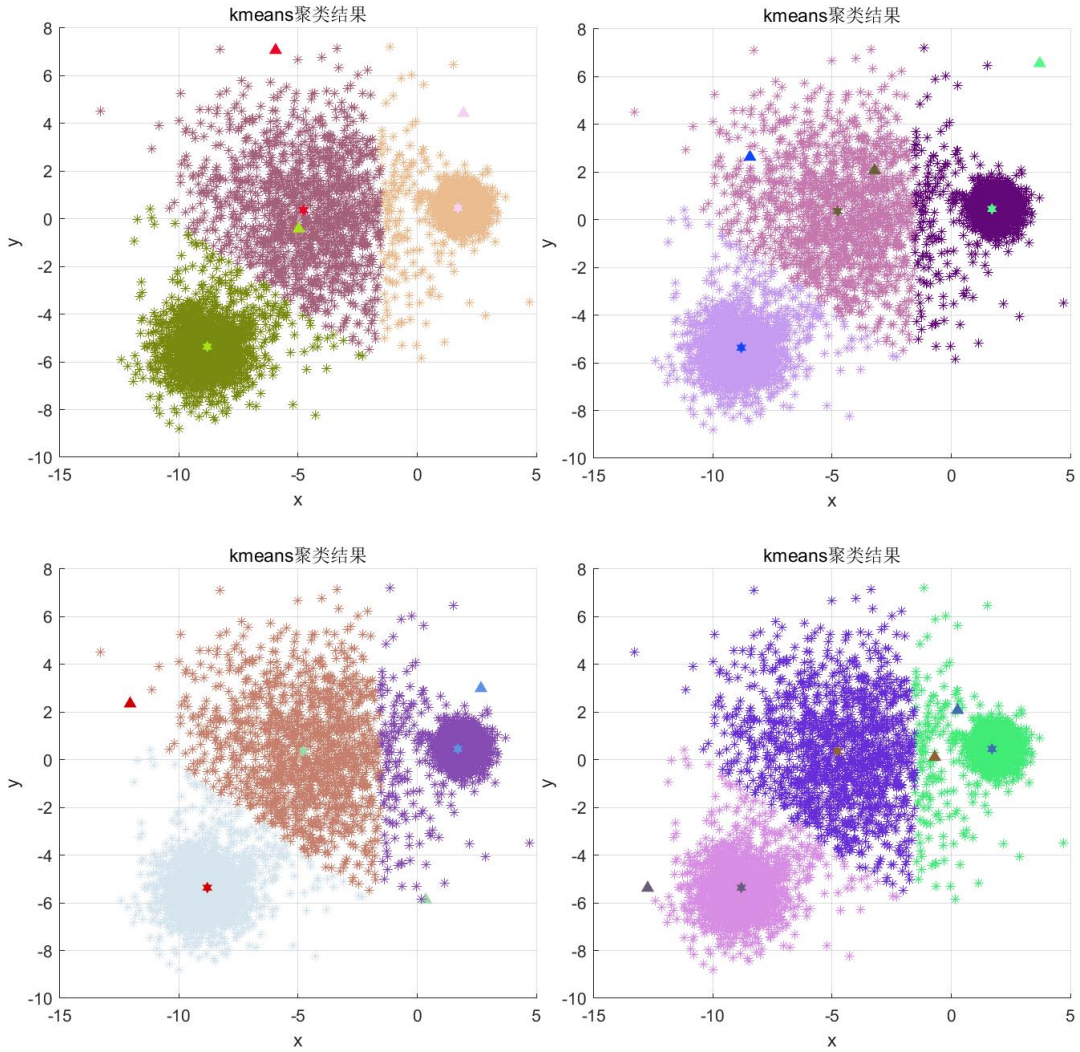


随着聚类数目的增大，对样本划分的更精细，Average distance 下降。由图

可知，当 k < 3 时，聚类数目增大，Average distance 下降明显。而当 k > 3 时，聚类数目增大，Average distance 下降缓慢，因此根据肘方法可知，最佳聚类数目为 3。

1.3 初始点的选择对结果的影响

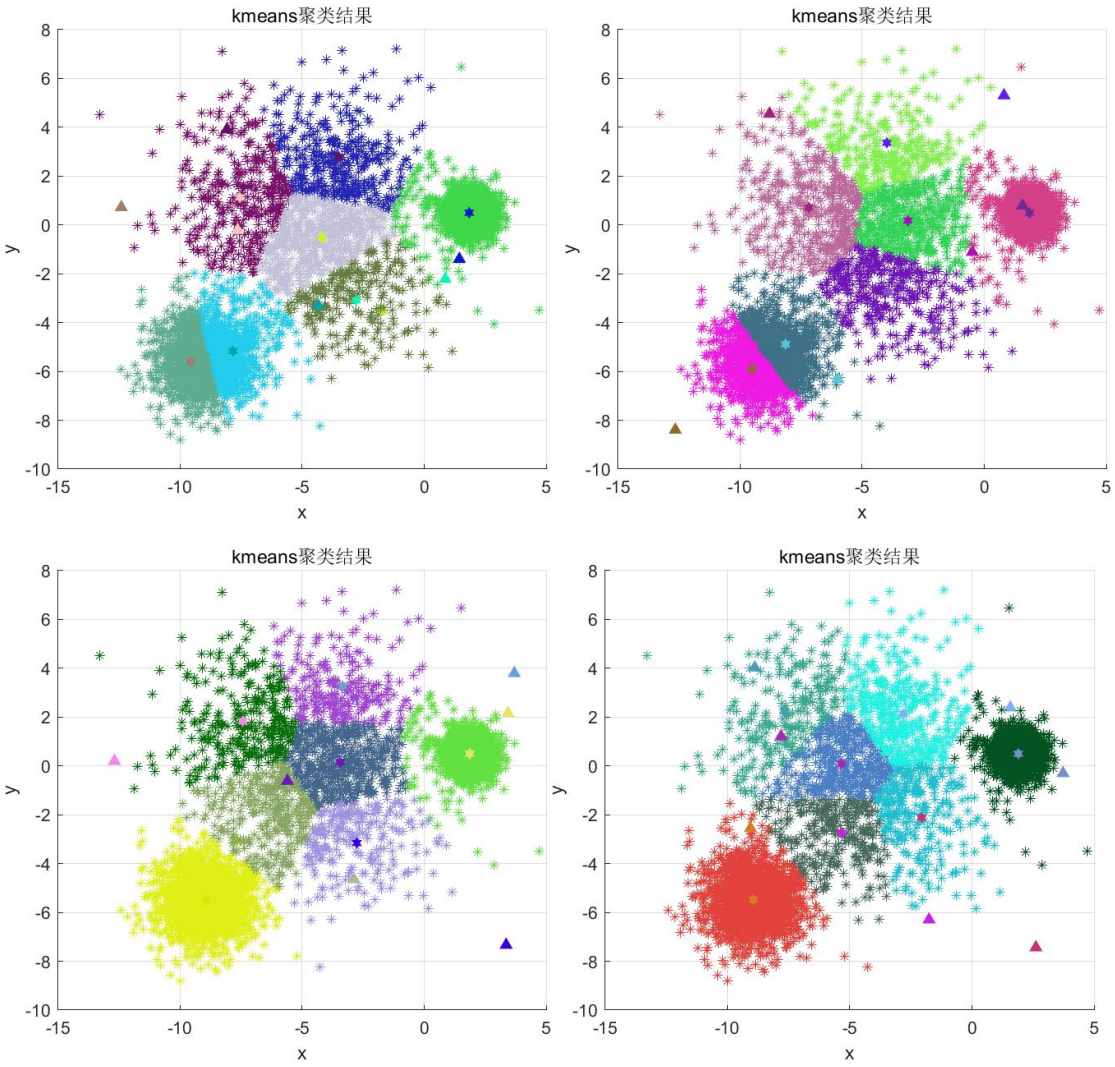
取 k = 3，选择不同的初始点多次实验，得到 data2 的 kmeas 聚类结果如下所示。



由图可知，k = 3 时，初始点的位置对最终聚类中心的位置影响不大，这是由于“最优分类”方式明确，近乎唯一，因此无论初始中心点的位置如何选择，均能迭代为近似一样的结果。且我设定的停止迭代时聚类中心点更新后与更新前距离差的阈值为 0.001，很小，因此初始点的选取对最终聚类中心位置的影响不大。但初始点的不同选择可能会影响算法收敛的速度。这是由于初始聚类中心到最终聚类中心距离不等，迭代次数不一。此外，初始点的选择也可能会影响部分

边界点的归属，这是由于聚类中心收敛的方向不一，停止迭代时偏移的方向不一，导致聚类边界点的归属可能不一致。但由于我设定的停止迭代阈值较小，聚类中心的位置偏移很小，因此对于边界点的归属影响不大。

取 k = 7，选择不同的初始点多次实验，得到 data2 的 kmeas 聚类结果如下所示。



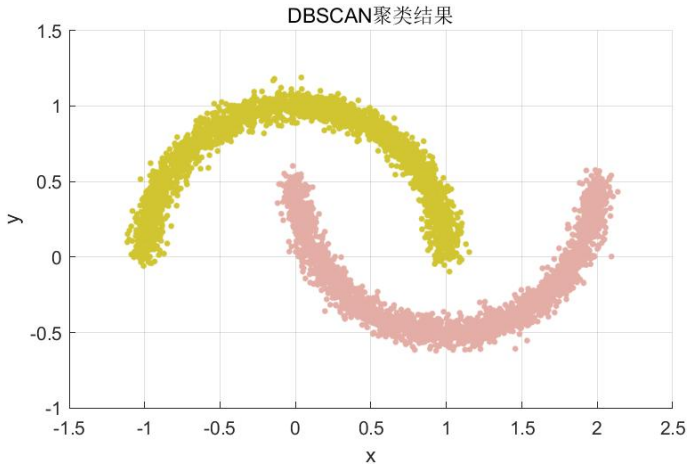
由图可知，k=7 时，初始点的选取会影响到最终聚类中心的位置，进而影响到聚类结果。这是由于划分类别较多，有多种分类方式能满足阈值条件，在初始中心点不同的情况下“最优”分类不同。

2 DBSCAN 聚类

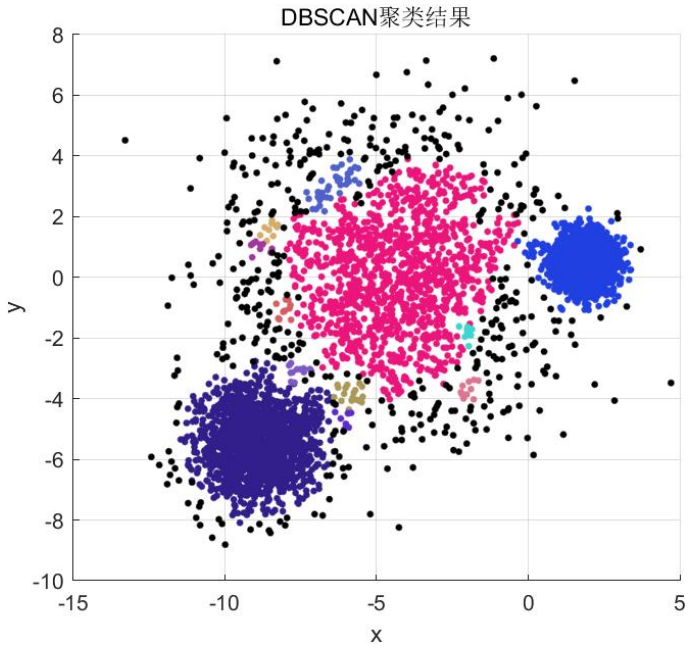
以下所有聚类结果图片中，不同聚类的点染为不同的随机颜色。黑色的点为噪声点。

2.1 聚类结果

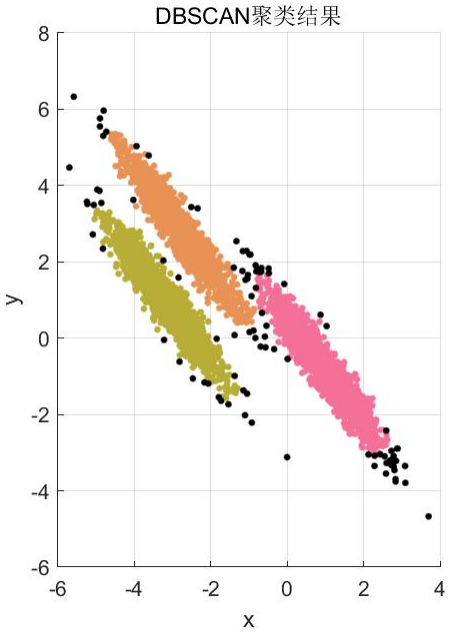
ε = 0.2，minPots = 10 时，data1 的 DBSCAN 聚类结果：



ε = 0.4，minPots = 10 时，data2 的 DBSCAN 聚类结果：

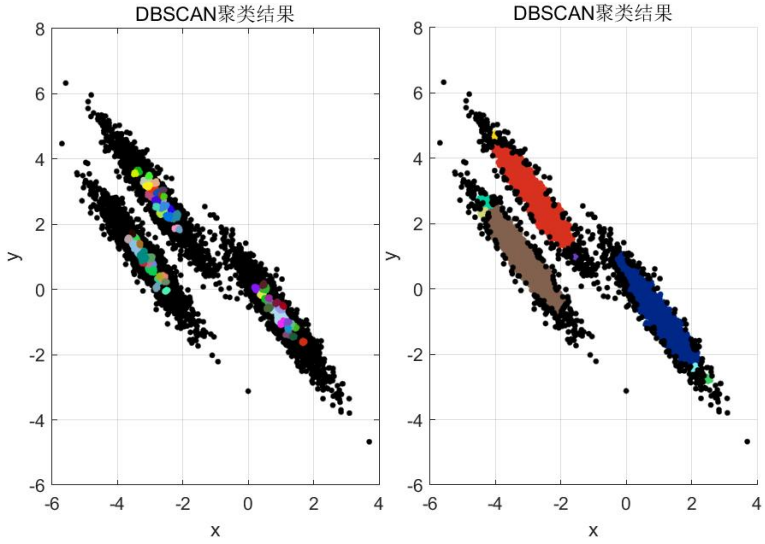


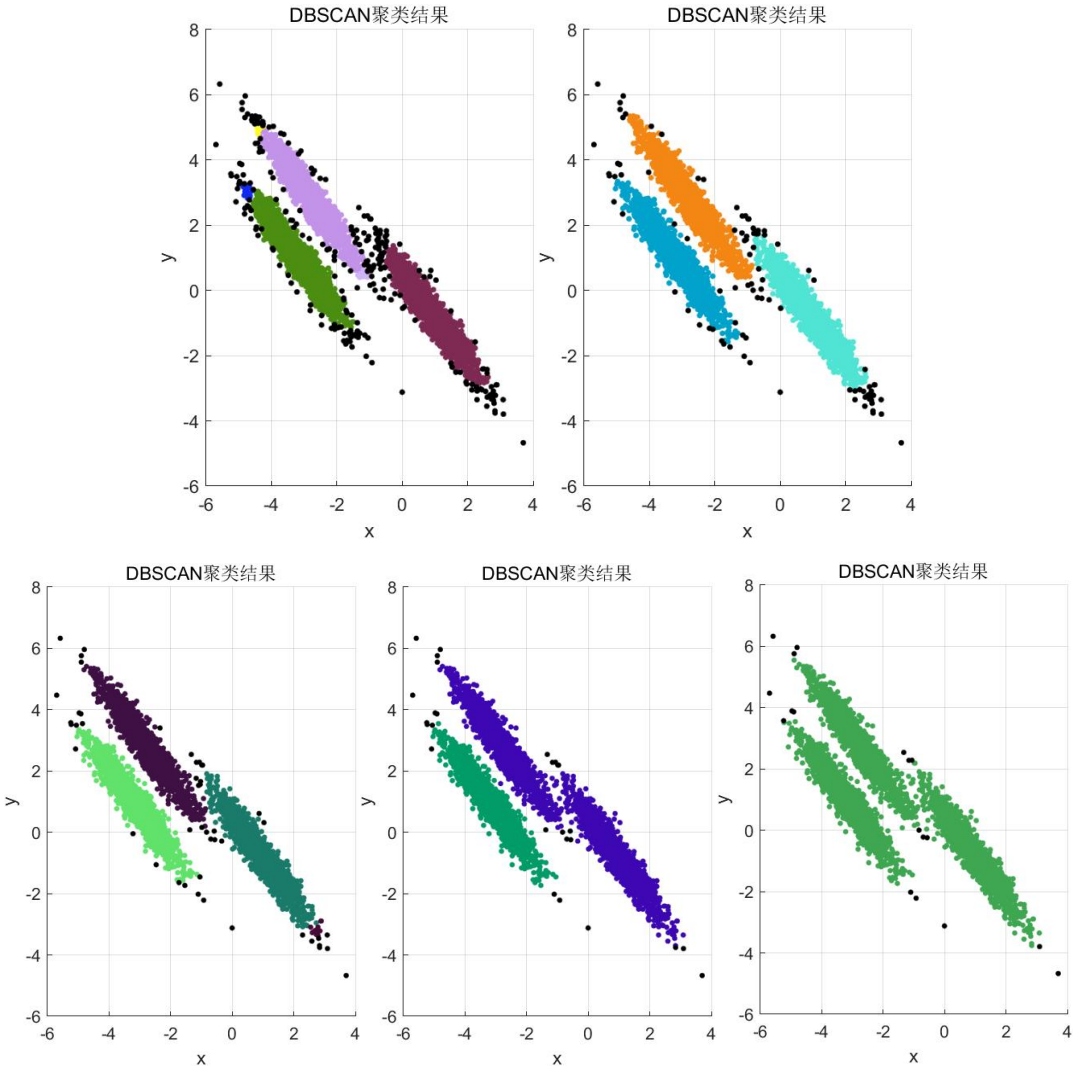
ε = 0.2，minPots = 10 时，data3 的 DBSCAN 聚类结果：



2.2 ε 对结果的影响

minPots = 10，ε = 0.05、0.10、0.15、0.20、0.25、0.30、0.35 时，data3 的 DBSCAN 聚类结果分别如下图所示。

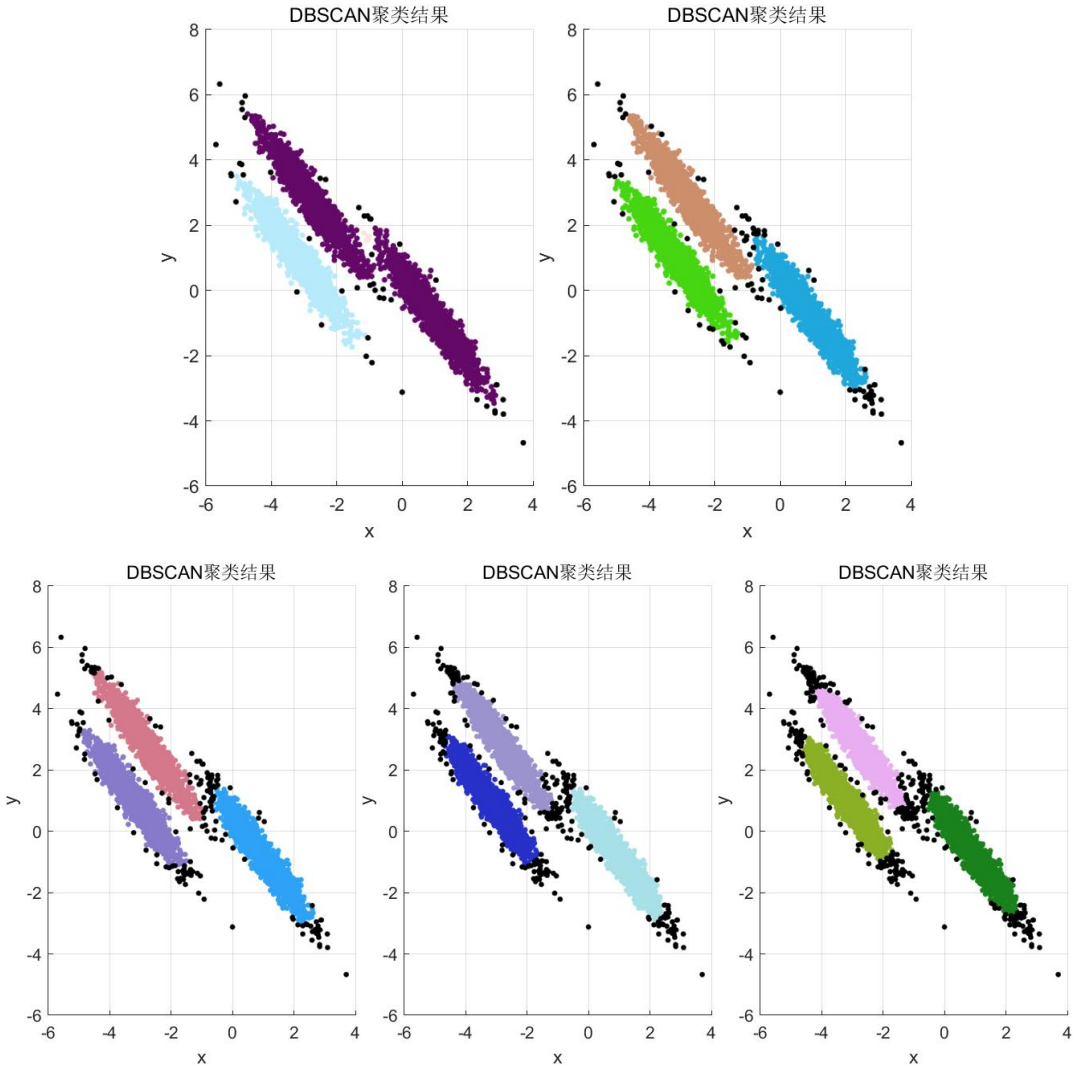




由图可知，当 ε 很小时，要求的密度过大，难以达到可达条件，因而会出现许多小聚类，同时噪声点很多。当 ε 逐渐增大时，邻域的范围增大，核心点的邻域内点的个数更容易达到 minPots，因此聚类变大，噪声点数变小。但当 ε 过大时，一些聚类交界处不能很好地分辨，可能会出现将一些聚类不合理地合并在一起。

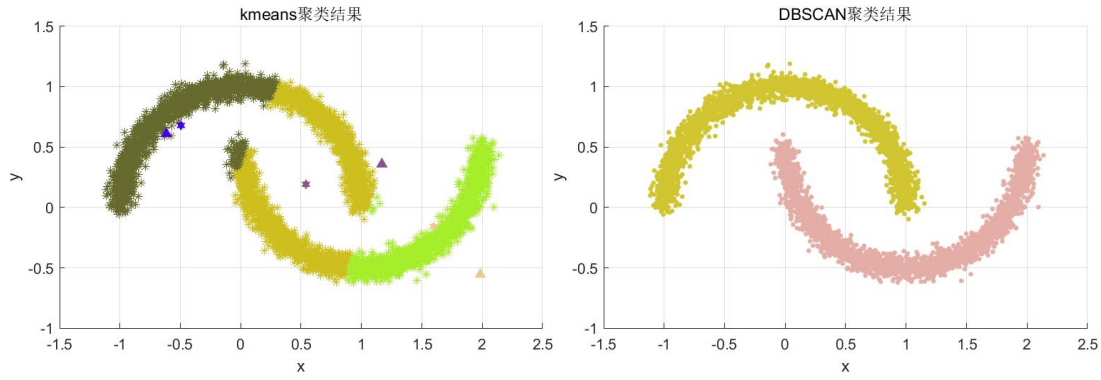
2.2 minPots 对结果的影响

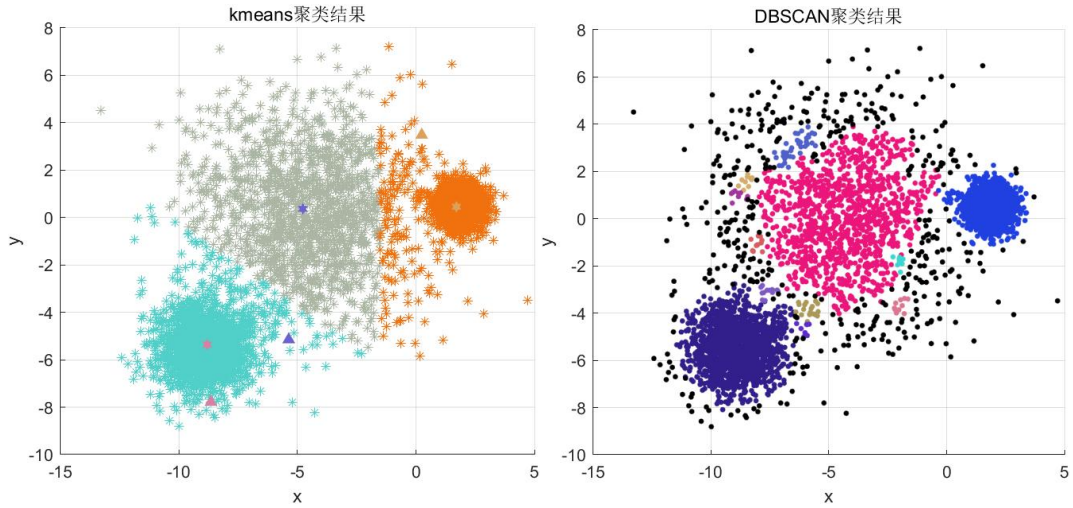
ε = 0.20，minPots = 5、10、15、20、25 时，data3 的 DBSCAN 聚类结果分别如下图所示。

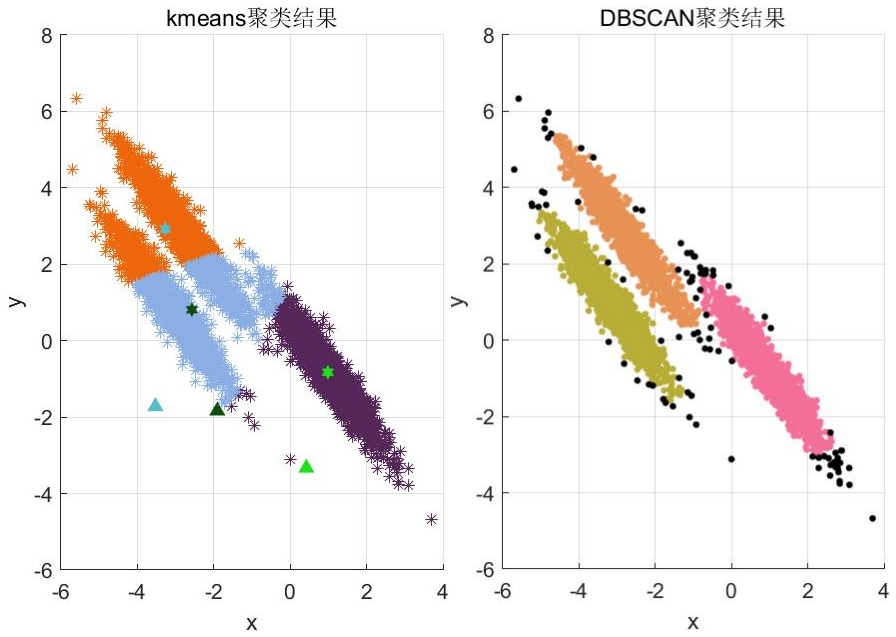


minPots 过小时对密度要求低，两点间容易可达，这会使得噪声点较少，但是可能会使得聚类间不合理地联通称为一个聚类，同时会出现一些点数很少（略大于 minPots）的小聚类。minPots 较大时相反，会使得噪声点较多。

3 聚类方法对比







对比可知，DBSCAN 算法的语义性更强，更适合不同类型之间区别明显的聚类划分问题，而 kmeans 算法更简单直接。

计算速度上来讲，kmeans 算法更快。

聚类形状上，欧式距离度量下，kmeans 聚类形状为圆形，难以适应 data1 双弧形结构等数据。DBSCAN 以小邻域逐步扩张聚类，可以适应多种结构的数据。

参数选取上，kmeans 算法中，k 及初始中心点的选取对结果会产生一定影响，可能影响结果稳定性；DBSCAN 算法中，同一数据集对于参数 ε 和 minPots 的敏感性相对不高，参数在合理区间内均能得到较好聚类效果。但不同的数据集需要调整出不同的参数，才能够合适。