清华大学《模式识别》

作业 8 - 非监督聚类算法

郭一隆

June 7, 2018

1 问题描述

- K-means 算法
  - 提供数据：`testSet.txt`
    文件包含 60 行 2 维数据，每行代表一个样本点，分布如图1所示。

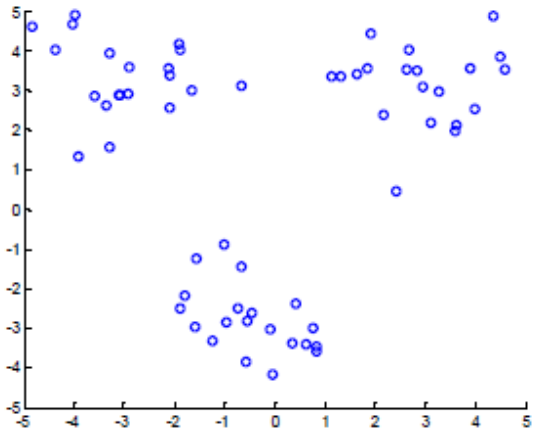


Figure 1: 数据集分布

  - 要求：

对这组数据进行 K-means 聚类，令 $k=2,3,4$。画出聚类结果及每类的中心点，观察聚类结果。记录使用不同初始点时的聚类结果，收敛迭代次数及误差平方和。

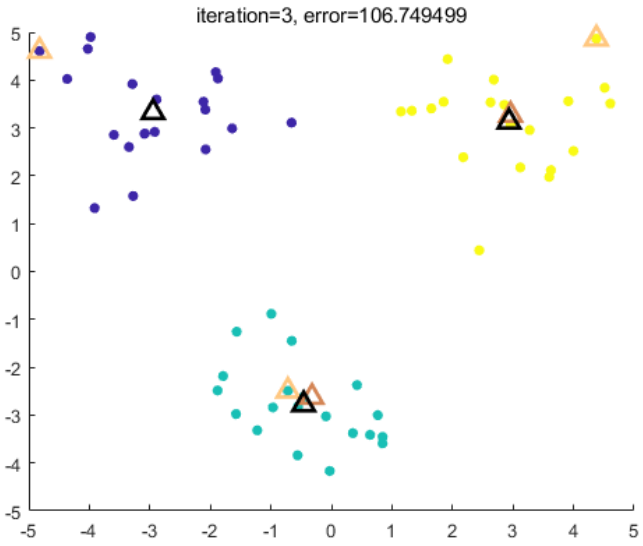
* $k=3$ 时，用给出几组初始点进行聚类
  - 初始点组 1：[−4.822 4.607; −0.7188 −2.493; 4.377 4.864]
  - 初始点组 2：[−3.594 2.857; −0.6595 3.111; 3.998 2.519]
  - 初始点组 3：[−0.7188 −2.493; 0.8458 −3.59; 1.149 3.345]
  - 初始点组 4：[−3.276 1.577; 3.275 2.958; 4.377 4.864]
* $k=2,4$ 时，自行给出初始点并聚类，观察聚类结果。

2 问题求解

2.1 $k=3$ 聚类

- 初始中心点：[−4.822 4.607; −0.7188 −2.493; 4.377 4.864]

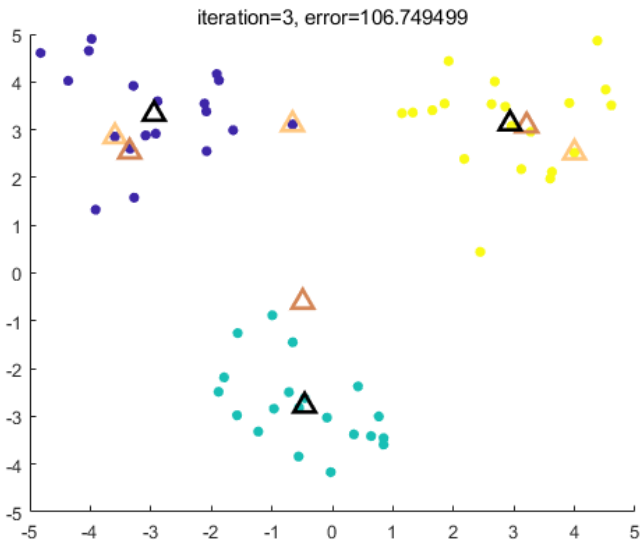
  3 次迭代后收敛，总误差平方和 =107



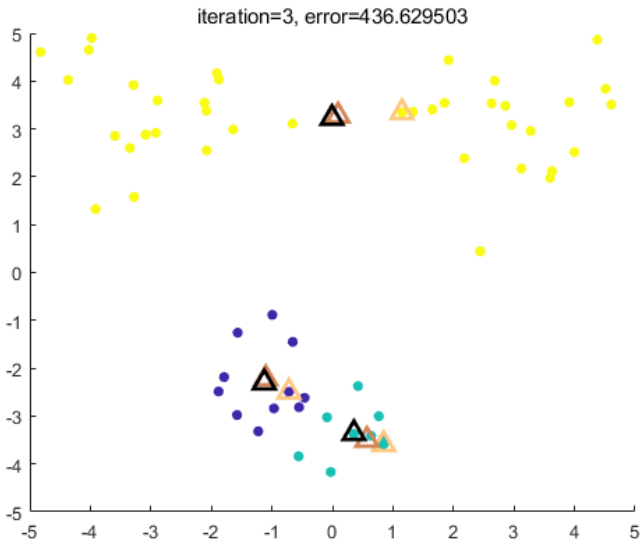
中心点随迭代次数变化的轨迹用“空心三角形”追踪显示，颜色变化采用 `copper`色系：

Figure 2: 初始 → 最终

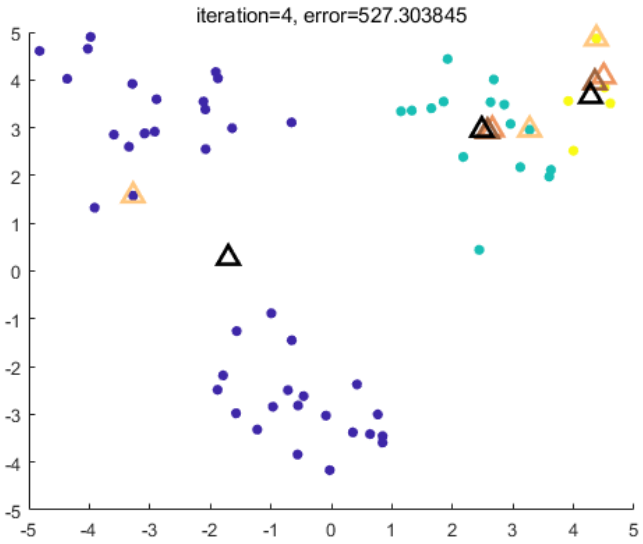
- 初始中心点：[−3.594 2.857; −0.6595 3.111; 3.998 2.519]
  3 次迭代后收敛，总误差平方和 =107，与初始点组 1 聚类结果相同。



- 初始中心点：[−0.7188 −2.493; 0.8458 −3.59; 1.149 3.345]
  3 次迭代后收敛，总误差平方和 =437，聚类结果不符合直觉。



- 初始中心点：[−3.276 1.577; 3.275 2.958; 4.377 4.864]
  4 次迭代后收敛，总误差平方和 =527，聚类结果同样不符合直觉。



2.2 $k=2$ 聚类

随机选取 k 个样本点作为初始中心点组，典型结果如下

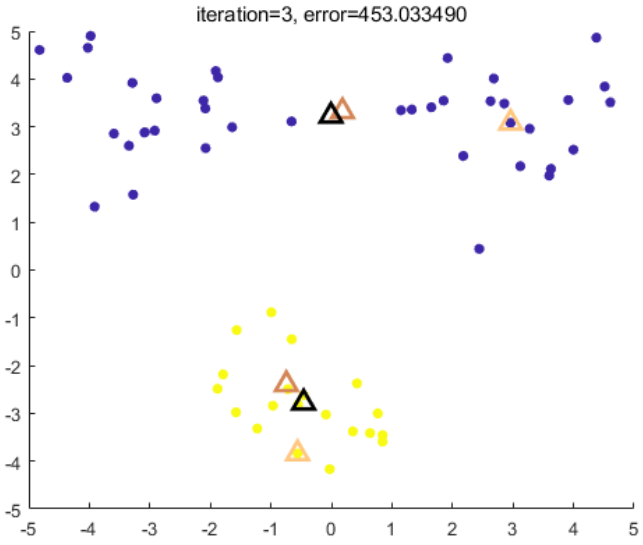


Figure 3: 多次运行所得到的误差平方和最小的结果

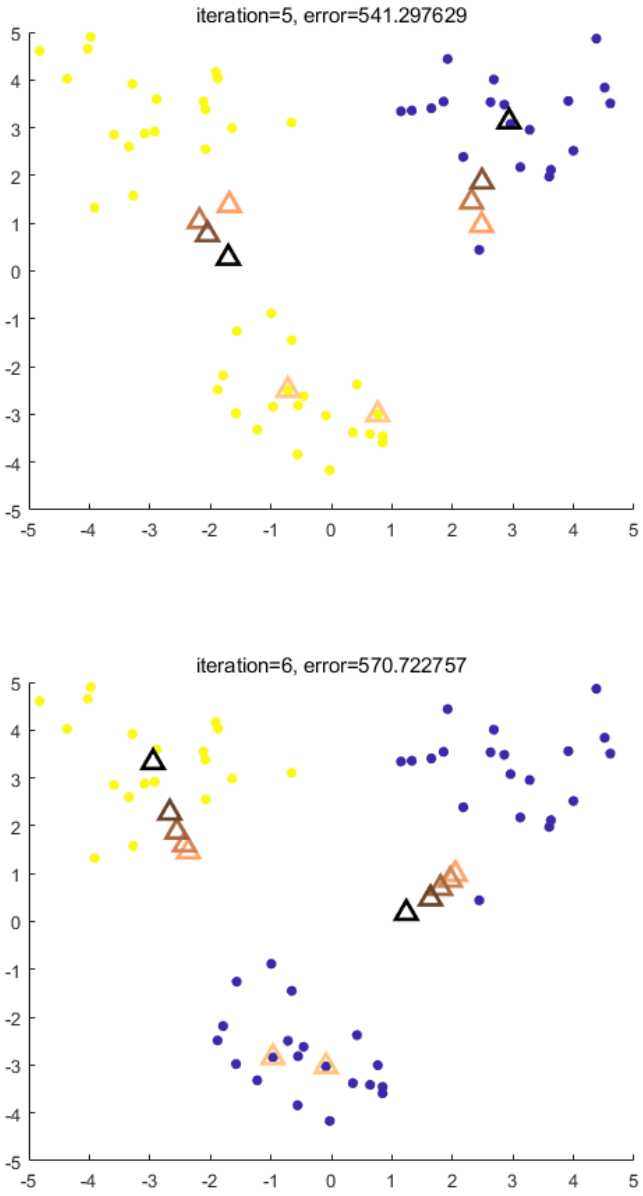
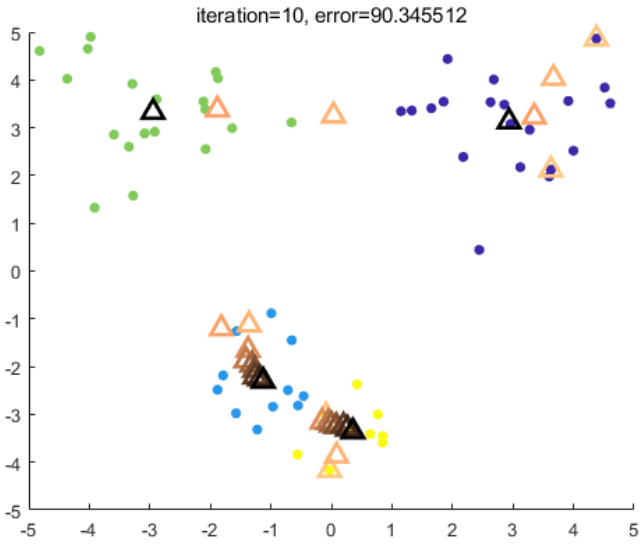
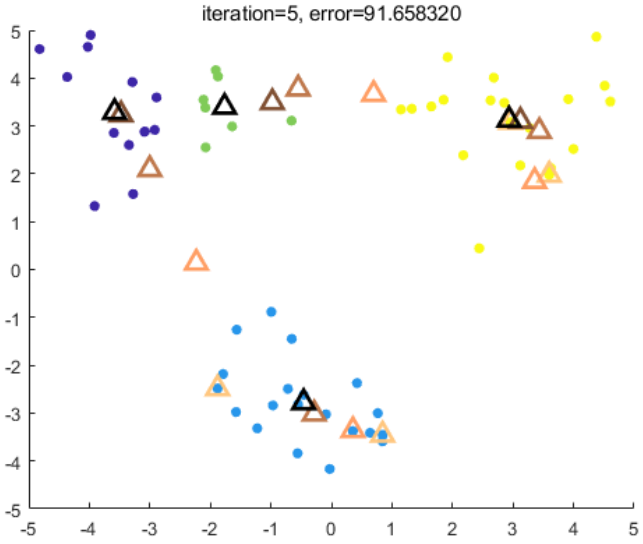


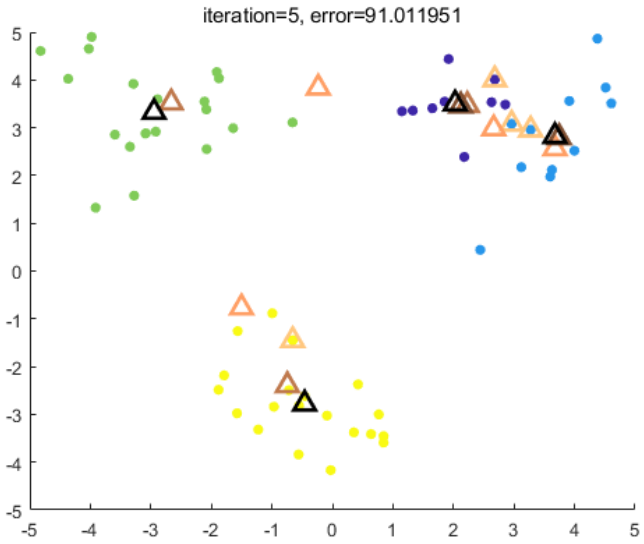
Figure 4: 即使初始中心点距离非常小，也会随着迭代次数被拉远

2.3 $k=4$ 聚类

对于 $k=4$ 的情况，结果如下


iteration=10, error=90.345512


iteration=5, error=91.658320



聚为 4 类的各结果误差平方和相差均不大，而且相比于 $k=2,3$ 的情况，误差平方和大幅下降（但这并不意味着分类效果好，需看实际是何种场景）。

3 总结

- 非监督聚类算法可以按一定的机理提取样本点的内在关联特征，有助于实际场景中发现有效的分类特征。
- 自然界中动物的学习模式大部分是无监督学习，希望有一天 SNN 等网络能在无监督学习领域取得突破，开创第三代神经网络新纪元。
- 回到 kmeans 算法，初值的选定对于聚类结果有很大的影响，尤其是对于本身类间区分度并不高的样本点。

附

- 仓库源码链接：https://github.com/Nuullll/pattern-recognition-assignment-8