

K-Means算法

聚类算法

聚类分析

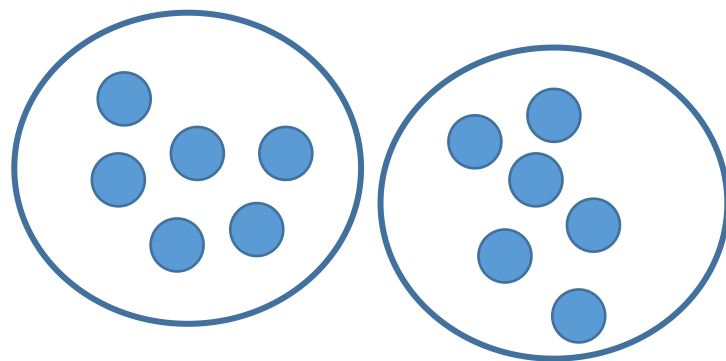
无监督学习： 需要数据，但没有数据标签（label）

发现数据中的规律（模式）

- 用户分群
- 行为聚类
-

聚类分析

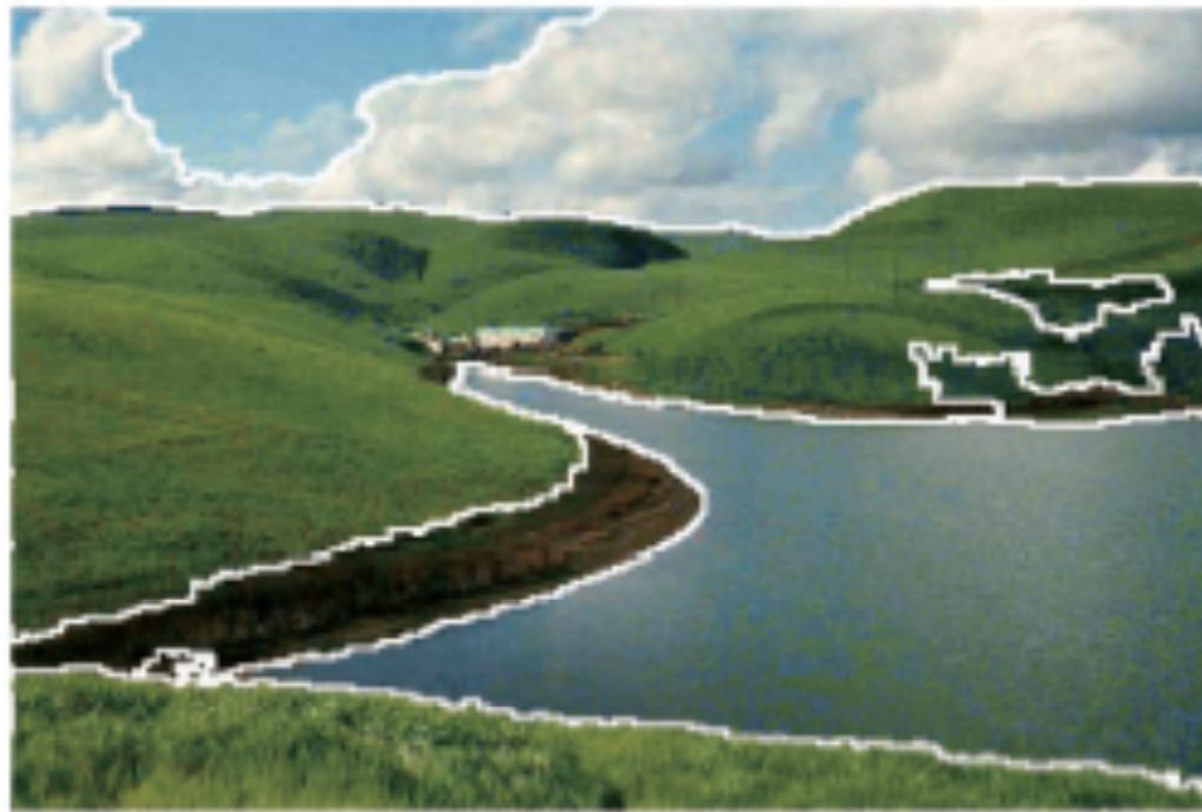
简单来讲，把相似的物体聚在一起



如何评估相似性？

应用场景

图像分割： 把图像分成相似的区域



应用场景

用户分层： 把拥有类似兴趣的用户聚在一起



K-means算法

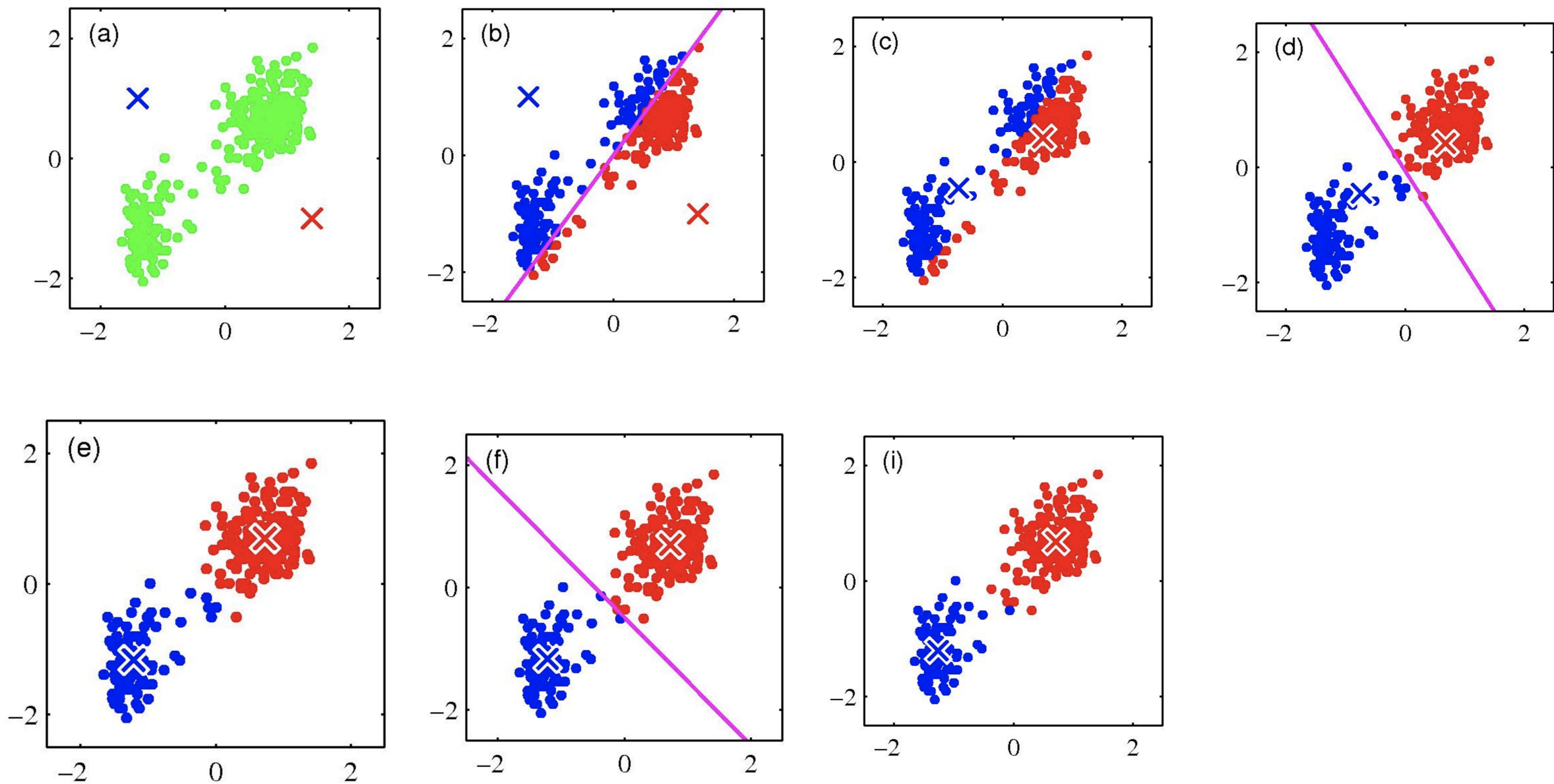
循环迭代式的算法

初始化：

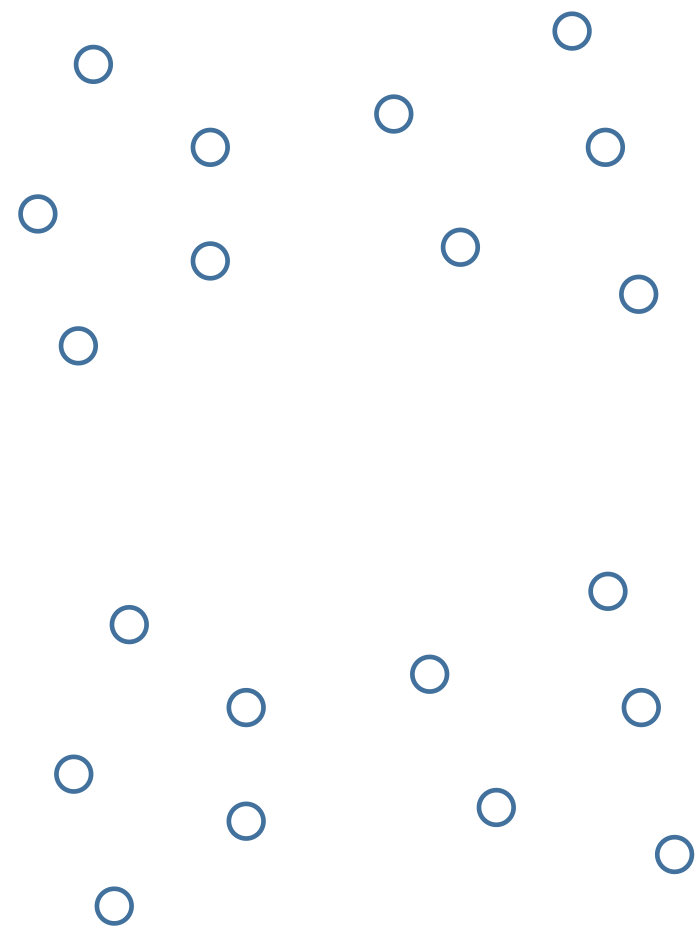
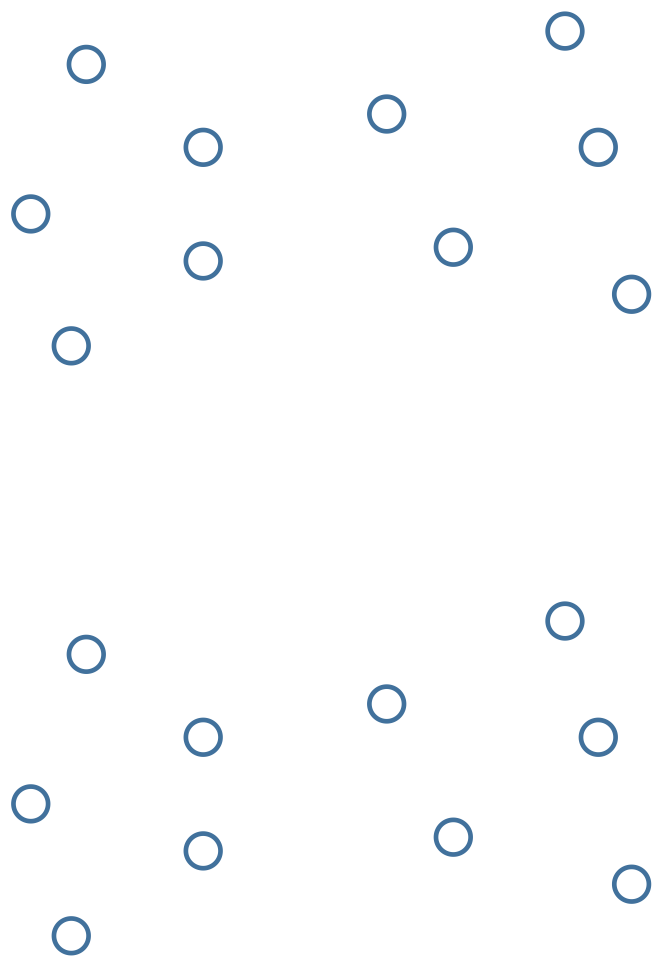
随机选择K个点，作为初始中心点，每个点代表一个group.

交替更新：

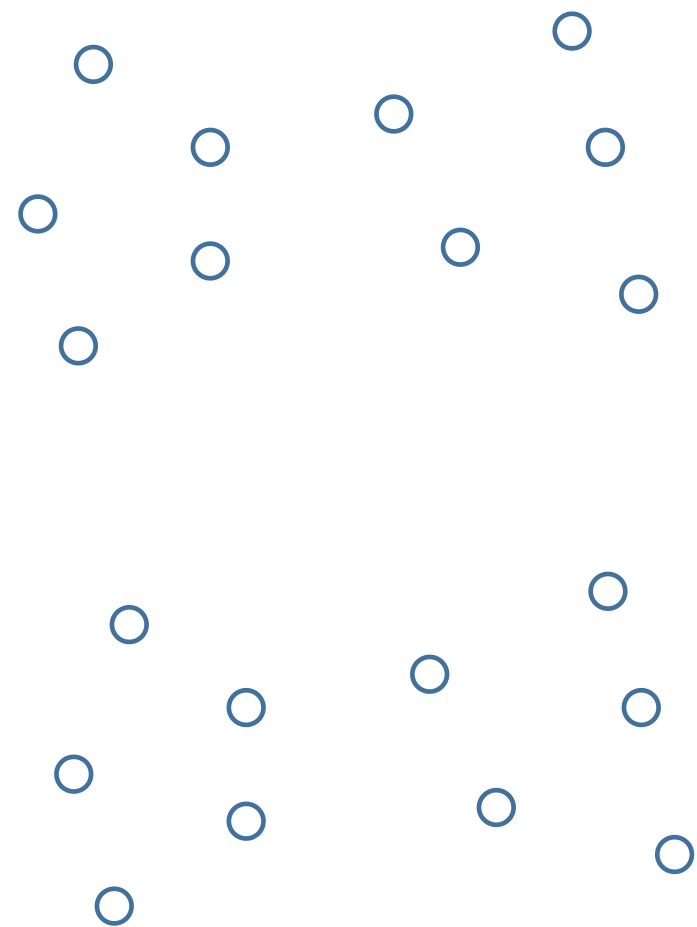
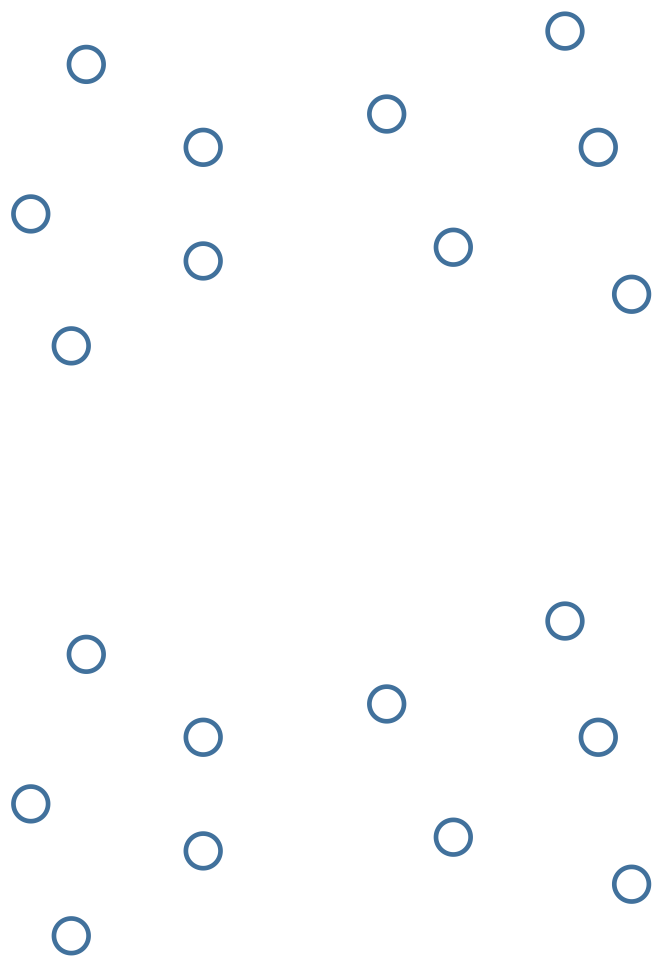
- 计算每个点到所有中心点的距离，把最近的距离记录下来并赋把group赋给当前的点
- 针对于每一个group里的点，计算其平均并作为这个group的新的中心点。



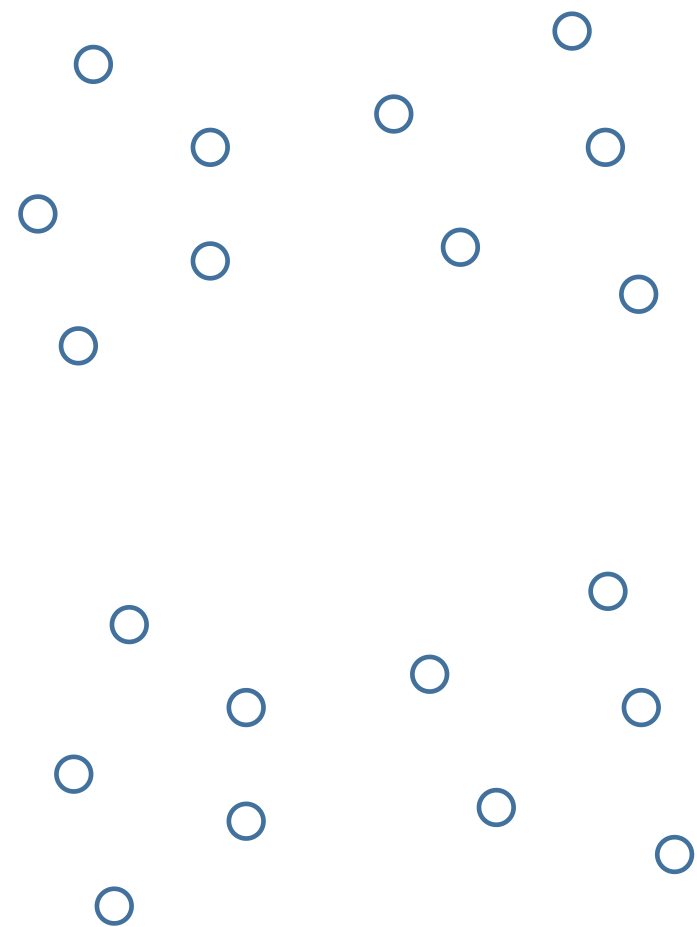
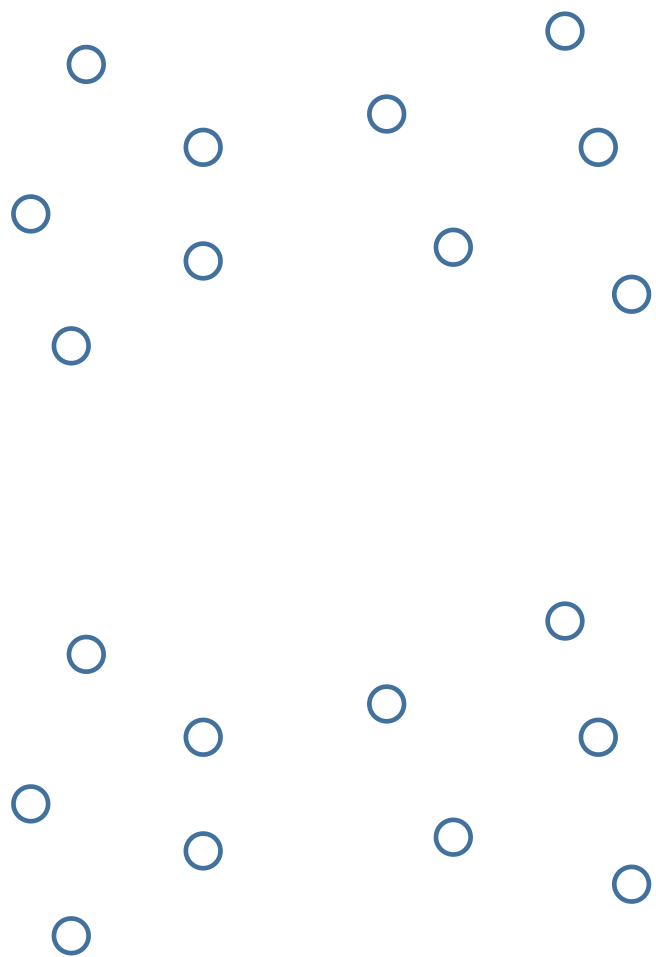
K-means 算法过程



K-means 算法过程



K-means 算法过程



K-means算法的一些特性

每一次迭代的复杂度

- 计算每个点到所有中心点的距离，把最近的距离记录下来并赋把group赋给当前的点

$O()$?

- 针对每一个group里的点，计算其平均并作为这个group的新的中心点

$O()$?

关于K-means几个问题

1. 一定会收敛吗？
2. 不同的初始化结果，会不会带来不一样的结果？
3. K-Means的目标函数是什么？
4. K如何选择？

问题1： 是否一定会收敛？

问题2：不同的初始化是否带来不一样的结果？

问题3： K-Means的目标函数

问题4：K值如何选择？