聚类

1、什么是无监督学习

- 一家广告平台需要根据相似的人口学特征和购买习惯将美国人口分成不同的小组，以便广告客户可以通过有关联的广告接触到他们的目标客户。
- Airbnb 需要将自己的房屋清单分组成不同的社区，以便用户能更轻松地查阅这些清单。
- 一个数据科学团队需要降低一个大型数据集的维度的数量，以便简化建模和降低文件大小。

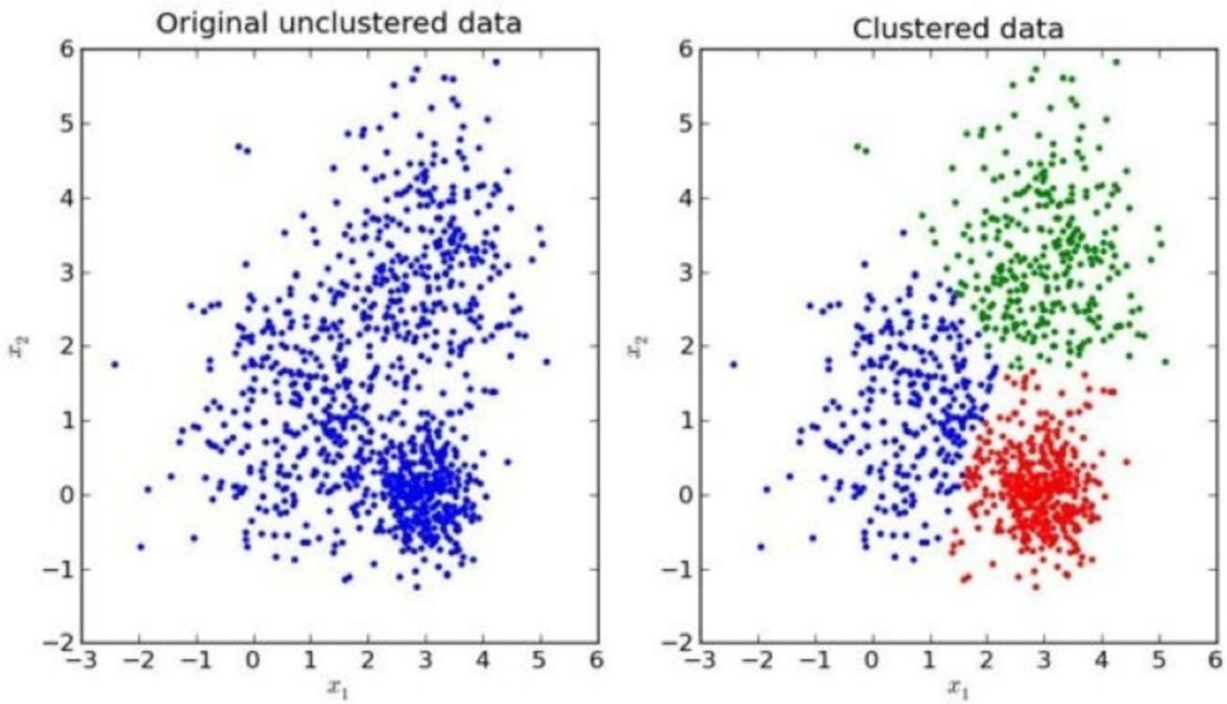
我们可以怎样最有用地对其进行归纳和分组？我们可以怎样以一种压缩格式有效地表征数据？**这都是无监督学习的目标，之所以称之为无监督，是因为这是从无标签的数据开始学习的。**

2、 无监督学习包含算法

- 聚类
  - K-means(K 均值聚类)
- 降维
  - PCA

3、 K-means 原理

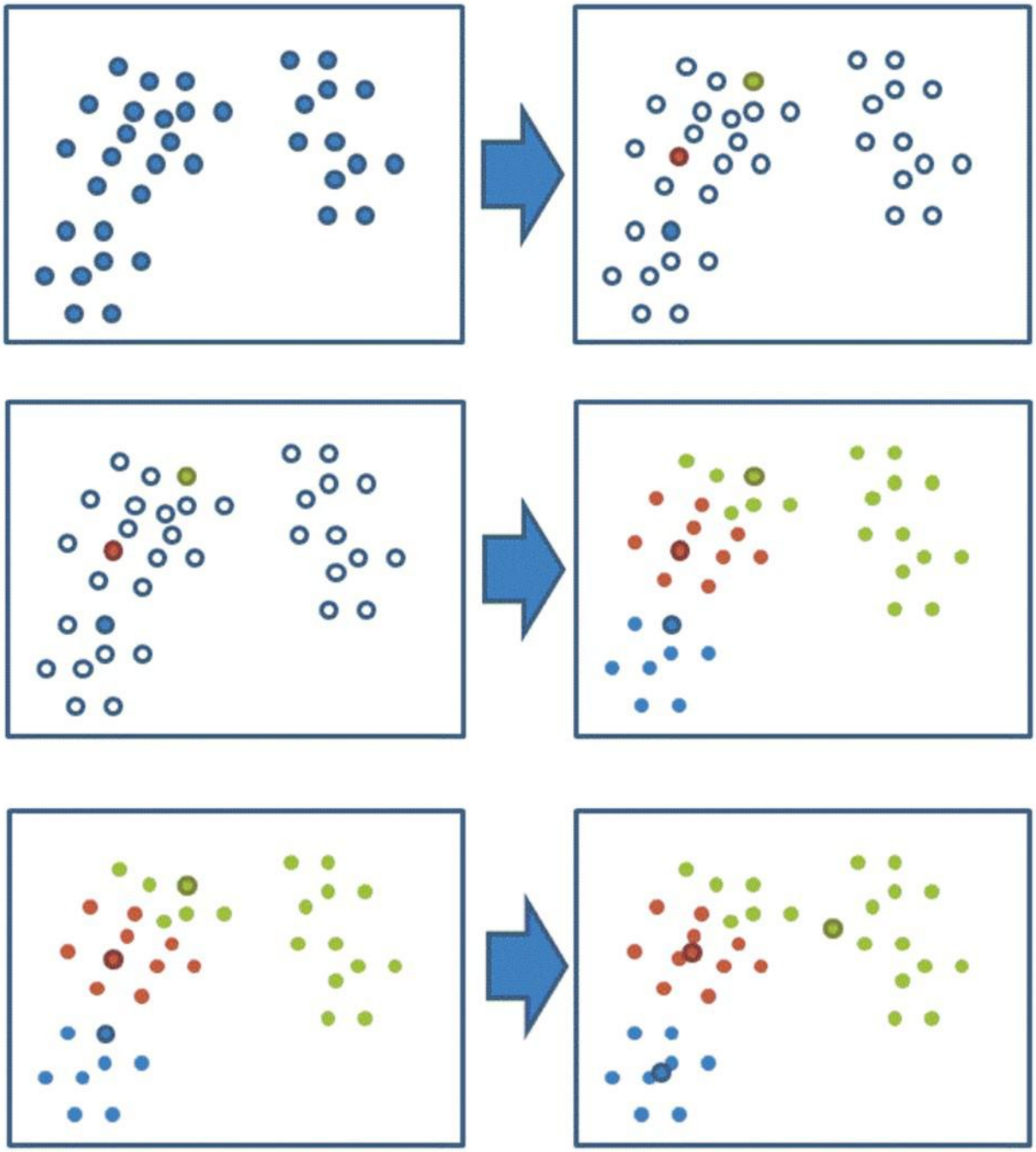
我们先来看一下一个 K-means 的聚类效果图



3.1 K-means 聚类步骤

- 1、随机设置 K 个特征空间内的点作为初始的聚类中心
- 2、对于其他每个点计算到 K 个中心的距离，未知的点选择最近的一个聚类中心点作为标记类别
- 3、接着对着标记的聚类中心之后，重新计算出每个聚类的新中心点（平均值）
- 4、如果计算得出的新中心点与原中心点一样，那么结束，否则重新进行第二步过程

我们以一张图来解释效果



4、K-meansAPI

- sklearn.cluster.KMeans(n_clusters=8,init='k-means++')
  - k-means 聚类
  - n_clusters:开始的聚类中心数量

- init:初始化方法，默认为'k-means ++’
- labels_:默认标记的类型，可以和真实值比较（不是值比较）