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。

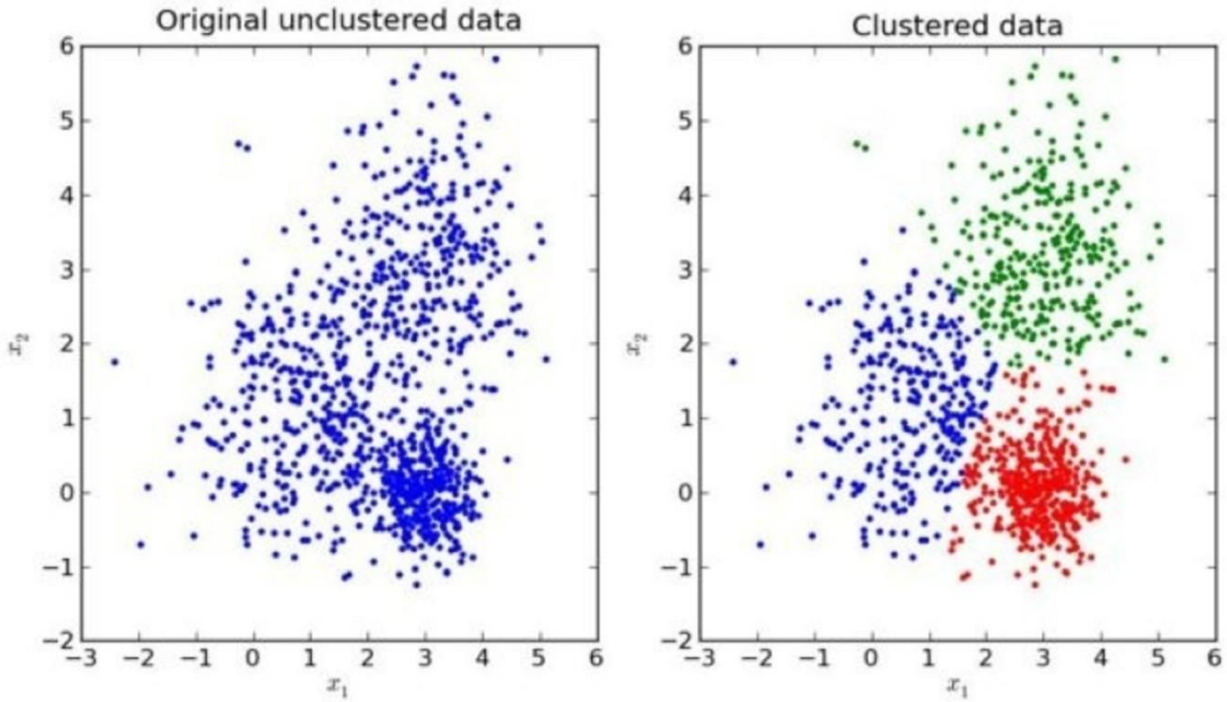
我们可以怎样最有用地对其进行归纳和分组？我们可以怎样以一种压缩格式有效地表征数据？**这都是无监督学习的目标，之所以称之为无监督，是因为这是从无标签的数据开始学习的。**

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

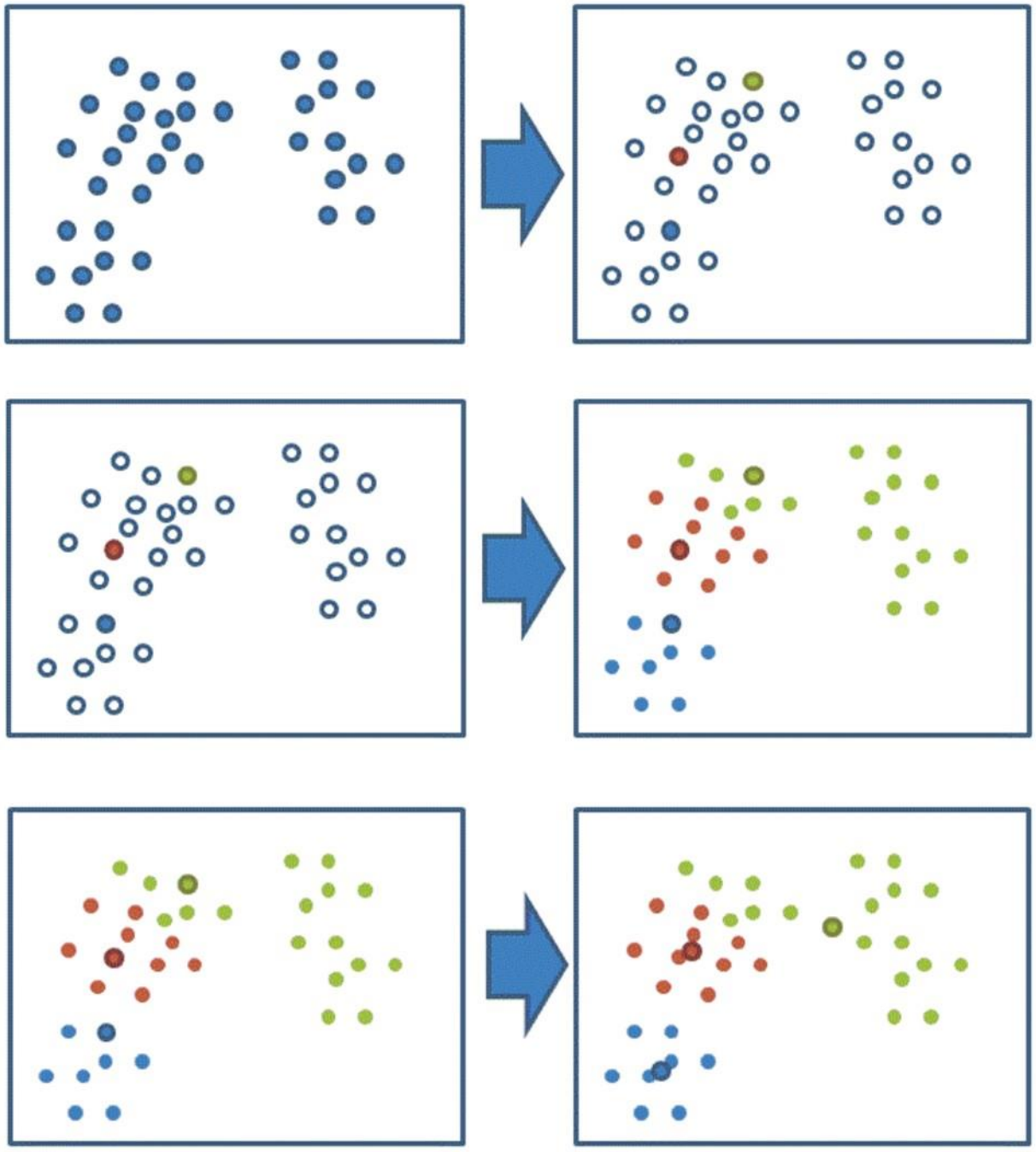
我们先来看一下一个 K-means 的聚类效果图



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）

5、聚类常用 api

1、K 均值

根据事先给定聚类数，为每个聚类随机分配中心点，计算所有样本与中心的距离，将每个样本分配到与其距离最近的中心点所在的聚类中，计算每个聚类的几何中心，用该中心作为新的聚类中心，重新划分聚类，直到计算出的几何中心与上一次的聚类使用的聚类中心重合或者足够接近为止。

示例代码：01-kmeans.py

特点：需要制定聚类数，聚类结果受到样本比例的影响，聚类中心的初始位置也会影响聚类结果。示例：图像量化 02-quant.py

2、均值漂移

把训练样本肯成是服从某种概率密度函数的随机分布，在不断迭代过程中试图寻找最佳的模式匹配，该密度函数的峰值点就是聚类中心，为该密度函数所覆盖的样本及隶属于该聚类。

特点：不需要事先给定聚类数，算法本身具有发现聚类数量的能力。

示例代码：03-shift.py

3-凝聚层次算法

凝聚层次聚类，可以是自下而上（聚），也可以自上而下（分）的。

1、在自下而上算法中，每个训练样本都是单独的集群，根据样本之间的相似度，将其不断的合并，直到集群达到事先指定的聚类数为止。

2、在自上而下算法中，所有的训练样本被看作是一个大的聚类，根据样本之间的差异度，将其不断拆分，直到集群数达到事先指定的聚类数为止。

特点：不同于其他基于中心的聚类算法，用它对一些在空间上具有明显连续性，但彼此间的距离未必最近的样本，可以优先聚集，这样所构成的聚类划分就能够表现出较强的连续性。

示例代码：04-spiral.py

4、DBSCAN 算法

朋友的朋友也是我的朋友

从任何一个训练样本出发，从一个事先给定的半径做圆，凡是不在此圆之外的样本都与圆心样本同类，再以这些同类样本为中心做圆，重复以上郭成，直到没有新的同类样本加入该聚类为止。

示例代码：05-dbscan.py

5、轮廓系数

表示聚类划分的内密外疏的程度。

轮廓系数主要以以下两个指标完成。

a:一个样本与其所在的聚类其他样本的平均距离（内密）

b:一个样本与其距离最近的另一个聚类中样本的平均距离（外疏）

轮廓系数 = (b – a)/ max(a,b)

**针对一个数据集，其轮廓系数就是其中所有样本的轮廓系数的平均值，轮廓系数介于（-1,1）之间。

1 表示完美聚类，-1 表示错误聚类，0 表示聚类重叠。

示例代码：06-score.py