# 第三讲 - 聚类分析

张建章

阿里巴巴商学院

杭州师范大学

2022-09-01

杭州師範大學
Hangzhou Normal University

新大陆
Newland

# 目录

# 目录

聚类是按照某个**特定标准** (如距离) 把一个数据集分割成不同的类或簇，使得**同一个簇内**的数据对象的**相似**性尽可能大，同时**不同簇**中的数据对象的**差异**性也尽可能地大。

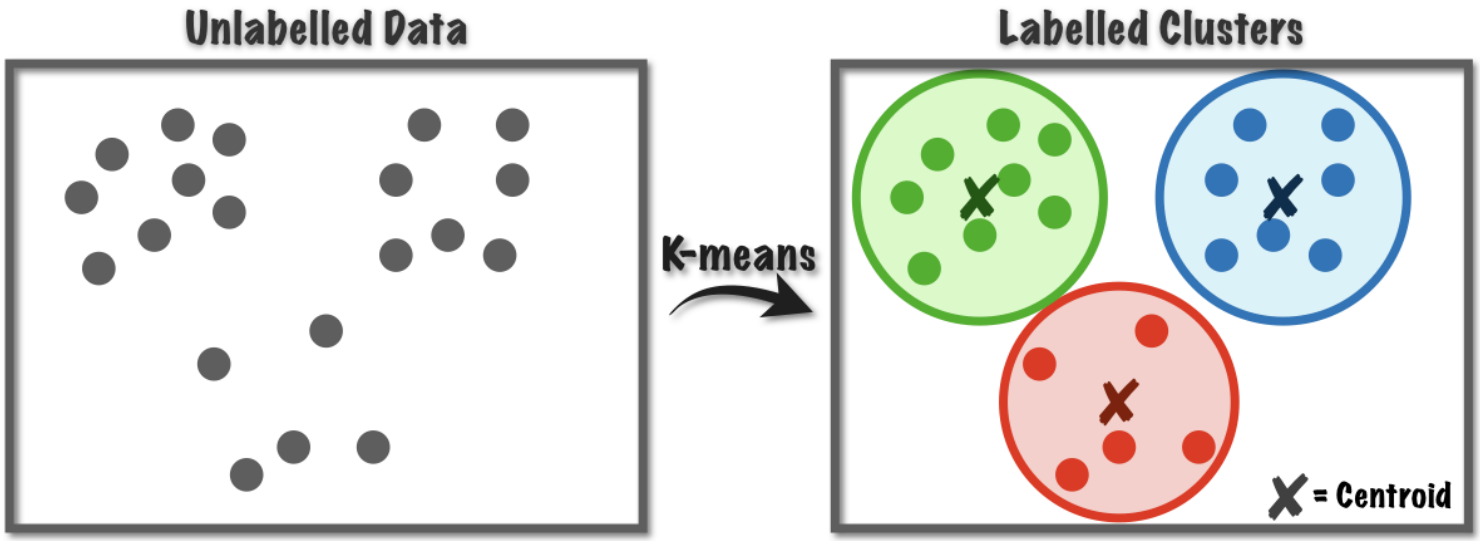


图 1: K 均值聚类示意图

由于不需要事先标注数据，聚类分析被归为**无监督学习**。

## 聚类与分类的区别

**相似：**分类和聚类都是根据样本特征，对样本进行归类。

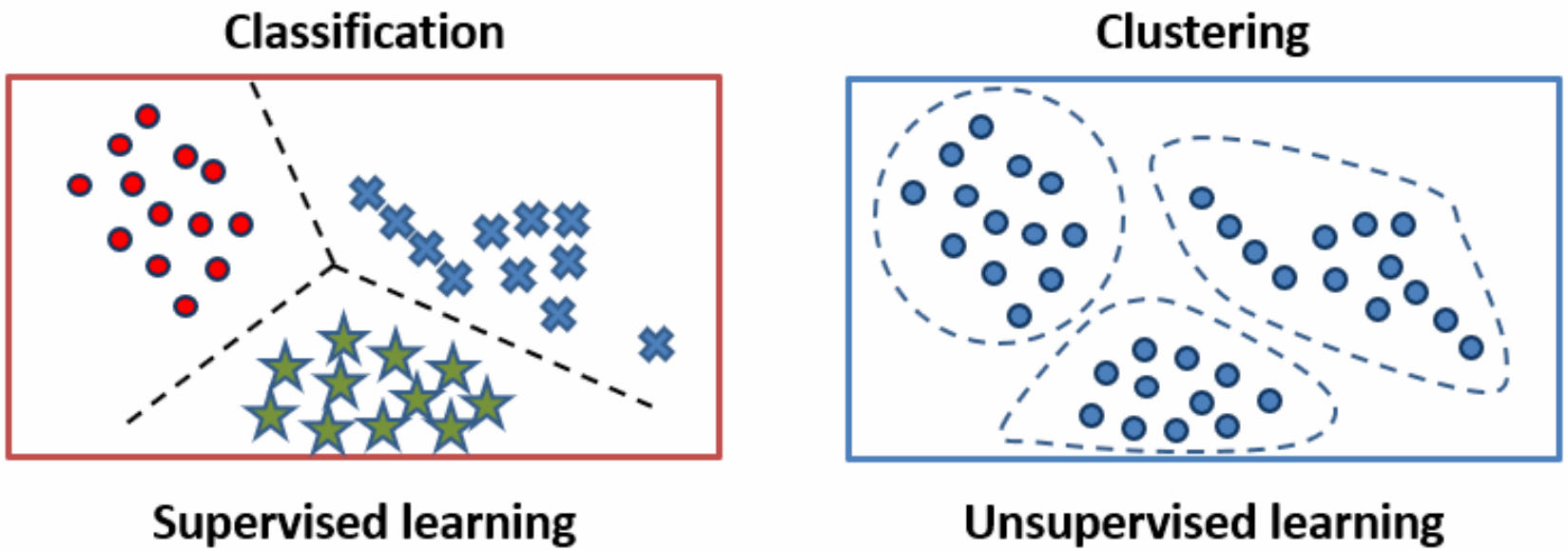


图 2: 聚类与分类的区别示意图

**差异：**① 分类需要事先定义类标签，聚类则不需要；② 分类需要标注训练数据集（监督学习），聚类则不需要（无监督学习）；③ 分类的类别标签和类别数量事先已知，聚类则不知道，因此聚类常用于探索性数据挖掘。

## 聚类分析的典型应用场景

- **目标用户的群体分类：**把目标群体划分成几个具有明显特征区别的细分群体，以进行精细化、个性化运营；
- **不同产品的价值组合：**把产品体系细分成具有不同价值、不同目的、多维度的产品组合，分别制定相应的产品开发计划；
- **探测、发现孤立点、异常值：**这些对象的行为特征与整体的数据行为特征很不一致，很多时候是风险的最大嫌疑和主要来源；

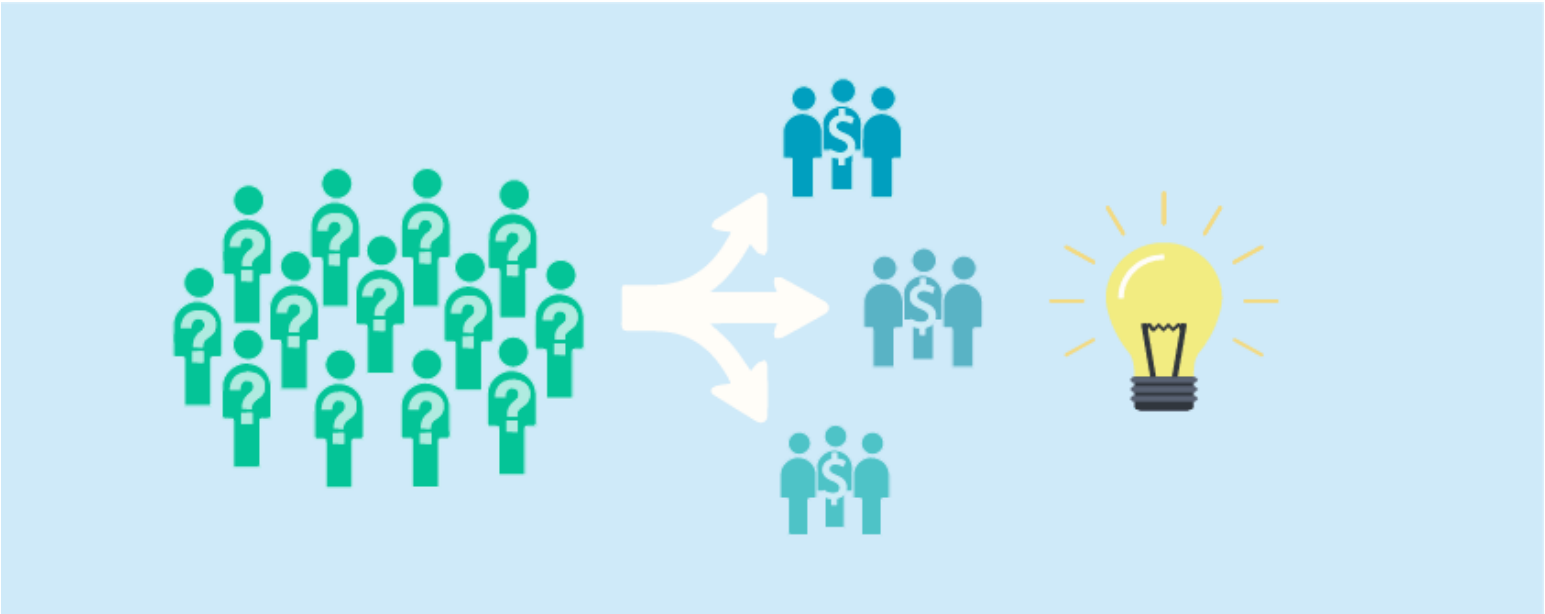

图 3: 客户聚类

# 目录

## 算法介绍

K 均值 (K-Means) 是目前最著名、使用最广泛的聚类算法，在给定一个数据集和需要划分的数目 k 后，该算法可以根据某个距离函数反复把数据划分到 k 个簇中，直到收敛为止，最常见的终止条件是误差平方和局部最小：

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

其中 $k$ 是类簇数，$x$ 是数据点，$S_i$ 为第 $i$ 个类簇点的集合，$\mu_i$ 为第 $i$ 个类簇的中心。

# 算法流程

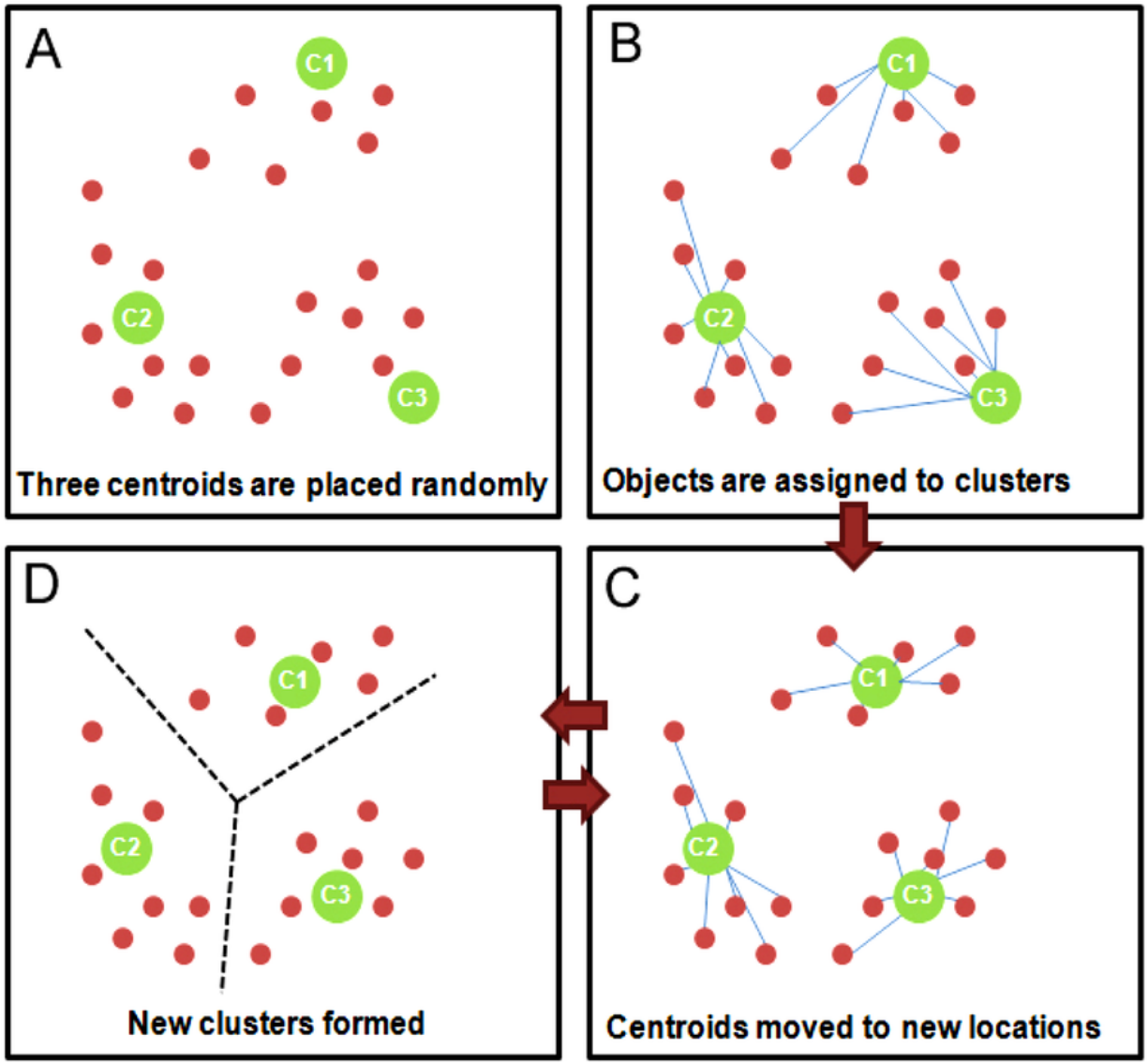


图 4: K 均值算法流程示意图

## 算法步骤

1. 选择聚类的个数 $k$;
2. 任意产生 $k$ 个聚类，并确定聚类中心，或者直接生成 $k$ 个中心;
3. 对每个点确定其聚类中心点;
4. 再计算其聚类新中心;
5. 重复以上步骤直到满足收敛要求。

$K$ 均值算法最常用的聚类函数为欧氏距离:

$$dis(p,q) = ||p-q||_2 = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

其中 $p$，$q$ 表示两个 $n$ 维数据点，$p_i$ 和 $q_i$ 为数据点的第 $i$ 个分量。

# 目录

## 1 - 准备工作

导入所需软件包 scikit-learn，pandas，使用 IPython 的魔法函数 `%matplotlib inline` 设置内嵌画图，加载并查看数据集。

```
import matplotlib.pyplot as plt
import pandas as pd
%matplotlib inline

customer_data = pd.read_csv('./data/ShoppingData.csv')
customer_data.head()
```

| | CustomerID | Genre | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|---|
| **0** | 1 | Male | 19 | 15 | 39 |
| **1** | 2 | Male | 21 | 15 | 81 |
| **2** | 3 | Female | 20 | 16 | 6 |
| **3** | 4 | Female | 23 | 16 | 77 |
| **4** | 5 | Female | 31 | 17 | 40 |

图 5: 前 5 条客户记录

## 2 - 数据转换

删除原始数据中的前三列，仅使用年收入和消费分两个维度对客户进行聚类，在二维平面上绘制数据点，通过观察数据点分布，选择合适的类簇数，由下图可以看出，将类簇数设置为 $k = 5$ 是一个合理的选择。

```
data = customer_data.iloc[:, 3:5].values
plt.scatter(data[:,0],data[:,1])
```

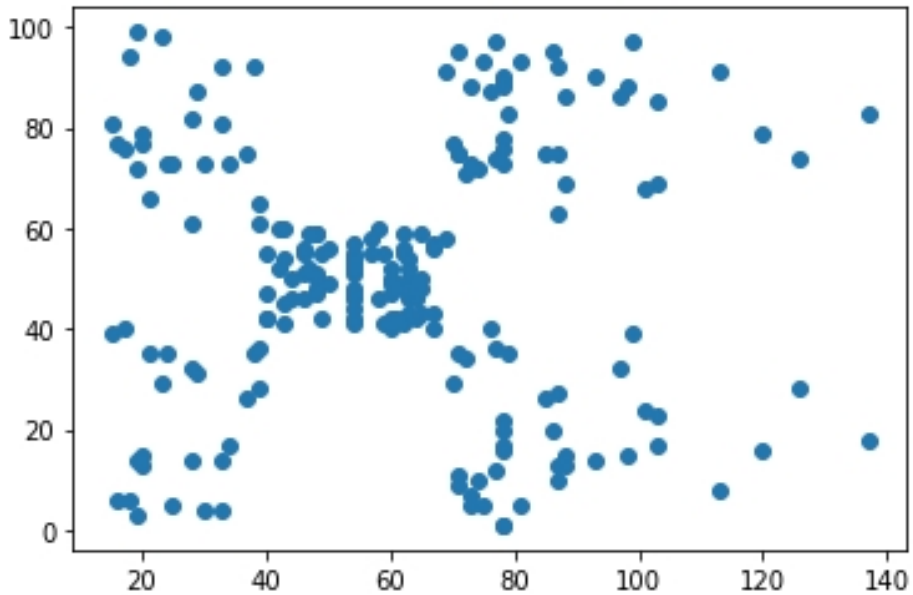


图 6: 原始客户数据点可视化

## 3 - K 均值聚类

使用 scikit-learn 中的 K-means 模型对客户数据点进行聚类，并为每个数据点分配类簇标签。

```
from sklearn.cluster import k_means
clusters = k_means(data,n_clusters=5)
clusters[1]
```

```
array([1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3,
       1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 4,
       1, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 0, 2, 0, 4, 0, 2, 0, 2, 0,
       4, 0, 2, 0, 2, 0, 2, 0, 2, 0, 4, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0,
       2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0,
       2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0,
       2, 0], dtype=int32)
```

图 7: 数据点类簇标签

聚类结果返回一个元组 (tuple)，第一个元素为类簇中心，第二个元素为每个数据点对应的类簇标签，需要进一步分析每个类簇中数据点的特征，来为每个类簇赋予有意义的解释。

## 4 - 聚类结果可视化分析

将数据点按照聚类结果绘制在二维平面上，为不同类簇的数据点着不同的颜色，横轴表示年收入，纵轴表示消费分。

```
plt.scatter(data[:,0],data[:,1], c=clusters[1], cmap='rainbow')
plt.title("Customer Clustering")
plt.xlabel("Annual Income")
plt.ylabel("Spending Score")
```

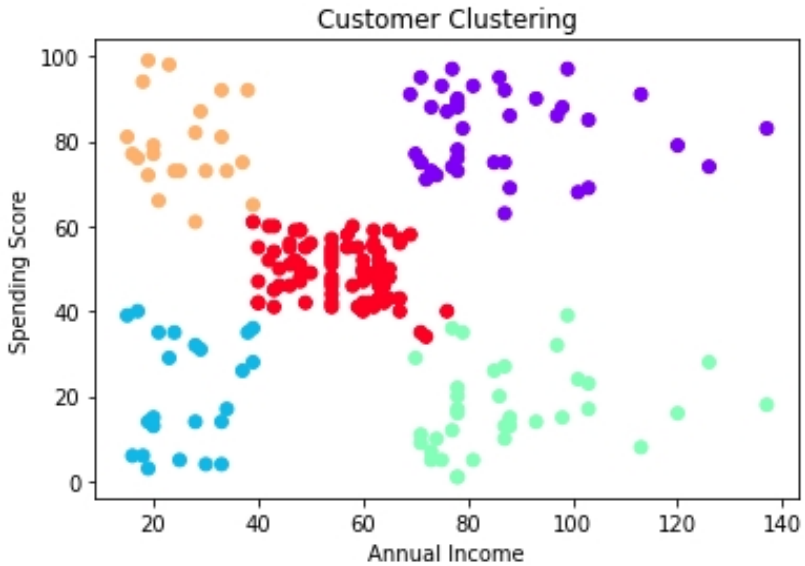


图 8: 客户聚类结果可视化

## 5 - 商业决策支持

从上图可以看到 5 种颜色的客户数据点代表 5 个不同的类簇，进一步分析年收入和消费分两个维度可知，5 个类簇对应 5 种不同特征的消费群体。

- 左下角的客户群体收入低，消费低，维持客户关系现状即可；
- 左上角的客户群体收入低，消费高，密切跟踪客户的资金状况，严格审批信用额度；

## 5 - 商业决策支持

- 右上角的客户群体收入高，消费高，是企业的现金牛客户，也是需要用心维系商业关系的大客户；
- 右下角的客户群体收入高，消费低，这类客户具有消费能力，需要有针对性地进行营销，激发其消费潜力；
- 正中间的客户群体收入中，消费中，且数据点众多，这类客户是企业的主要客户群体，需要积极维护客户关系，保持其持续的消费动力。

# 目录

# 汽车产品聚类分析

使用 UCI 汽车数据集 (数据详情和下载地址)，对 205 种汽车进行聚类，对于指定的车型，通过聚类分析找到其竞品车型，为产品定位提供决策支持。建议流程如下：

- 理解数据各字段含义，理解每条记录含义；
- 数据预处理，对不同类型的字段进行数值化和规范化；
- 选择合适的类簇数目；
- 数据聚类；
- 分析各类簇的车辆特点，为指定车型提供竞品分析。

**Automobile Data Set**

*Download*: Data Folder, Data Set Description

**Abstract**: From 1985 Ward's Automotive Yearbook

THE END