小灶商业数据分析Python训练营

聚类分析

目录

聚类方法的目标与应用

• 什么是聚类分析：

没有给定划分类别的情况下，根据样本相似度进行样本分组的一种方法，是一种非监督的学习算法

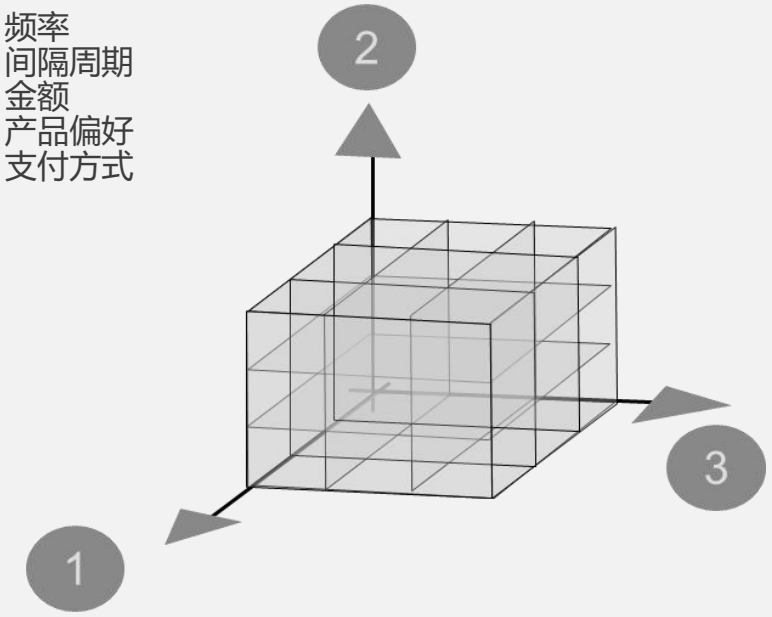
• 聚类的输入是一组未被标记的样本，聚类根据数据自身的距离或相似度划分为若干组，划分的原则是组内距离最小化而组间距离最大化

• 主要应用场景：客户分群，画像，标签体系

现有数据与可应用场景

购买记录数据

- 频率
- 间隔周期
- 金额
- 产品偏好
- 支付方式



个人信息

- 年龄
- 性别
- 建卡时间
- 建卡信息
- 车辆信息

需求和兴趣

- 网站咨询明细
- 来电信息（频次/内容）
- 客户工单信息

可应用场景

- **高价值客户识别**：基于现有数据中工单和购物等消费数据为核心，结合顾客车产信息，可识别高价值用户群体，进行针对性的营销活动
- **潜客转化分析**：针对客户咨询，进行对应产品推荐，提升潜客转化率和客单价
- **提升CRM质量**：结合顾客建卡和与公司往来情况，建立全生命周期管理流程，实现客户流失预警和其他KPI监控流程

小灶能力派

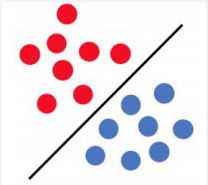
实现方法

聚类分析

- 收集和评估现有客户数据覆盖度和丰富性
- 基于现有数据，衍生新变量，如：客户最近一次访问/订单时间距今天数，历史平均客单价等
- 通过机器学习方法，结合业务经验选择合适的群组数量
- 分析各群组有代表性的特征，进行用户画像

倾向性模型

- 收集和评估现有客户数据覆盖度和丰富性
- 判断业务模型目标，通过机器学习方法，筛选关键变量，建立预测模型
- 模型应用于营销场景，并持续迭代优化，提升营销的各流程的精准度



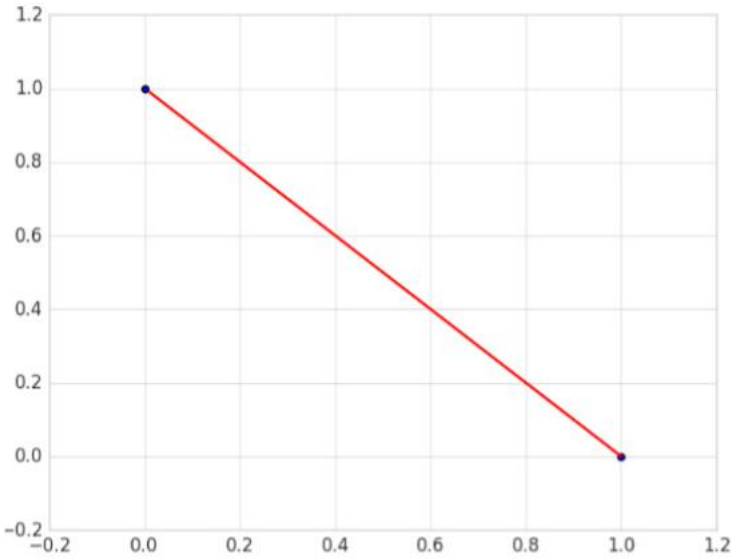
数据的标准化与距离计算

- **标准化数据**

  - 不同类型数据之间的差别可能很大，比如用户收入之间可能差别上万，但是对于年龄来说，是不可能有这么大的差距的
  - Z评分是将数据标准化的方法

$$Z_i = \frac{x_i - mean\ (x)}{std(x)}$$

- **计算数据的距离**



计算这三个用户之间的“距离”

Customer	Age (in years)	Income (in USD)
1	40	$40,000
2	40	$30,000
3	30	$40,000

和传统分群方法相比的优势

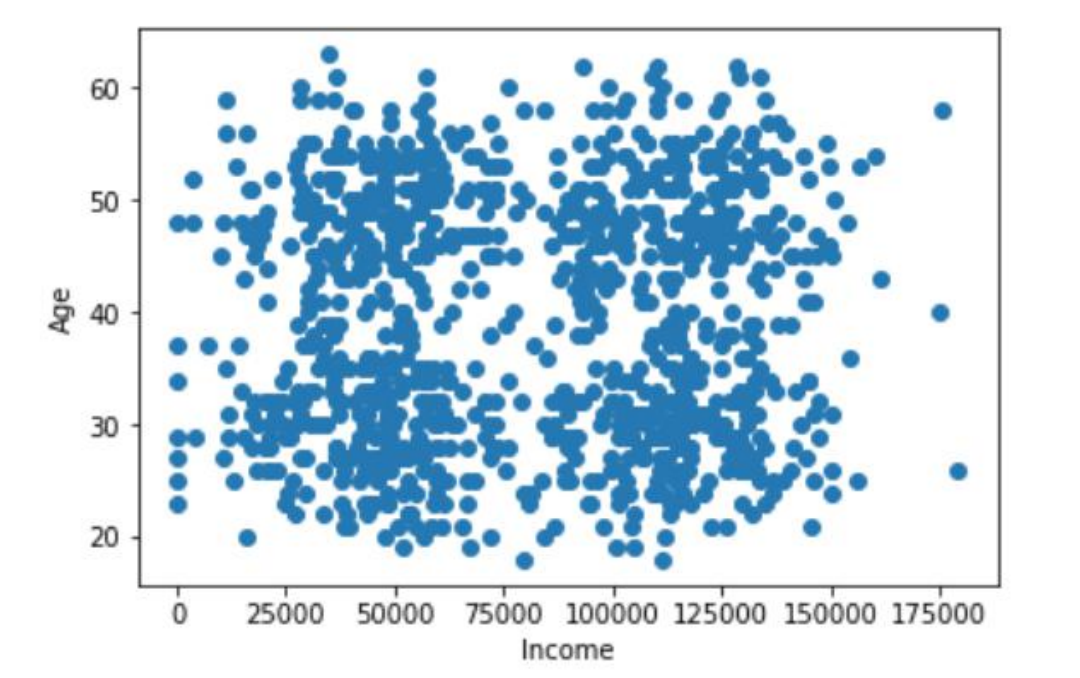
- **传统分群方法依据业务思路**
  - 比如RFM方法
  - 比如客户价值的20-80原则
- **聚类方法则可以**
  - 发现之前未知的因素，减少业务理解的狭隘性
  - 能对新的数据快速复制应用
- **聚类方法仍需要需要对分群结果进行解读，通过业务合理性来选择分群的数量**

kmeans方法的逻辑

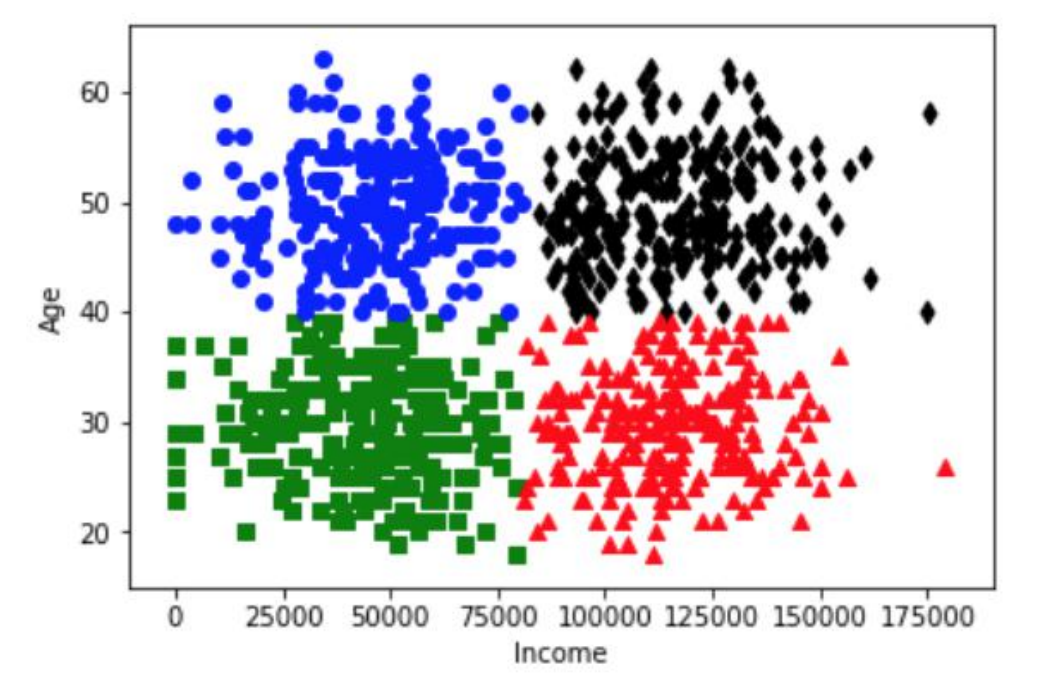
- 最常见的聚类方法,速度快，效率高，能适用于各种量级的数据
- 在最小化误差函数的基础上将数据划分为预定的类数K，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大
- 实际上能跟用户调整的就是k，有一定的局限性
- 在sklearn中快速调用kmeans

```
>>> from sklearn.cluster importKMeans
# 设置类数k
>>> k = 3
# 设置最大迭代次数
>>> iteration = 500
# 创建kmeans对象
>>> model = KMeans(n_clusters=k,n_jobs=4,max_iter=iteration)
# 使用数据训练训练model
>>> model.fit(data_zs)
```

Kmeans的小案例



聚类前



聚类后

小灶能力派

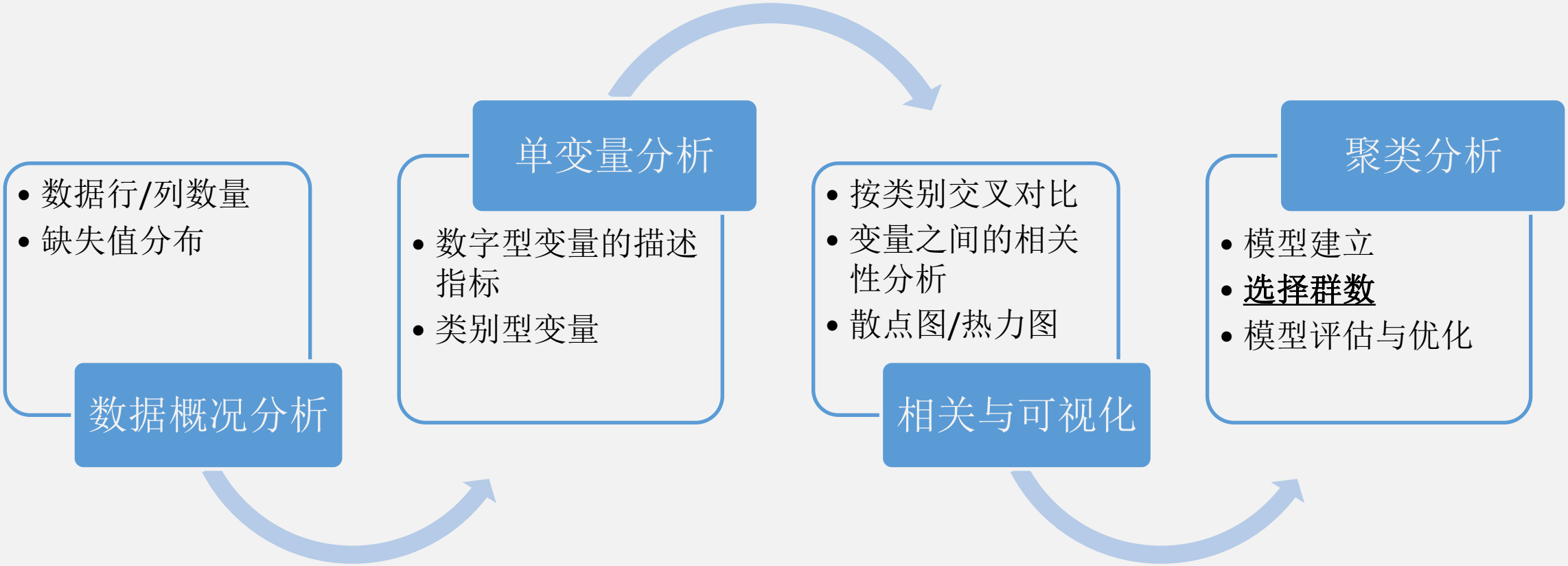
案例：airbnb数据环境下的客户分层

- Airbnb在全球拥有广泛丰富的用户出行场景。自身在app和网页端，以及通过各种营销渠道会收集到用户非常全面的行为数据。
- 通过这些数据，锁定潜在的目标客群并制定相应的营销策略是爱彼迎发展的重要基石。

数据解释

- id: 唯一的用户id
- date_account_created: 用户创建日期
- date_first_booking: 第一次订房日期
- Gender 性别
- Age 年龄
- signup_method 注册方式
- signup_flow: 注册来源页面
- Language: 语言偏好
- affiliate_channel: 是否通过某种营销渠道而来
- affiliate_provider: 营销渠道的名字
- first_affiliate_tracked: 引入的活动名称
- signup_app
- first_device_type 第一个设备
- first_browser 第一个浏览器
- country_destination 目的地国家
- Married 已婚
- Children 小孩数量

本课题分析流程



Kmeans实现客户分群

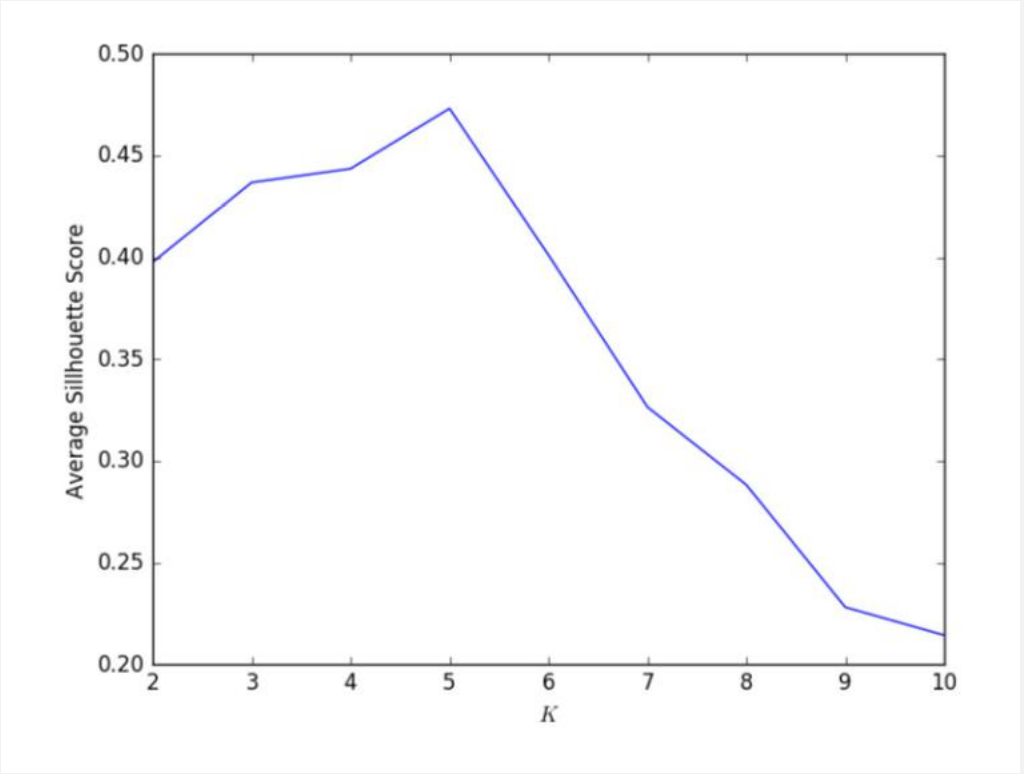
选择合理的分群数量

- **三个方面**
  - 可视化以及业务理解:分群之后是否有好的业务意义
  - 肘方法(elbow)：看各个群组是否有足够的区分
  - 标准评分：根据silhouette 分数来选择最好的k，取值范围为[-1,1],越大越好

小灶能力派

Silhouette评分

- 评分计算的是聚类模型的“效率”，这里分数就是越高越好。在右图对几种k值的聚类中也会发现最高值。



案例总结

最终会分为几个群体

- Ios深度用户
- 中文用户
- 喜欢安卓和moweb用户