Artificial Intelligence
人工智能实验

机器学习基础

中山大学计算机学院
2025年春季

目录

1.1 无监督学习

- 无监督学习和无监督学习的区别
  - **监督学习**
    - 在一个典型的监督学习中， **训练集有标签y** 。
    - 目标是找到能够区分正样本和负样本的决策边界， 需要据此拟合一个假设函数。
    - 监督学习常用于分类问题。
  - **无监督学习**
    - 在无监督学习中， 数据**没有附带任何标签y**。
    - 样本数据类别未知，需要根据样本间的相似性对样本集进行聚类，试图使类内差距最小化，类间差距最大化。
    - 无监督学习常用于聚类问题。

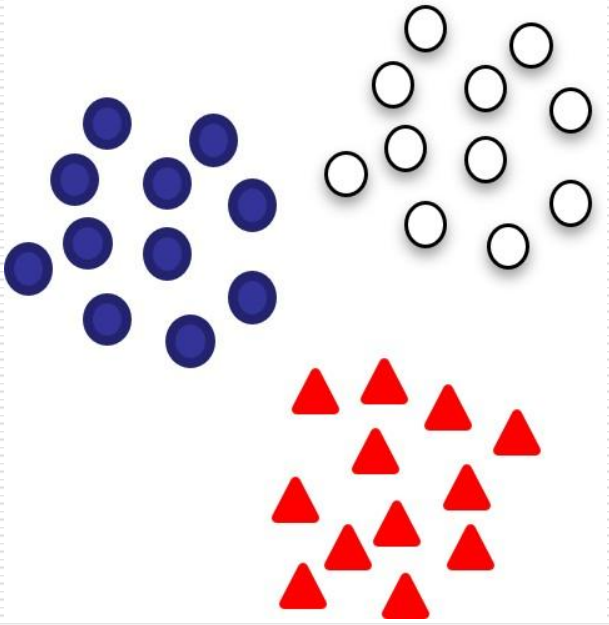
1.1 无监督学习

- 聚类
  - **主要算法**
    - K-means、密度聚类、层次聚类等
  - 主要应用
    - 市场细分、 文档聚类、 图像分割、 图像压缩、 聚类分析、 特征学习或者词典学习、 确定犯罪易发地区、 保险欺诈检测、 公共交通数据分析、 IT资产集群、 客户细分、 识别癌症数据、 搜索引擎应用、 医疗应用、 药物活性预测……

1.2 K-means聚类

背景知识

- 图中的数据可以分成三个分开的点集(称为簇)， 一个能够分出这些点集的算法， 就被称为聚类算法。



聚类算法示例

1.2 K-means聚类

算法概述

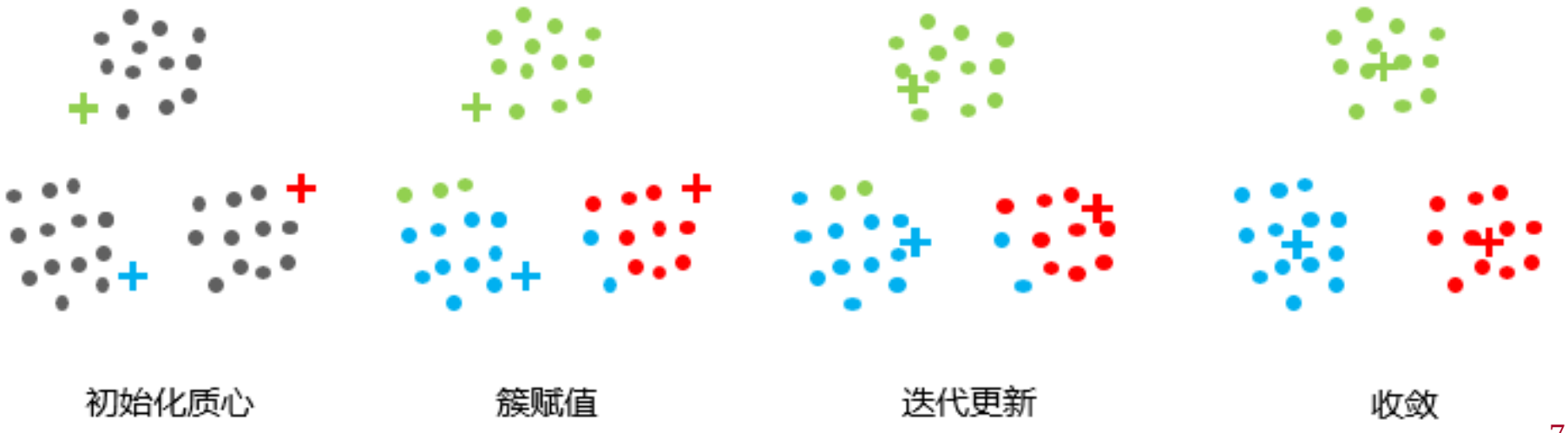
- K-means算法是一种无监督学习方法。
- 使用一个没有标签的数据集，然后将数据聚类成不同的组。
- 通过迭代将数据点分配到K个簇中，使得每个数据点与其所属簇的中心(质心)之间的距离平方和最小化。
- 距离度量：
  - 闵可夫斯基距离（Minkowski distance）
    - P=2即为欧氏距离
    - P=1则为曼哈顿距离
    - 当p取无穷时的极限情况下，可以得到切比雪夫距离

$$d(x,y)=(\sum_{i}|x_i-y_i|^p)^{\frac{1}{p}}$$

1.2 K-means聚类

算法流程

1. 选择K个点作为初始质心。（初始化簇质心为任意点）
2. 将每个点指派到最近的质心， 形成K个簇。（遍历所有数据点，计算所有质心与数据点的距离，根据距离选择簇）
3. 对于上一步聚类的结果， 对所有簇计算平均距离， 得出该簇的新的聚类中心（新的质心）。
4. 重复上述两步/直到迭代结束：质心不发生变化。



1.2 K-means聚类

- 优点
  - 原理简单，实现容易，收敛速度快
  - 聚类效果较优
  - 算法的可解释度比较强
  - 主要需要调参的参数仅仅时簇数K

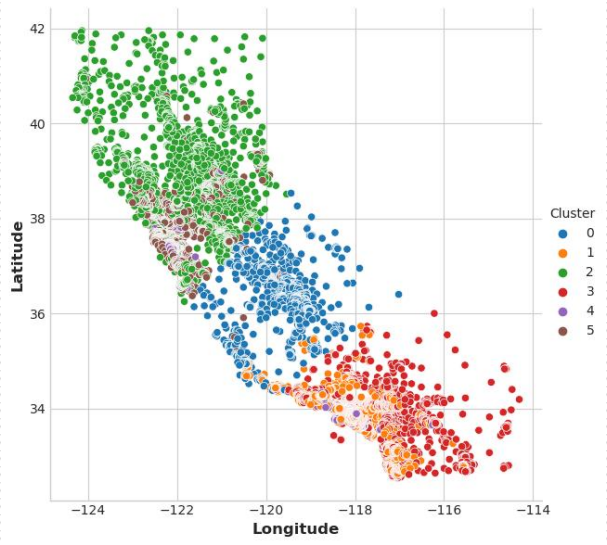
1.2 K-means聚类

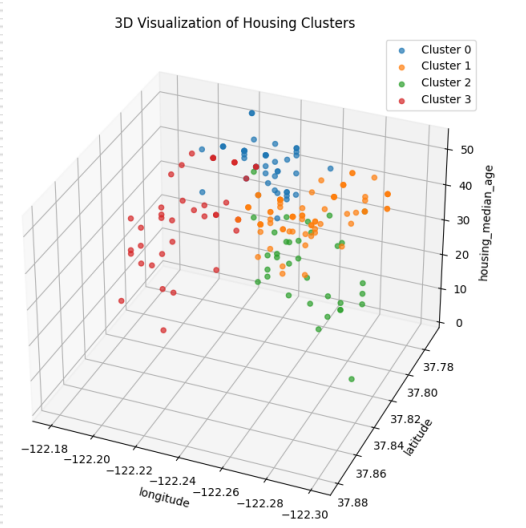
- 缺点
  - 需要预先指定簇的数量
  - 无法区分高度重叠的数据
  - 欧几里得距离限制了能处理的数据变量类型
  - 随机选择质心并不能带来理想的结果
  - 无法处理异常值和噪声数据
  - 不适用于非线性数据
  - 对特征尺度敏感
  - 如果遇到非常大的数据集，计算机可能会崩溃

2 实验任务

聚类任务（无需提交）

- 在给定数据集上，设计合适的k以及距离度量函数，利用k-means算完成数据的聚类。
- 要求：
  - 尝试分别在前200条、前1000条、前10000条数据完成聚类。
  - 画出聚类后的数据可视化图。
- 例图：





3. 作业提交说明

- 压缩包命名为：“学号_姓名_作业编号”，例：20250414_张三_实验4。
- 每次作业文件下包含两部分：code文件夹和实验报告PDF文件。
  - code文件夹：存放实验代码；
  - PDF文件格式参考发的模板。
- 如果需要更新提交的版本，则在后面加_v2，_v3。如第一版是“学号_姓名_作业编号.zip”，第二版是“学号_姓名_作业编号_v2.zip”，依此类推。
- 截至日期：**2025年4月28日晚24点**。
- 提交邮箱：zhangyc8@mail2.sysu.edu.cn。