机器学习基础

2022.05.19

目录

- 无监督学习
- K-means聚类
- 实验任务

1. 无监督学习

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， 我们的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。
- ✓ 无监督学习
  - ✓ 与此不同的是， 在无监督学习中， 我们的数据没有附带任何标签y， 无监督学习主要分为聚类、 降维、 关联规则、 推荐系统等方面

1. 无监督学习

主要方法

- ✓ 聚类（Clustering）
  - ✓ 如何将教室里的学生按爱好、身高划分为5类？
- ✓ 降维（Dimensionality Reduction）
  - ✓ 如何将将原高维空间中的数据点映射到低维度的空间中？
- ✓ 关联规则（Association Rules）
  - ✓ 很多买尿布的男顾客，同时买了啤酒，可以从中找出什么规律来提高超市销售额？
- ✓ 推荐系统（Recommender systems）
  - ✓ 很多客户经常上网购物，根据他们的浏览商品的习惯，给他们推荐什么商品呢？

1. 无监督学习

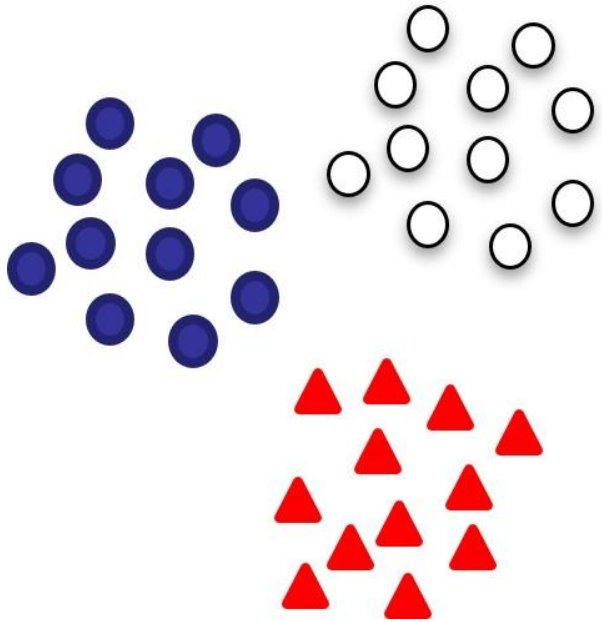
聚类

- ✓ 主要算法
  - ✓ K-means、密度聚类、层次聚类
- ✓ 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……

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

2. K-means聚类

算法概述

- ✓ K-means算法是一种无监督学习方法，是最普及的聚类算法，算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。K-means算法具有一个迭代过程，在这个过程中，数据集被分组成若干个预定义的不重叠的聚类或子组，使簇的内部点尽可能相似，同时试图保持簇在不同的空间，它将数据点分配给簇，以便簇的质心和数据点之间的平方距离之和最小，在这个位置，簇的质心是簇中数据点的算术平均值

2. K-means聚类

距离度量

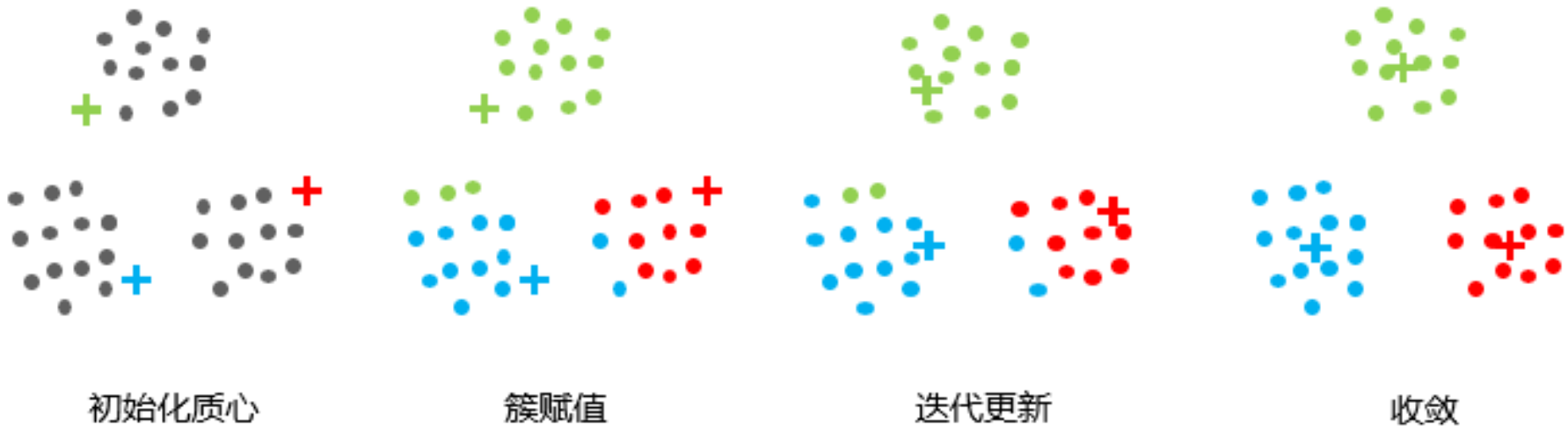
- ✓ 闵可夫斯基距离（Minkowski distance）
  - ✓ P取1或2时的闵可夫斯基时最为常用的
  - ✓ P=2即为欧氏距离
  - ✓ P=1则为曼哈顿距离
  - ✓ 当p取无穷时的极限情况下，可以得到切比雪夫距离

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

2. K-means聚类

算法流程

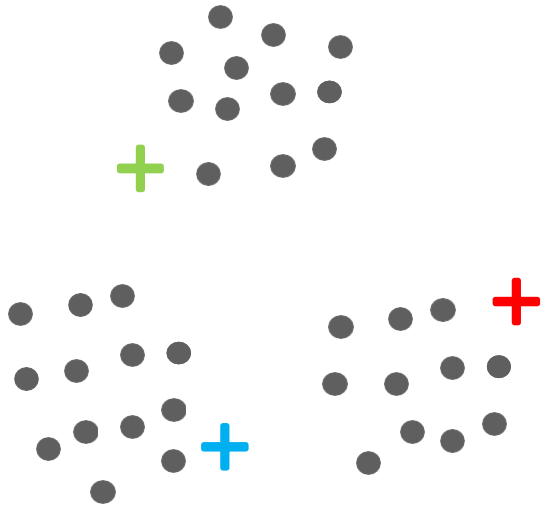
1. 选择K个点作为初始质心。
2. 将每个点指派到最近的质心， 形成K个簇。
3. 对于上一步聚类的结果， 进行平均计算， 得出该簇的新的聚类中心（新的质心）。
4. 重复上述两步/直到迭代结束：质心不发生变化。



2. K-means聚类

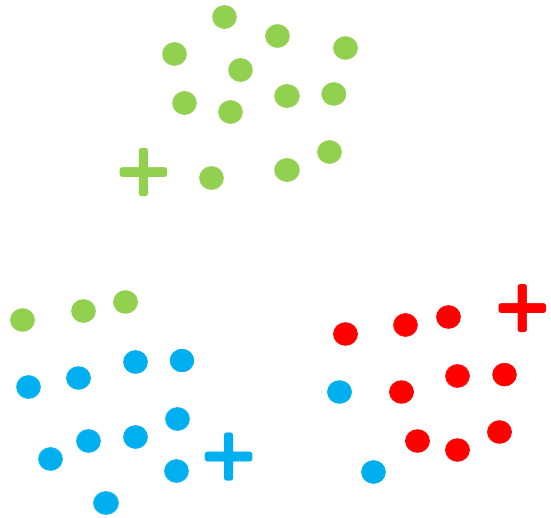
算法流程

1.初试化簇质心为的任意点。初试化时，必须注意簇的质心必须小于训练数据点的数目。



初始化质心

2.遍历所有数据点，计算所有质心与数据点的距离。这些簇将根据质心的最小距离而形成。



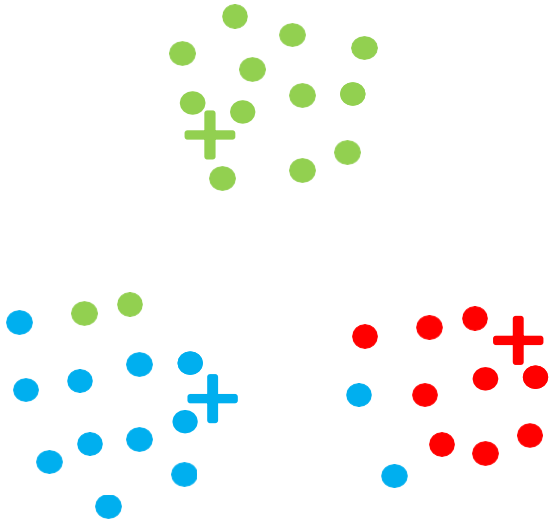
簇赋值

2. K-means聚类

算法流程

3.移动质心，因为上面步骤中形成的簇没有优化，所以需要形成优化的簇。为此需要迭代地将质心移动到一个新位置。取一个簇的数据点，计算平均值，然后将簇的质心移动到这个新位置。所有簇重复相同的步骤。

4.重复上述步骤，直至收敛 。



迭代更新

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

2. K-means聚类

缺点

- ✓ 需要预先指定簇的数量
- ✓ 如果有两个高度重叠的数据，那么它就无法区分，也不能判断有两个簇
- ✓ 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，那么计算机可能回崩溃

3. 实验任务

- 在给定数据集完成k-means算法聚类，画出聚类后的数据可视化图。（数据集课上发布）

- 要求
  - 设计合适的k以及距离度量函数，画出聚类后的数据可视化图。
  - 需要提交简要报告+代码
  - 压缩包：学号_姓名_作业编号.zip，如 20331234_张三_实验7.zip。
  - 截止日期：2022.5.25　23:59

附录

Matplotlib可视化教程:
https://www.runoob.com/matplotlib/matplotlib-tutorial.html

Thanks!