聚类分析KMEANS

目录

CONTENTS

01

基　本　概　念

4 学 习 类 型

监督学习

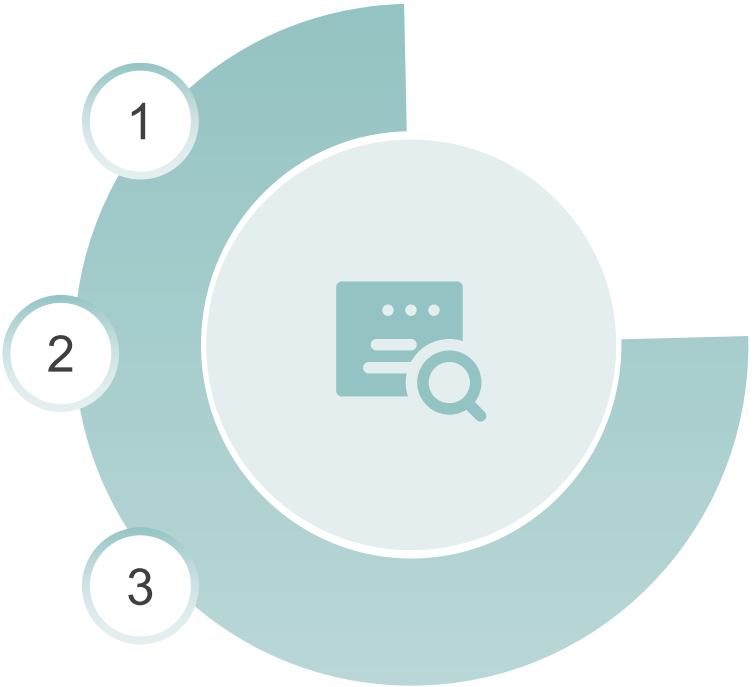
监督学习是指从有标注的数据中学习模型；比如：
分类模型
回归模型

无监督学习

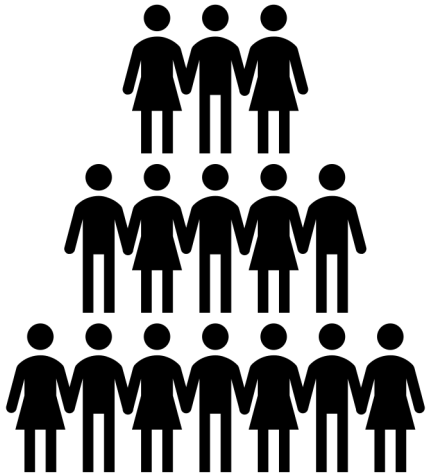
无监督学习是指从没有标注的数据中学习模型；比如：
聚类、降纬

半监督学习

版监督学习是指从部分有标注，部分没标注的数据中学习
模型。



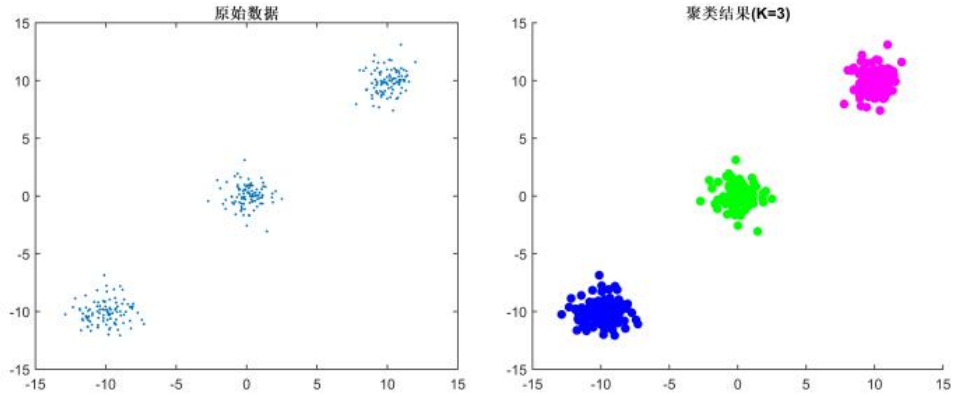
5 聚 类 应 用



聚类分群

6 聚 类 概 述

聚类是针对给定的样本，依据它们的特征的相似度或距离，将其归并到若干个“类”或“簇”的数据分析问题。一个类是给定样本集合的一个子集。



7 聚 类 案 例

Wholesale customers data

Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
2	3	12669	9656	7561	214	2674	1338
2	3	7057	9810	9568	1762	3293	1776
2	3	6353	8808	7684	2405	3516	7844
1	3	13265	1196	4221	6404	507	1788
2	3	22615	5410	7198	3915	1777	5185
2	3	9413	8259	5126	666	1795	1451
2	3	12126	3199	6975	480	3140	545
2	3	7579	4956	9426	1669	3321	2566
1	3	5963	3648	6192	425	1716	750
2	3	6006	11093	18881	1159	7425	2098
2	3	3366	5403	12974	4400	5977	1744
2	3	13146	1124	4523	1420	549	497

- Wholesale的案例数据表明了其消费者对不同类别商品的年度消费金额；
- 目标是根据消费记录客户进行类别划分

- FRESH: annual spending (m.u.) on fresh products (Continuous);
- MILK: annual spending (m.u.) on milk products (Continuous);
- GROCERY: annual spending (m.u.)on grocery products (Continuous);
- FROZEN: annual spending (m.u.)on frozen products (Continuous);
- DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous);
- DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
- CHANNEL: customers Channel – Horeca (Hotel/Restaurant/Cafe) or Retail channel (Nominal);
- REGION: customers Region – Lisnon, Oporto or Other (Nominal);

02

距 离 计 算

9 闵可夫斯基距离

定义： 给定样本集合X， X是 m 维实数向量空间$\mathbb{R}^m$中点的集合，其中，

$$x_i, x_j \in X, x_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}, x_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}}$$

样本x_i， x_j的闵可夫斯基距离定义为：

$$d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p \right)^{\frac{1}{p}}$$

闵可夫斯基距离越大相似度越小，距离越小相似度越大

当$p = 2$时称为欧式距离

$$d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^2 \right)^{\frac{1}{2}}$$

当$p = 1$时称为曼哈顿距离

$$d_{ij} = \sum_{k=1}^{m} |x_{ki} - x_{kj}|$$

10 马哈拉诺比斯距离

马氏距离：考虑各个特征之间的相关性并与各个特征的尺度无关。马氏距离越大相似度越小，距离越小相似度越大。

定义： 给定一个样本集合X， $X=[x_{ij}]_{m\times n}$，其协方差矩阵记作S。样本x_i和样本x_j之间的马氏距离d_{ij}定义为

$$d_{ij}=\left[(x_i-x_j)^{\mathrm{T}}S^{-1}(x_i-x_j)\right]^{\frac{1}{2}}$$

其中

$$x_i=(x_{1i},x_{2i},\cdots,x_{mi})^{\mathrm{T}},\quad x_j=(x_{1j},x_{2j},\cdots,x_{mj})^{\mathrm{T}}$$

当S为单位矩阵时，即样本数据的各个分量互相独立且各自分量的方差为1时，马氏距离就是欧式距离，所以马氏距离是欧式距离的推广。

11 相 关 系 数

相关系数：样本之间的相似度使用相关系数来表示。相关系数的值越接近1，表示样本越相似；越接近0，表示样本越不相似。

定义：样本x_i和样本x_j之间的相关系数

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\left[\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{m}(x_{kj}-\bar{x}_j)^2\right]^{\frac{1}{2}}}$$

12 余 弦 距 离

余弦距离：样本之间的夹角余弦来表示。夹角余弦的值越接近1，表示样本越相似；越接近0，表示样本越不相似。

定义：样本x_i和样本x_j之间夹角余弦

$$s_{ij}=\frac{\sum_{k=1}^{m}x_{ki}x_{kj}}{\left[\sum_{k=1}^{m}x_{ki}^2\sum_{k=1}^{m}x_{kj}^2\right]^{\frac{1}{2}}}$$

聚　类　算　法

14 聚 类 算 法

K均值聚类的两步骤：

1. 选择k个类的中心，将样本逐个指派与其最近的中心类中，得到一个聚类结果
2. 更新每个类的样本均值，作为类的新的中心

重复以上步骤，直到收敛为止。

15 聚 类 算 法

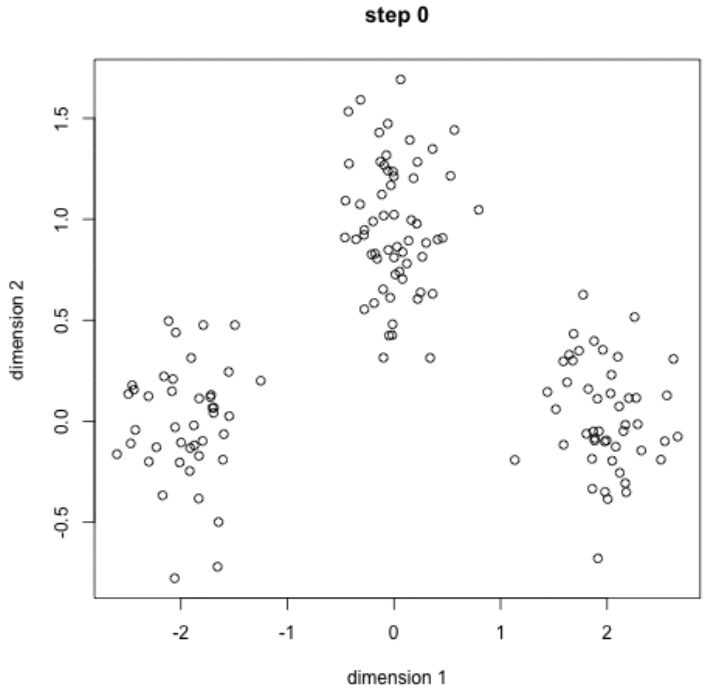
输入：n个样本的集合X

输出：样本集合的聚类

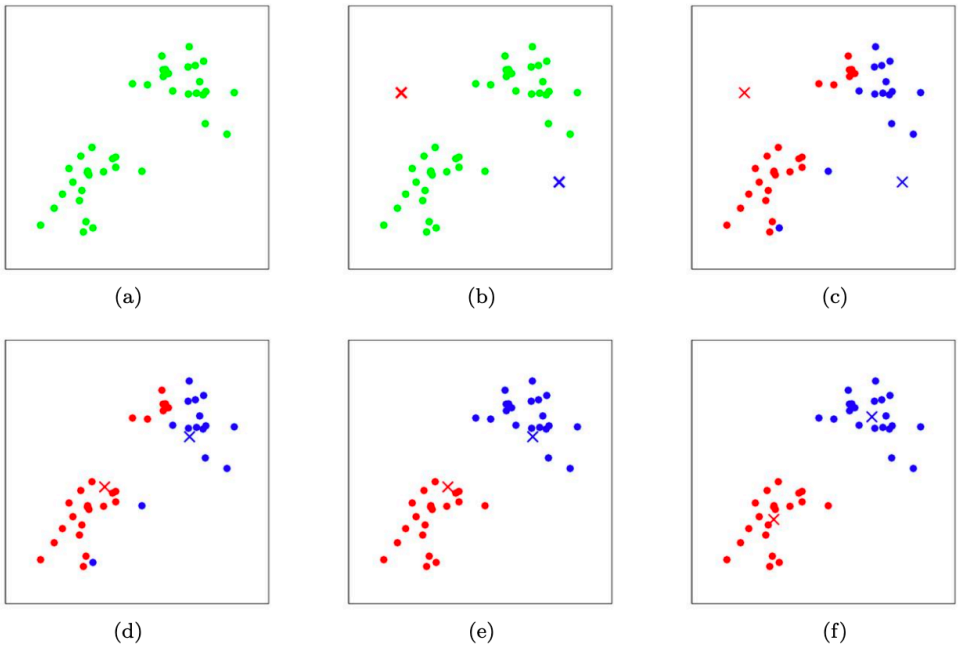
1. 初始化。令 t = 0，随机选择 k 个样本点作为初始聚类中心$\mathrm{m}^{(0)} = \left(\mathrm{m}_1^{(0)}, \mathrm{m}_2^{(0)}, \ldots, \mathrm{m}_\mathrm{k}^{(0)}\right)$
2. 对样本进行聚类。对固定的类中心$\mathrm{m}^{(\mathrm{t})} = \left(\mathrm{m}_1^{(\mathrm{t})}, \mathrm{m}_2^{(\mathrm{t})}, \ldots, \mathrm{m}_\mathrm{l}^{(\mathrm{t})}, \ldots, \mathrm{m}_\mathrm{k}^{(\mathrm{t})}\right)$，其中$\mathrm{m}_\mathrm{l}^{(\mathrm{t})}$为类中心，计算每个样本到类中心的距离，将每个样本指派到与其最近的中心的类中，构成聚类结果$\mathrm{C}^{(\mathrm{t})}$
3. 计算新的类的中心。对聚类结果$\mathrm{C}^{(\mathrm{t})}$，计算当前各个类中的样本的均值，作为新的类中心

   $\mathrm{m}^{(\mathrm{t}+1)} = \left(\mathrm{m}_1^{(\mathrm{t}+2)}, \mathrm{m}_2^{(\mathrm{t}+1)}, \ldots, \mathrm{m}_\mathrm{l}^{(\mathrm{t}+1)}, \ldots, \mathrm{m}_\mathrm{k}^{(\mathrm{t}+1)}\right)$,
4. 如果迭代收敛或符合停止条件，输出聚类结果；否则，令t = t+1，返回步骤2

16 聚类算法动态演示



17 聚 类 过 程



初 始 化

(a) 原始数据集

(b) 随机初始化聚类中心

样 本 聚 类

(c) 样本进行类别分配

(d) 重新计算聚类中心

(e) 重新进行样类别分配

聚 类 模 型

(f) 两次迭代，最终聚类结束，形成两个类

18 KMEANS优缺点

优点：

1. 原理比较简单，实现也是很容易，收敛速度快。
2. 当结果簇是密集的，而簇与簇之间区别明显时，它的效果较好。
3. 主要需要调参的参数仅仅是簇数k。

缺点：

1. K值需要预先给定，很多情况下K值的估计是非常困难的。
2. K-Means算法对初始选取的质心点是敏感的，不同的随机种子点得到的聚类结果完全不同，对结果影响很大。
3. 对噪音和异常点比较的敏感。用来检测异常值。
4. 采用迭代方法，**可能只能得到局部的最优解，而无法得到全局的最优解**。