第二十六节：必备模型及其应用（三）

主讲老师 侯梓熙

聚类分析

本节课程内容

聚类分析

1.什么是聚类分析

2.聚类分析应用场景

3.相似程度的度量

4.K-Means cluster

什么是聚类分析?

根据对象之间的相似程度把对象分成不同的类别，这些类不是事先给定的，而是直接根据数据的特征确定的。其原理是根据数据自身的距离或者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。

聚类分析应用场景

01 城市：一线城市，新一线城市，二线城市…

02 用户会员：钻石会员，铂金会员，黄金会员，白银会员…

03 识别客户购买模式：一大早来买蔬菜和鲜瘦肉，习惯周末时一次性大采购等

04 探测，发现离群点，异常值：如某B2C电商平台上，比较昂贵，频繁的交易，就有可能隐含欺诈的风险尘封，需要风控部门提前关注，监控。

05 不同产品的组合：企业可以按照不同的商业目的，并依照特定的指标标量来为众多的产品种类进行聚类分析，把企业的产品体系进一步细分成具有不同价值，不同目的的多维度的产品组合，并且在此基础分别制定和相应的开发计划，运营计划和服务规划（如哪些产品畅销毛利率又高，哪些产品滞销且毛利又低）。

聚类分析类型

R型聚类

根据样本对变量进行分类

Q型聚类

根据变量对样本进行分类

聚类分析

R型聚类分析的主要作用：

1、不但可以了解个别变量之间的关系的亲疏程度，而且可以了解各个变量组合之间的亲疏程度。
2、根据变量的分类结果以及它们之间的关系，可以选择主要变量进行回归分析或Q型聚类分析。

Q型聚类分析的主要作用：

1、可以综合利用多个变量的信息对样本进行分类；
2、分类结果是直观的，聚类谱系图非常清楚地表现其数值分类结果；
3、聚类分析所得到的结果比传统分类方法更细致、全面、合理。

相似性的度量

01 样本相似性的度量

02 变量相似性的度量

样本相似性的度量

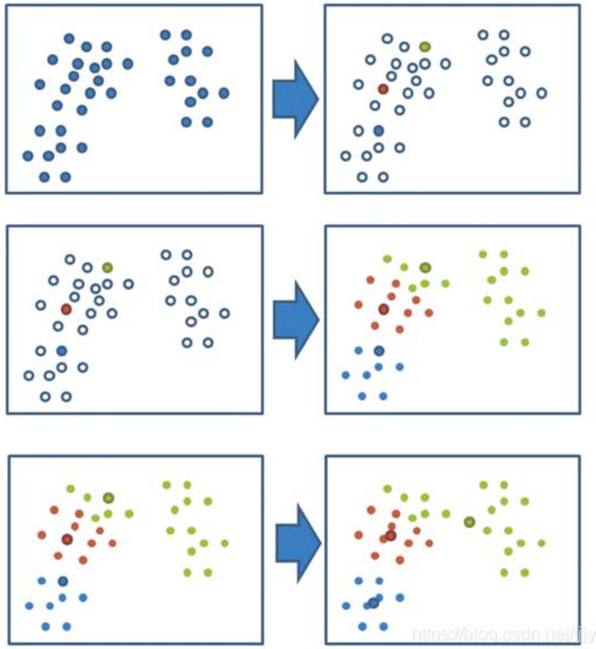
名称	公式
Euclidean distance	$\sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Squared Euclidean distance	$\sum_{i=1}^{p}(x_i - y_i)^2$
Block distance	$\sum_{i=1}^{p}\|x_i - y_i\|$
Chebychev distance	$max\|x_i - y_i\|$
Minkovski distance	$\sqrt[q]{\sum_{i=1}^{p}(x_i - y_i)^2}$

变量相似性的度量

名称	公式
夹角余弦	$cos\theta_{xy} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$
Pearson 相关系数	$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

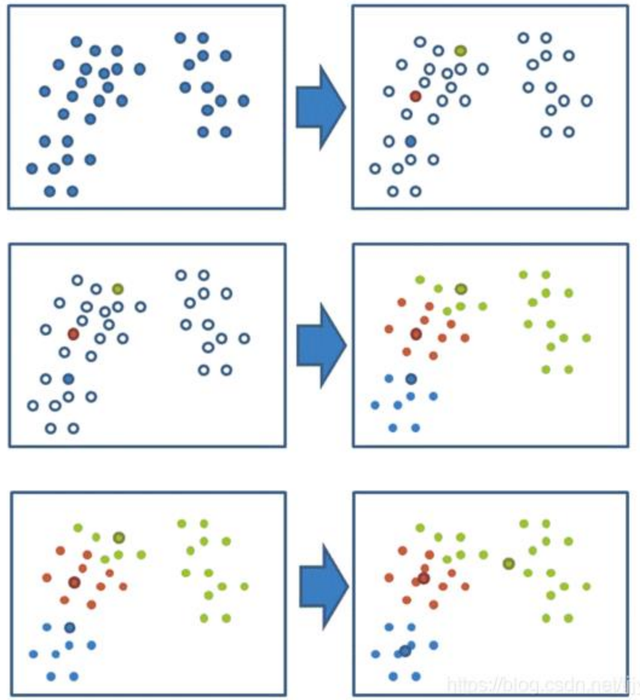
K-Means cluster
（K-均值聚类）

是要求研究者先指定需要划分的类别个数，然后确定各聚类中心，在计算出各样本到聚类中心的距离，最后按距离的远近进行分类。



K-Means cluster
（K-均值聚类）

1. 确定要分的类别数目 K。
2. 确定 K 个类别的初始聚类中心。
3. 根据确定的 K 个初始聚类中心，依次计算每个样本到 K 个聚类中心的欧式距离，并根据距离最近的原则将所有的样本分到事先确定的 K 个类别中。
4. 根据所分成的 K 个类别，计算出各类别中每个变量的均值，并以均值点作为新的 K 个类别中心。根据新的中心位置，重新计算每个样本到新中心的距离，并重新进行分类。
5. 重复第四步，直到满足终止聚类的条件为止。



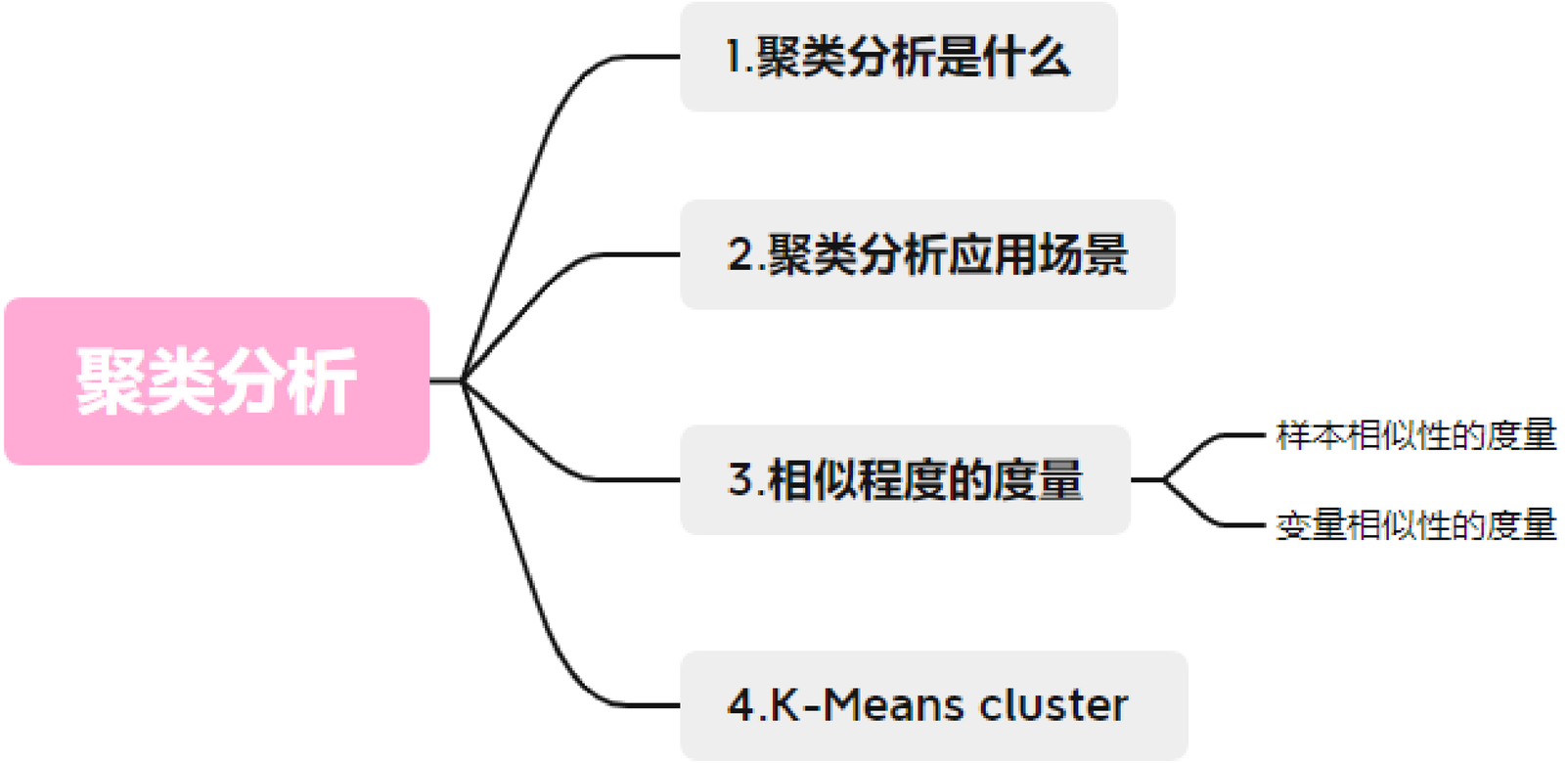
终止聚类的条件包括：

A. 最后一次聚类结果与上一次的聚类结果相同；

B. 迭代次数达到研究者事先指定的最大迭代次数；

C. 新确定的聚类中心点与上一次迭代形成的中心点的最大偏移量小于指定量。

本节课程回顾


聚类分析
1.聚类分析是什么
2.聚类分析应用场景
3.相似程度的度量
样本相似性的度量
变量相似性的度量
4.K-Means cluster

THANK YOU

感谢观看

主讲老师 侯梓熙