社会统计学及SPSS软件应用

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年11月5日

CONTENTS

1 聚类分析概述
- 无监督学习
- 聚类的基本概念

2 迭代聚类法

CONTENTS

1 聚类分析概述

2 迭代聚类法

- 迭代聚类法概述
- k-均值算法
- Stata软件操作

1. 人工智能的核心是机器学习，其首要任务是对事物进行辨识和区分。
2. Herbert Simon：如果一个系统能够通过执行某个过程改进它的性能，这就是学习。
3. 机器学习（statistical machine learning）分为监督学习和无监督学习两大类。

1 监督学习（supervised learning）的主要任务是分类，即用大量已标记数据完成对新数据的区分。

(1) 分类（classification）
(2) 标注（tagging）
(3) 回归（regression）

2 无监督学习（unsupervised learning）的主要任务是聚类，即在没有任何人工干预的情况下对数据进行区分。

1 人和动物的学习在很大程度上是无监督的；我们通过观察发现世界的结构，而不是对每个事物命名。

Lecun, Yann, Y. Bengio, and G. Hinton. "Deep learning." Nature 521.7553(2015):436.

https://www.nature.com/articles/nature14539

2 无监督学习包括：

(1) 聚类分析（Clustering）
(2) 主成分分析（Principle Components Analysis）
(3) 因子分析（Factor Analysis）
(4) 典型相关分析（Canonical Correlation Analysis）

1 方以类聚，物以群分

2 自然的事物总是按一定的规律组织起来的，人们通过认识这些组织的结构特征获得知识，从而做出决策。

3 以生物为例，人们根据生物的相似程度（包括形态结构和生理功能等）

(1) 把生物划分为种和属等不同的等级
(2) 并对每一类群的形态结构和生理功能等特征进行科学的描述
(3) 以弄清不同类群之间的亲缘关系和进化关系

聚类分析的功能

1. 发现数据的潜在结构：深入洞察数据、产生假设、检测异常、确定主要特征。
2. 对数据进行自然分组：确定不同组织之间的相似程度（系统关系）。
3. 对数据进行压缩：将聚类原型作为组织和概括数据的方法。

1. 聚类的概念最早出现在1954年的一篇处理人类学数据的论文中。
2. 最流行并且最简单的算法是1955年发表的k-均值。

聚类分析：直观定义

聚类分析（clustering analysis）的直观定义：将研究对象（例如上述的拼图、顾客、金融产品）根据一些 **特征指标**（如拼图的颜色、顾客的消费习惯和人口特征、金融产品的收益和波动等）的信息，把比较 **相似** 的研究对象，按 **一定的方式** 归为同类。

- **要点1**：根据不同的特征指标聚出的类是不同的
- **要点2**：定义什么叫做“相似的研究对象”
- **要点3**：如何归类

1 聚类：把一个数据对象的集合划分成簇（子集），使簇内对象彼此相似，簇间对象不相似的过程。

2 聚类的核心概念是相似度（similarity）或距离（distance）。

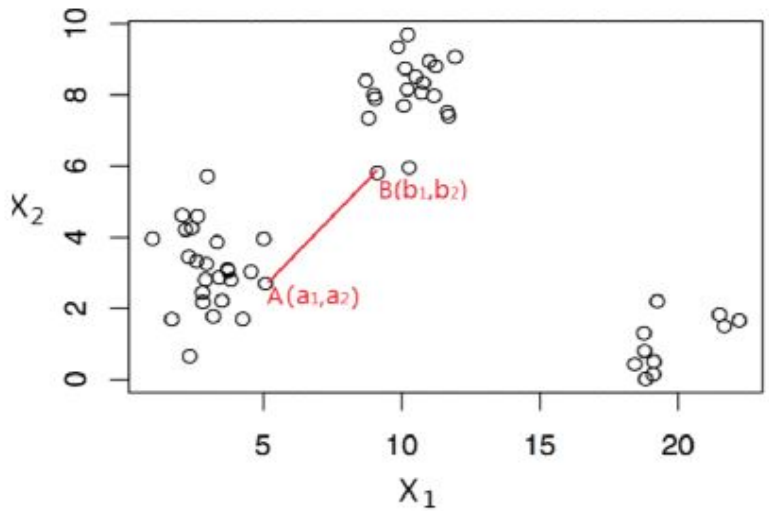
定义“相似度”/距离

单一指标：值越接近的研究对象越相似

两个指标X_1和X_2：把它们作为坐标轴，将所有的研究对象在图中画出来，每一个点对应一个个体。欧式距离越小，即图中点A和点B之间红色线段的长度越短，相似度越高

多个指标：欧氏距离仍适用。还有其他很多种距离的定义方式，比如马氏距离、闵式距离等。

思考：如果指标的取值范围相差很大（比如员工人数和企业的净利润），会对欧氏距离产生怎样的影响，有什么解决办法么？



相异度——数值型数据

明考夫斯基距离（Minkowski distance）

$$d_{ij}=(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^{q})^{1/q},q>0$$

明考夫斯基距离简称明氏距离，依据q的不同取值可以分成

- 绝对距离（$q=1$，Absolute distance）
- 欧氏距离（$q=2$，Euclidean distance）
- 切比雪夫距离（$q=\infty$，Chebyshev distance）

1 相似度（similarity）或距离（distance）
   (1) Minkowski distance
   (2) Mahalanobis distance
   (3) cosine similarity
   (4) correlation coefficient
2 类或簇（cluster）
3 类与类之间的距离，也称为连接（linkage）
   (1) 最短距离或单连接（single linkage）
   (2) 最长距离或完全连接（complete linkage）
   (3) 中心距离
   (4) 平均距离

聚类分析经典算法

1 基于划分的算法 （迭代聚类法）

2 层次聚类法

(1) 聚合聚类（agglomerative clustering）

(2) 分裂聚类（divisive clustering）

1. 基本思想：对一个包含了n个样本的原始数据集，采用某种方法将其划为k个划分，其中的每个划分均表示一堆。

2. 划分的过程：首先创建一个初始的划分（可以是随机的）；之后，通过迭代的方式反复将样本重新分配到更合适的堆中，从而改善划分的整体质量，直到满足划分精度的要求。

3. 评价标准：处于同一划分中的样本之间的差异尽可能小，处于不同划分中的样本之间差异尽可能大。

1 输入：所有数据点A，聚类个数k
2 输出：k个划分

(1) 随机选取k个初始的中心点
(2) repeat
(3) 计算每个点与各中心点之间的距离，将点分派到其距离最近的中心点所属的堆中。
(4) 按照不同的准则进行计算，更新堆的中心点
(5) until 中心点不发生变化

1 在众多基于划分的聚类算法中，使用最为频繁、应用最为广泛的当数k-均值算法。k-均值算法最大特点是它的简洁高效。

2 k-均值算法分别由Steinhaus于1955年、Lyoyd于1957年、Ball和Hall 于1965年、McQueen于1967年提出。

K-均值算法的迭代过程

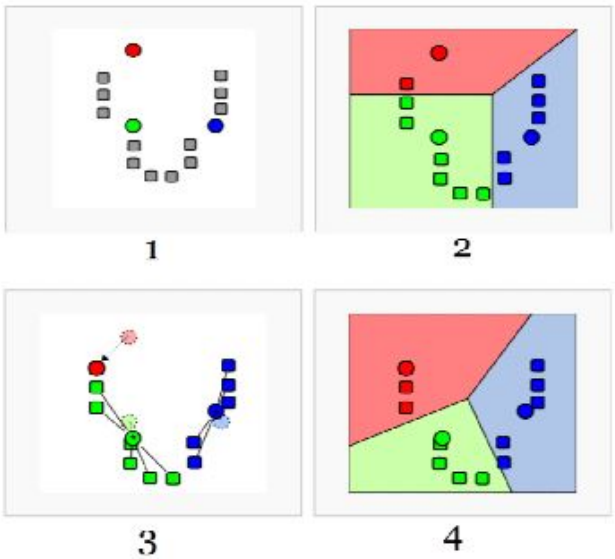
1. 第一步是分配，在分配过程中，每个数据样本都要被分配到与它距离最近的类中心点所属的类中；
2. 第二步是更新，在更新过程中，利用分配到这一类别中的所有数据样本，对类中心点进行更新。

K-均值算法的迭代过程

1. 自行设置聚类个数k，从数据集中选择k个初始的中心点；
2. 将每一个数据对象**指派**到最近的中心点所属的类中，从而构成k堆；
3. 在所有对象都指派完成后，根据每一堆中的数据对象计算平均值作为该堆新的中心点的值；
4. 重复进行指派和更新两个步骤，直至堆中样本的分布不再发生变化。

K均值聚类

K均值聚类：先确定类别数K。确定之后，选取K个“种子”（下图1中红绿蓝三个圆形点），然后看每个个体离哪个种子最近就归到哪一类（下图2）。归类之后原来的种子就被每一个新类的“中心”代替（下图3新的三个圆形点）。再重复上述的归类步骤。直到每个个体所属的类别不再变动为止（下图4）。最终的种子（即最终聚类的中心点）可以用来刻画这一类的特征。



1. 在聚类前，通常需要对变量进行标准化。
2. 由于变量都是以不可比的单位进行测量的，具有极为不同的方差。
3. 方差大的变量比方差小的变量对距离或相似度的影响更大。
4. 对数据进行标准化，可以避免使结果受到具有很大方差变量的影响。

1 变量的标准化

(1) egen newx1=std(x1)
(2) egen newx2=std(x2)
(3) ...

2 迭代聚类分析，事先指定聚类数k

(1) cluster kmeans newx1 newx2 newx3 newx4 newx5, k(3)
(2) 或 cluster kmedians newx1 newx2 newx3 newx4 newx5, k(3)

3 对聚类结果进行排序

(1) sort _clus_1

1 欧式距离（软件默认）

(1) cluster kmeans newx1 newx2 newx3 newx4 newx5, k(3) measure(L2)

2 欧式距离的平方

(1) cluster kmeans newx1 newx2 newx3 newx4 newx5, k(3) measure(L2squared)

3 绝对值距离

(1) cluster kmeans newx1 newx2 newx3 newx4 newx5, k(3) measure(L1)

4 最大值距离

(1) cluster kmeans newx1 newx2 newx3 newx4 newx5, k(3) measure(Linfinity)

5 相关系数相似性度量

(1) cluster kmeans newx1 newx2 newx3 newx4 newx5, k(3) measure(correlation)

- 给定含有5个样本的集合

$$X = \begin{bmatrix} 0 & 0 & 1 & 5 & 5 \\ 2 & 0 & 0 & 0 & 2 \end{bmatrix}$$

- 试用k均值聚类算法将样本聚到2个类中。

（1）选择两个样本点作为类的中心。假设选择 $m_1^{(0)} = x_1 = (0,2)^{\mathrm{T}}$，$m_2^{(0)} = x_2 = (0,0)^{\mathrm{T}}$。

（2）以 $m_1^{(0)}$，$m_2^{(0)}$ 为类 $G_1^{(0)}$，$G_2^{(0)}$ 的中心，计算 $x_3 = (1,0)^{\mathrm{T}}$，$x_4 = (5,0)^{\mathrm{T}}$，$x_5 = (5,2)^{\mathrm{T}}$ 与 $m_1^{(0)} = (0,2)^{\mathrm{T}}$，$m_2^{(0)} = (0,0)^{\mathrm{T}}$ 的欧氏距离平方。

对 $x_3 = (1,0)^{\mathrm{T}}$，$d(x_3, m_1^{(0)}) = 5$，$d(x_3, m_2^{(0)}) = 1$，将 x_3 分到类 $G_2^{(0)}$。

对 $x_4 = (5,0)^{\mathrm{T}}$，$d(x_4, m_1^{(0)}) = 29$，$d(x_4, m_2^{(0)}) = 25$，将 x_4 分到类 $G_2^{(0)}$。

对 $x_5 = (5,2)^{\mathrm{T}}$，$d(x_5, m_1^{(0)}) = 25$，$d(x_5, m_2^{(0)}) = 29$，将 x_5 分到类 $G_1^{(0)}$。

(3) 得到新的类 $G_1^{(1)} = \{x_1, x_5\}$, $G_2^{(1)} = \{x_2, x_3, x_4\}$, 计算类的中心 $m_1^{(1)}$, $m_2^{(1)}$:

$$m_1^{(1)} = (2.5, 2.0)^{\mathrm{T}}, \quad m_2^{(1)} = (2, 0)^{\mathrm{T}}$$

(4) 重复步骤 (2) 和步骤 (3)。

将 x_1 分到类 $G_1^{(1)}$, 将 x_2 分到类 $G_2^{(1)}$, x_3 分到类 $G_2^{(1)}$, x_4 分到类 $G_2^{(1)}$, x_5 分到类 $G_1^{(1)}$。

得到新的类 $G_1^{(2)} = \{x_1, x_5\}$, $G_2^{(2)} = \{x_2, x_3, x_4\}$。

由于得到的新的类没有改变, 聚类停止。得到聚类结果:

$$G_1^* = \{x_1, x_5\}, \quad G_2^* = \{x_2, x_3, x_4\}$$

参考文献

1 李航，2019，《统计学习方法》（第2版），北京：清华大学出版社。
2 潘蕊等，2018，《数据思维实践：从零经验到数据英才》，北京：北京大学出版社。
3 张宪超，2018，《数据聚类》，北京：科学出版社。