分类方法

授课教师:　赵春晖
联系方式:
Email: chhzhao@zju.edu.cn
Phone: 13588312064
Room: 工控新楼308室

内容

1 引言

2 聚类分析

3 判别分析

4 分类实例应用

第一章 引言

什么是聚类分析？

- 物以类聚，人以群分
- 出自《战国策·齐策三》《周易·系辞上》。
- 比喻同类的东西常聚在一起，志同道合的人相聚成群，反之就分开。

第

什

同类的东西 成群，反之
就分开。

第一章 引言

什么是判别分析？

买水果：挑挑拣拣

第一章 引言

什么是判别分析？

Bayes判别（模糊思维）

办公室新来了一个雇员小王，小王是好人还是坏人大家都在猜测。按人们主观意识，一个人是好人或坏人的概率均为0.5。坏人总是要做坏事，好人总是做好事，偶尔也会做一件坏事，一般好人做好事的概率为0.9，坏人做好事的概率为0.2，一天，小王做了一件好事，**小王是好人的概率有多大，你现在把小王判为何种人**。

第一章 引言

两种分类问题

一种是对当前所研究的问题已知它的类别数目, 且知道各类的特征(如分布规律, 或知道来自各类的训练样本), 我们的目的是要将另一些未知类别的个体正确归属于其中某一类, 这是判别分析所要解决的问题.

另一种是事先不知道研究的问题应分为几类, 更不知道观测到的个体的具体分类情况, 我们的目的正是需要通过对观测数据所进行的分析处理，选定一种度量个体接近程度的量, 确定分类数目, 建立一种分类方法, 并按亲近程度对观测对象给出合理的分类. 这种问题在实际中大量存在, 它正是聚类分析所要解决的问题.

第一章 引言

聚类分析和判别分析的区别与联系

- 都是研究分类的
- 在进行聚类分析前，对总体到底有几种类型不知道（研究分几类较为合适需从计算中加以调整）。
- 判别分析则是在总体类型划分已知，在各总体分布或来自总体训练样本基础上，对当前新样本判断它们属于哪个总体。
- 如我们对研究的多元数据的特征不熟悉，当然要先进行聚类分析，才能考虑判别分析问题。

第二章 聚类分析

§ 2.1 聚类分析的思想

§ 2.2 相似性度量

§ 2.3 类间距离度量

§ 2.4 K-均值聚类

第二章 聚类分析

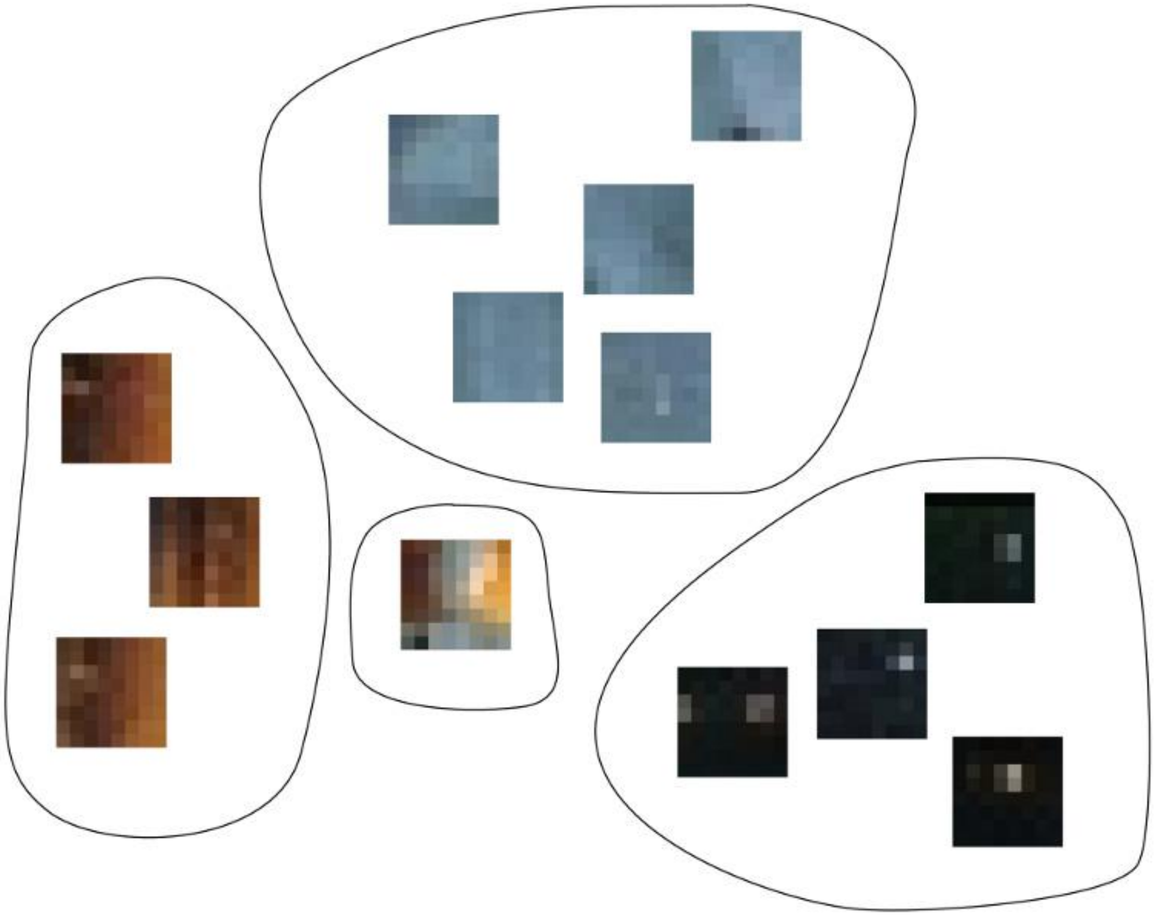
聚类思想

- 我们去参观一个画展，我们完全对艺术一无所知，但是欣赏完多幅作品之后，我们也能把它们分成不同的派别，比如哪些更朦胧一点，哪些更写实一些，即使我们不知道什么叫做朦胧派，什么叫做写实派，但是至少我们能把他们分为两个类。

- 无监督学习（也有人叫非监督学习）与监督学习的不同之处，在于我们事先没有任何训练样本，而需要直接对数据进行建模。这听起来似乎有点不可思议，但是在我们自身认识世界的过程中很多处都用到了无监督学习。比如无监督学习里典型的例子就是聚类了。

一个简单的聚类例子

- 这是按照颜色进行一维聚类。
- 实践中，维度经常多于一个。



第二章 聚类分析

聚类分析的思想

- 对样品的分类常称为Q型聚类分析，对变量的分类常称为R型聚类分析。
- 与多元分析的其他方法相比，聚类分析的方法是很粗糙的，理论上还不完善，但由于它能解决许多实际问题，很受人们的重视（待见）。

聚类分析的思想

- 【例1】若我们需要将下列11户城镇居民按户主个人的收入进行分类，对每户作了如下的统计，结果列于表3.1。
- 在表中，“标准工资收入”、“职工奖金”、“职工津贴”、“性别”、“就业身份”等称为指标，每户称为样品。

第二章 聚类分析

聚类分析的思想

表 3.1 某市 2001 年城镇居民户主个人收入数据

X1	职工标准工资收入	X5	单位得到的其他收入
X2	职工奖金收入	X6	其他收入
X3	职工津贴收入	X7	性别
X4	其他工资性收入	X8	就业身份

X1	X2	X3	X4	X5	X6	X7	X8
540.00	0.0	0.0	0.0	0.0	6.00	男	国有
1137.00	125.00	96.00	0.0	109.00	812.00	女	集体
1236.00	300.00	270.00	0.0	102.00	318.00	女	国有
1008.00	0.0	96.00	0.0	86.0	246.00	男	集体
1723.00	419.00	400.00	0.0	122.00	312.00	男	国有
1080.00	569.00	147.00	156.00	210.00	318.00	男	集体
1326.00	0.0	300.00	0.0	148.00	312.00	女	国有
1110.00	110.00	96.00	0.0	80.00	193.00	女	集体
1012.00	88.00	298.00	0.0	79.00	278.00	女	国有
1209.00	102.00	179.00	67.00	198.00	514.00	男	集体
1101.00	215.00	201.00	39.00	146.00	477.00	男	集体

第二章 聚类分析

聚类分析的思想

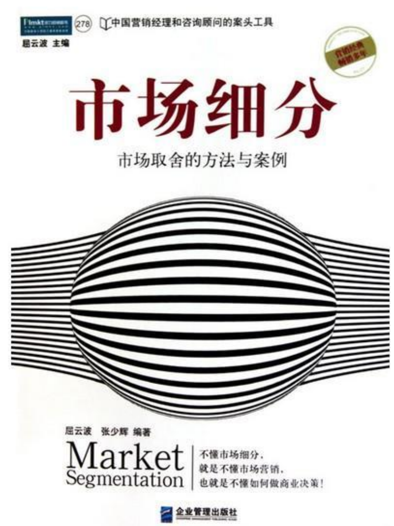
- 例1中的8个指标，前6个是定量的，后2个是定性的。如果分得更细一些，**指标的类型有三种尺度**：
- 间隔尺度：变量用连续的量来表示。 （定量变量）
- 有序尺度：指标用有序的等级来表示，有次序关系，但没有数量表示。
- 名义尺度：指标用一些类来表示，这些类之间没有等级关系也没有数量关系。 （定性变量）

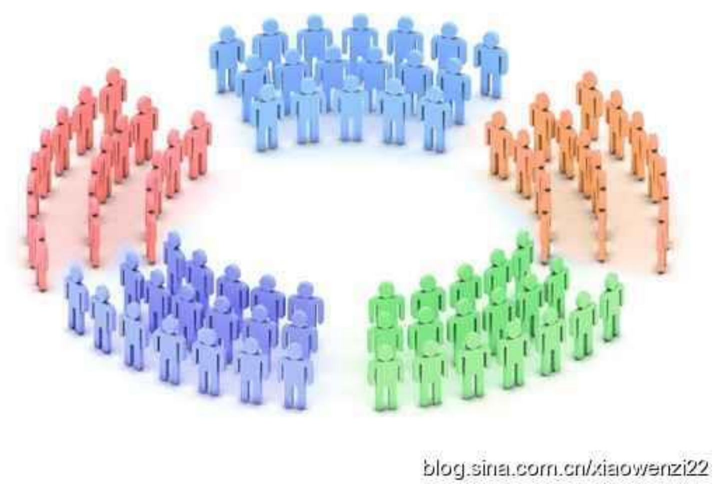
- 不同类型的指标，在聚类分析中，处理的方式是大不一样的。总的来说，提供给间隔尺度的指标的方法较多，对另两种尺度的变量处理的方法不多。

第二章 聚类分析

聚类分析的最典型应用领域

- 客户分群，进而制定差异化的营销方案



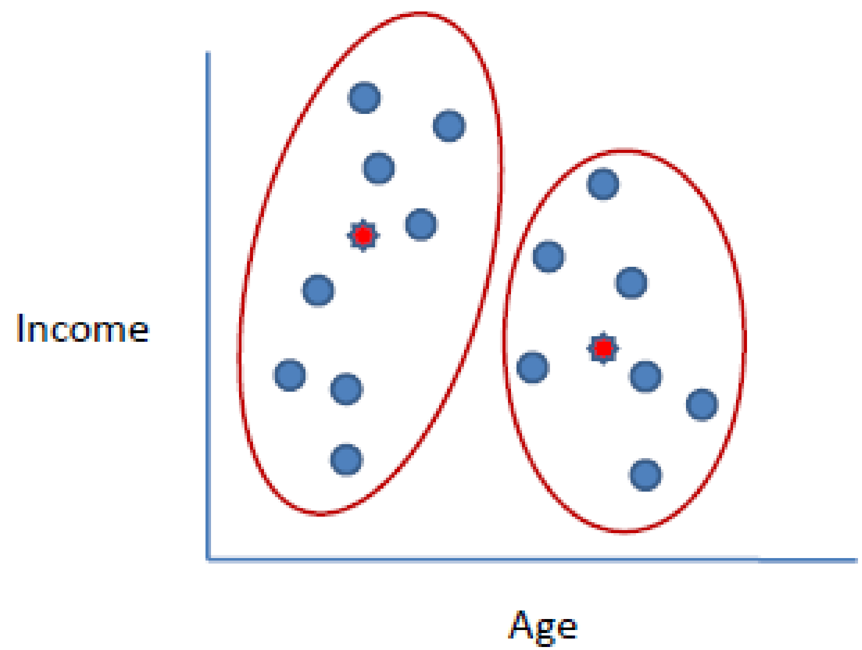


第二章 聚类分析

聚类分析的最典型应用领域

- 客户分群，进而制定差异化的营销方案

例子：如图，按照收入和年龄把客户聚类为两类



第二章 聚类分析

聚类分析在携程的应用举例

- 比如在携程做用户细分时，对用户做春季度假游促销

- 选取一段时间内的：用户ID，星级，用户最后一次消费行为，度假游金额，出发时间，返回时间，目的地，出行人数，入住酒店星级，房间类型，间夜数，是否是商旅客户，是否拒绝邮件，去年同期是否有春季游的相关度假项目，是否是催眠唤醒客户……

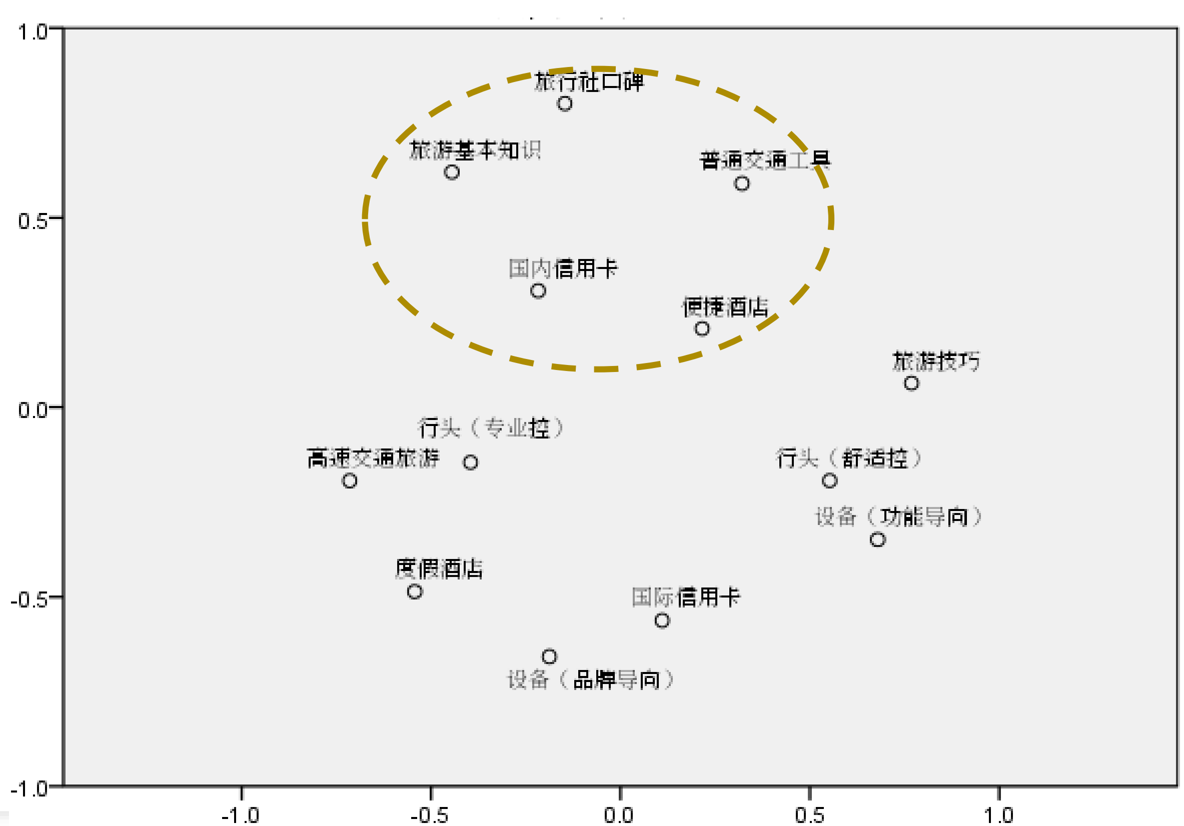
聚类分析在携程的应用举例

- 首先将字段进行预处理
- 然后通过系统聚类模型做出谱系图
- 依据谱系图和具体业务内容，用k-means模型将用户分成3类
  - 普通个人
  - 商务族
  - 积极旅游族

第二章 聚类分析

聚类分析在携程的应用举例

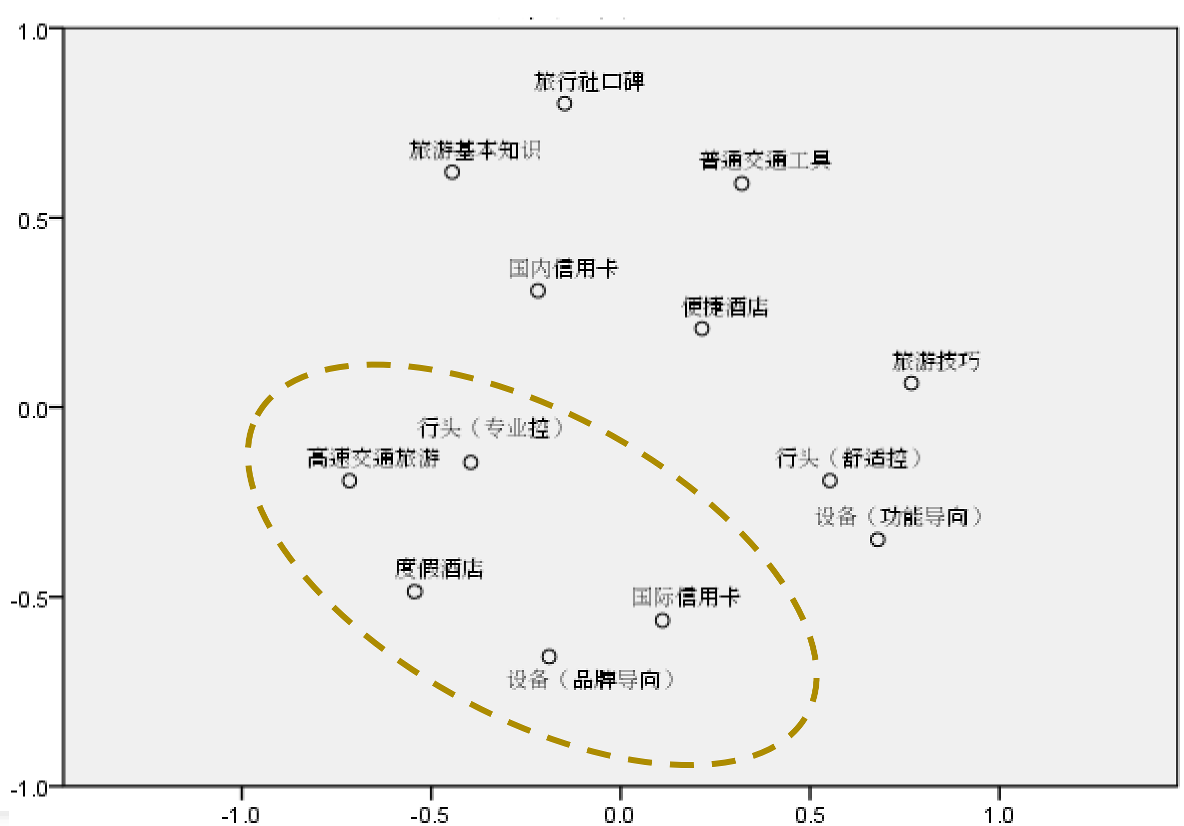
普通个人 关注网站口碑和旅行基本信息，在意旅游安排



第二章 聚类分析

聚类分析在携程的应用举例

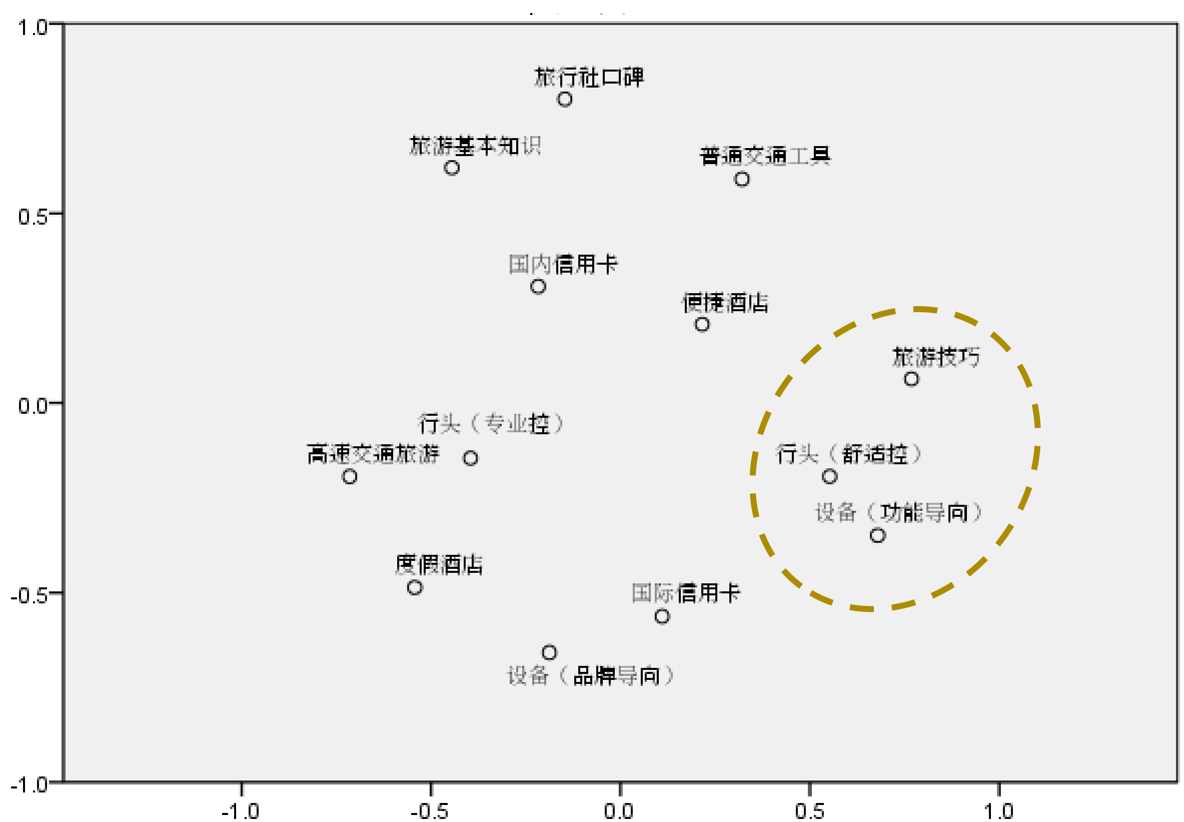
商务族 从出行到住宿皆从容，有固定的购买渠道



第二章 聚类分析

聚类分析在携程的应用举例

积极旅游族 热爱旅游，有经验和技巧，在意出行、住宿性价比



聚类分析在携程的应用举例

- 营销方案：
  - 针对积极旅游族，制作目的地优势显著的度假游，增加新项目，比如：潜水类
  - 针对普通个人：选择更多合作酒店，提供度假抵用券，增加客户粘性
  - 针对商务型：不促销

第二章 聚类分析

聚类分析的思想

- 聚类分析中“类”的特征：
  - 聚类所说的类不是事先给定的，而是根据数据的相似性和距离来划分
  - 聚类的数目和结构都没有事先假定
- 聚类方法的目的是寻找数据中：
  - 潜在的自然分组结构a structure of “natural” grouping
  - 感兴趣的关系relationship

相似性度量

- 从一组复杂数据产生一个相当简单的类结构，必然要求进行“相关性”或“相似性”度量。

- 在相似性度量的选择中，常常包含许多主观上的考虑，但是最重要的考虑是指标（包括离散的、连续的和二态的）性质或观测的尺度（名义的、次序的、间隔的和比率的）以及有关的知识。
- 一般来说，当对样品进行聚类时，“靠近”往往由某种距离来刻画。另一方面，当对指标聚类时，根据相关系数或某种关联性度量来聚类。

表 3.2 数据矩阵

No	$\mathbf{x}_1$	$\mathbf{x}_2$	…	$\mathbf{x}_p$
1	x_{11}	x_{12}	…	x_{1p}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}

相似性度量

- 如何衡量样本点或变量之间的距离或相似程度？
  - 距离
  - 相似系数

常用的距离的计算方法

- 设每个样品有p个指标（变量）。把n个样品看成p维空间中的n个点，则两个样品间相似程度就可用p维空间中的两点距离公式来度量。
- 两点距离公式可以从不同角度进行定义。
- 当变量的测量值相差悬殊时，要先进行标准化，以消除计量单位对计算结果的影响。

第二章 聚类分析

常用的距离的计算方法

- 欧氏距离（Euclidean）

$$\sqrt{\sum (x_{ik} - x_{jk})^2}$$

- 平方欧氏距离Squared Euclidean

$$\sum (x_{ik} - x_{jk})^2$$

- 切比雪夫距离（Chebyshev）

$$\max | x_{ik} - x_{jk} |$$

第二章 聚类分析

常用的距离的计算方法

闵柯夫斯基距离

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

按q的取值不同可以包括多种距离计算方法。例如：

（1）绝对距离（$q=1$）**：** $d_{ij}(1) = \sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|$

（2）欧氏距离（$q=2$）**：** $d_{ij}(2) = \left(\sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|^{2}\right)^{1/2}$

相似系数的计算方法

- 变量间的相似性可以从它们的方向趋同性或“相关性”进行考察， “夹角余弦法”和“相关系数”两种主要度量方法，统称为相似系数。

(1) 夹角余弦

两变量X_i与X_j看作p维空间的两个向量，这两个向量间的夹角余弦可用下式进行计算

$$\cos\theta_{ij}=\frac{\sum_{k=1}^{p}X_{ik}X_{jk}}{\sqrt{(\sum_{k=1}^{p}X_{ik}^{2})(\sum_{k=1}^{p}X_{jk}^{2})}}$$

显然， $|\cos\theta_{ij}|\leq 1$。

相似系数的计算方法

（2）相关系数

相关系数经常用来度量变量间的相似性。变量X_i与X_j的相关系数定义为

$$r_{ij}=\frac{\sum_{k=1}^{p}(X_{ik}-\bar{X}_i)(X_{jk}-\bar{X}_j)}{\sqrt{\sum_{k=1}^{p}(X_{ik}-\bar{X}_i)^2\sum_{k=1}^{p}(X_{jk}-\bar{X}_j)^2}}$$

显然也有，$|r_{ij}|\le 1$。

第二章 聚类分析

聚类分析的思想

通常认为，聚类作为一种无监督式的机器学习方法，它的过程是这样的：

在未知样本类别的情况下，通过计算样本彼此间的距离（欧式距离，汉明距离，余弦距离等）来估计样本所属类别。

从结构性来划分，聚类方法分为自上而下和自下而上两种方法，前者的算法是先把所有样本视为一类，然后不断从这个大类中分离出小类，直到不能再分为止；后者则相反，首先所有样本自成一类，然后不断两两合并，直到最终形成几个大类。

第二章 聚类分析

聚类分析的思想

- 聚类分析给人们提供了丰富多采的方法进行分类，这些方法大致可归纳为：

(1)系统聚类法。(2)模糊聚类法。(3) K-均值法。

(4) 有序样品的聚类。(5) 分解法。(6) 加入法。

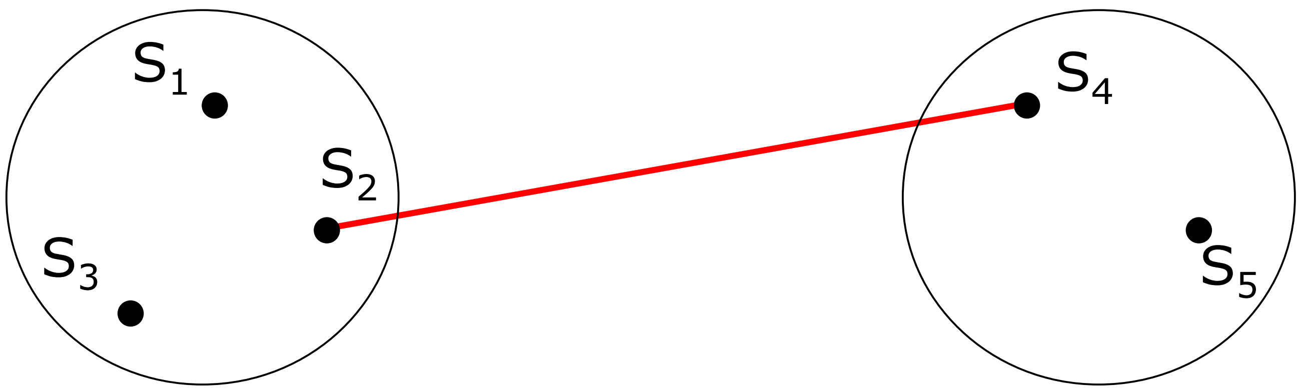
系统聚类法（分层聚类）

- 开始时，有多少样本点就是多少类。
- 第一步先把最近的两类（点）合并成一类;
- 然后再把剩下的最近的两类合并成一类;
- 这样下去，每次都少一类，直到最后只有一大类为止。显然，越是后来合并的类，距离就越远。

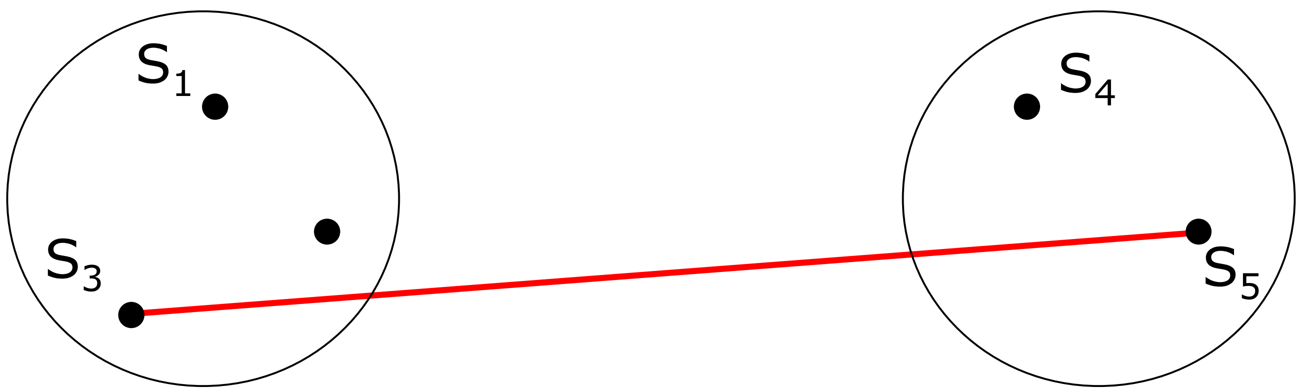
如何计算类与类之间的距离？

- 最短距离法
- 最长距离法
- 重心法
- Ward法（离差平方和法）
- 等等

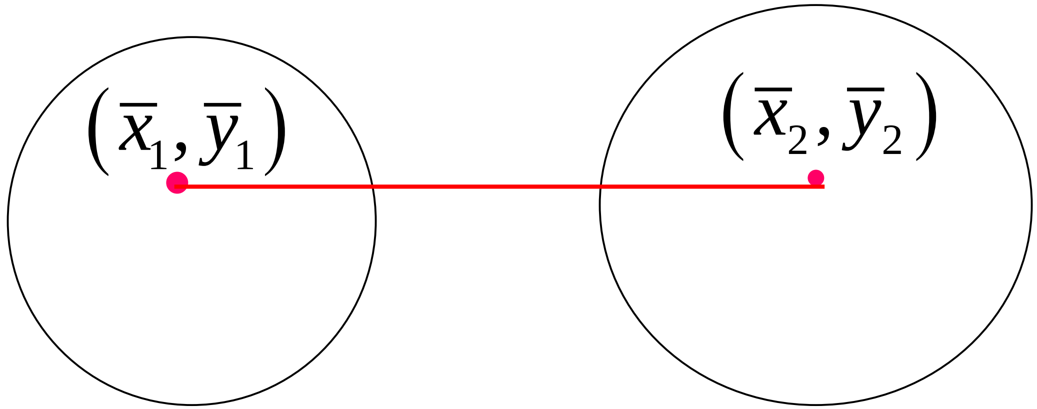
最短距离



最长距离

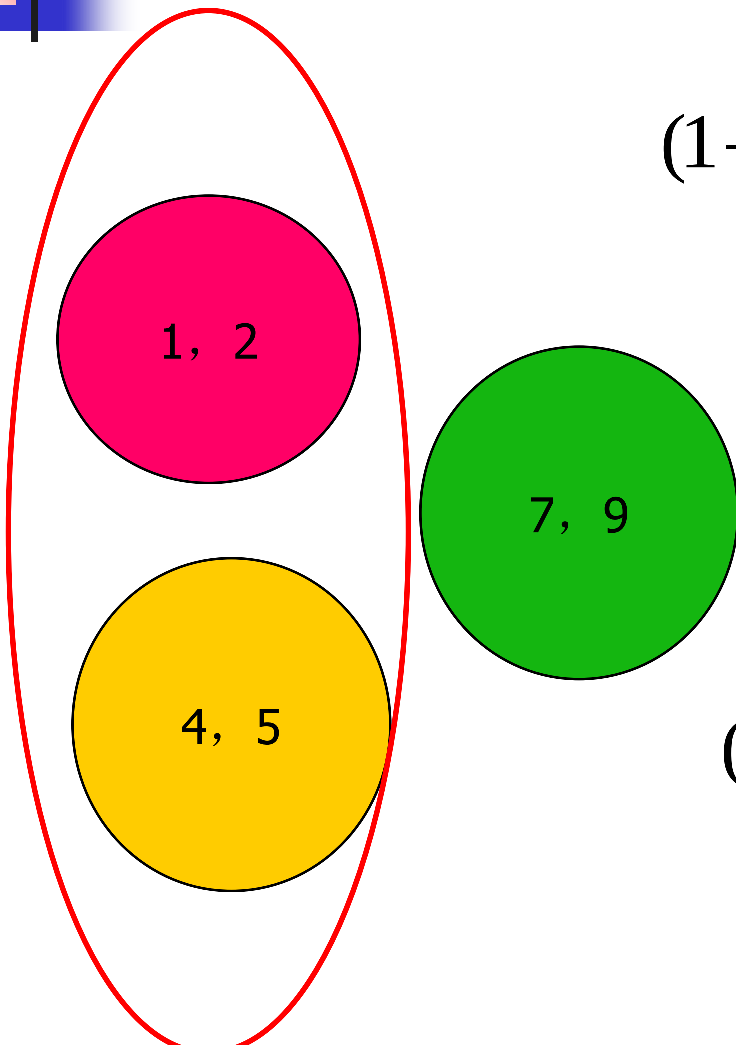


重心法（Centroid clustering):均值点的距离



离差平方和法（Ward法）：

合并离差平方和变动最小的两个类



$$(1-1.5)^2+(2-1.5)^2=0.5$$

$$(7-8)^2+(9-8)^2=2$$

$$(4-4.5)^2+(5-4.5)^2=0.5$$

第二章 聚类分析

例2.哪些少数民族的生存状况更接近?

民族	原始数据	
	标化死亡率(‰)	出生时期望寿命(岁)
满族	5.80	70.59
朝鲜族	7.44	67.14
蒙古族	8.11	65.48
维吾尔族	10.21	58.88
藏族	9.51	59.24
哈萨克族	9.81	60.47

*标化死亡率是根据相同的人口年龄结构（标准组）计算的，因而更具可比性。

第二章 聚类分析

6个不同民族的聚类:

民族	原始数据		标准化数据	
	标化死亡率(‰)	出生时期望寿命(岁)	标化死亡率(‰)	出生时期望寿命(岁)
满族	5.80	70.59	**-1.59**	**1.44**
朝鲜族	7.44	67.14	**-0.62**	**0.73**
蒙古族	8.11	65.48	**-0.22**	**0.38**
维吾尔族	10.21	58.88	**1.03**	**-0.99**
藏族	9.51	59.24	**0.61**	**-0.91**
哈萨克族	9.81	60.47	**0.79**	**-0.66**

第二章 聚类分析

各民族之间的欧氏距离

		满族 G1={S1}	朝鲜族 G2={S2}	蒙古族 G3={S3}	维吾尔族 G4={S4}	藏族 G5={S5}	哈萨克族 G6={S6}
满族	G1={S1}	0					
朝鲜族	G2={S2}	1.208	0				
蒙古族	G3={S3}	1.732	0.526	0			
维吾尔族	G4={S4}	3.570	2.374	1.851	0		
藏族	G5={S5}	3.224	2.048	1.539	0.422	0	
哈萨克族	G6={S6}	3.173	1.973	1.448	0.406	0.311	0

第二章 聚类分析

最短距离法：

- （1）首先合并G5、G6，再计算新类与其他类之间的距离。

		满族 G1={S1}	朝鲜族 G2={S2}	蒙古族 G3={S3}	维吾尔族 G4={S4}	藏族 G5={S5}	哈萨克族 G6={S6}
满族	G1={S1}	0					
朝鲜族	G2={S2}	1.208	0				
蒙古族	G3={S3}	1.732	0.526	0			
维吾尔族	G4={S4}	3.570	2.374	1.851	0		
藏族	G5={S5}	3.224	2.048	1.539	0.422	0	
哈萨克族	G6={S6}	3.173	1.973	1.448	0.406	0.311	0

第二章 聚类分析

最短距离法：

（2）根据计算结果合并G4，G7

	G1={S1}	G2={S2}	G3={S3}	G4={S4}	G7={S5,S6}
G1={S1}	**0**				
G2={S2}	**1.208**	**0**			
G3={S3}	**1.732**	**0.526**	**0**		
G4={S4}	**3.570**	**2.374**	**1.851**	**0**	
G7={S5,S6}	**3.173**	**1.973**	**1.448**	**0.406**	**0**

最短距离法：

（3）根据表中的结果合并G2,G3

	G1={S1}	G2={S2}	G3={S3}	G8={S4,S5,S6}
G1={S1}	0			
G2={S2}	1.208	0		
G3={S3}	1.732	0.526	0	
G8={S4,S5,S6}	3.173	1.973	1.448	0

最短距离法：

（4）根据表中的结果合并G1,G9

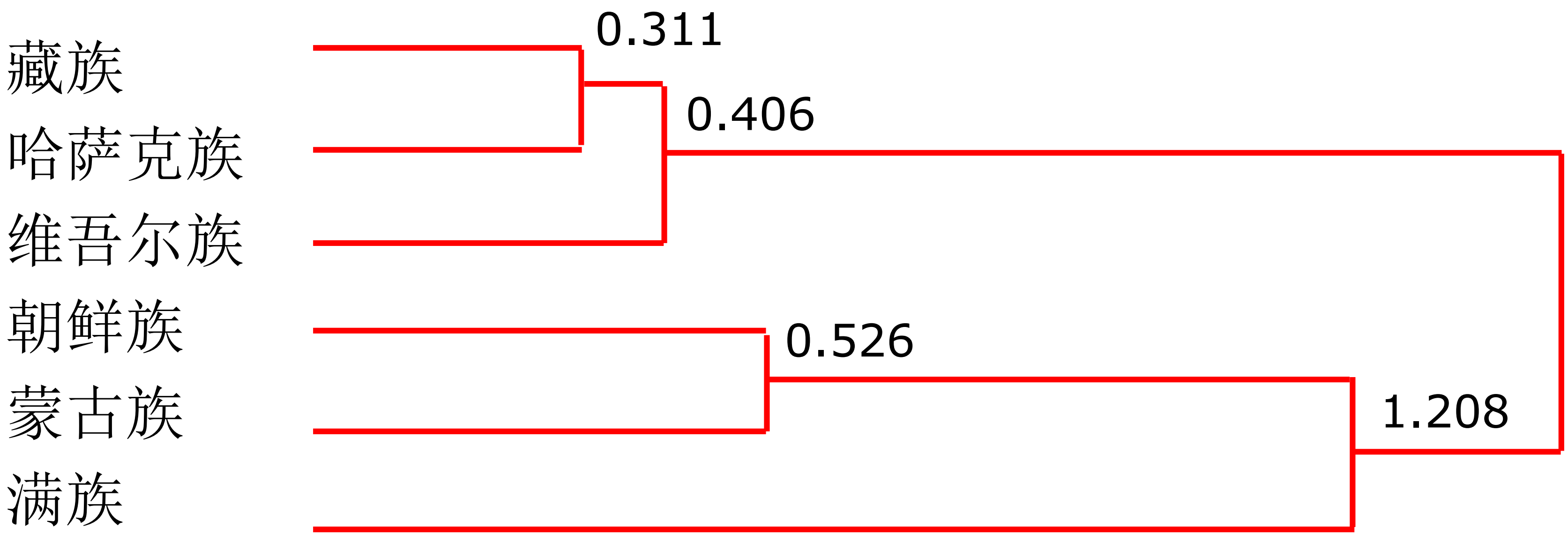
	G1={S1}	G9={S2,S3}	G8={S4,S5,S6}
G1={S1}	0		
G9={S2,S3}	1.208	0	
G8={S4,S5,S6}	3.173	1.448	0

最短距离法：

（5）最后合并成一类

	G10={S1,S2,S3}	G8={ S4,S5,S6}
G10={S1,S2,S3 }	0	
G8={ S4,S5,S6}	1.448	0

聚类结果的谱系聚类图(最短距离法)



K-均值聚类

- 系统聚类法需要计算出不同样品或变量的距离，还要在聚类的每一步都要计算“类间距离”，相应的计算量自然比较大；特别是当样本的容量很大时，需要占据非常大的计算机内存空间，这给应用带来一定的困难。
- k-均值聚类（k-means cluster）可以避免上述问题，适用于样本点很多的情况，但要求你先确定要分多少类。

K-均值聚类

K-均值法，又叫快速聚类法，是Macqueen于1967年提出的，其思想是通过把每个样品聚集到其最近形心（均值）类中去，从而把样品（而不是变量）聚集成K个类的集合。类的个数K可以预先给定，或者在聚类过程中确定。

或者一开始就对元素分组，或者从一个构成各类核心的“种子”集合开始。

选择好的初始构形，将能免除系统的偏差。一种方法是从所有项目中随机地选择“种子”点或者随机地把元素分成若干个初始类。

K-均值聚类

这个聚类过程由下列三步所组成：

- 把样品粗略分成K个初始类；
- 进行修改，逐个分派样品到其最近均值的类中去（通常用标准化数据或非标准化数据计算欧氏距离）。重新计算接受新样品的类和失去样品的类的形心（均值）；
- 重复第2步，直到各类无元素进出。

K-均值聚类

几个影响的关键因素：

最初类中心的选取（包括位置与个数）；
相似度度量指标及阈值；
最多聚类个数；
每类内样本个数限制；

第二章 聚类分析

K-均值聚类的实现很简单

原始数据{x1,x2,...,xn}，这些数据没有被标记的。
初始化k个随机数据u1,u2,...,uk。这些xn和uk都是向量。
根据下面两个公式迭代就能求出最终所有的u,这些u就是最终所有类的中心位置。

$$c^{(i)} = \arg\min_j \| x^{(i)} - u_j \|^2$$

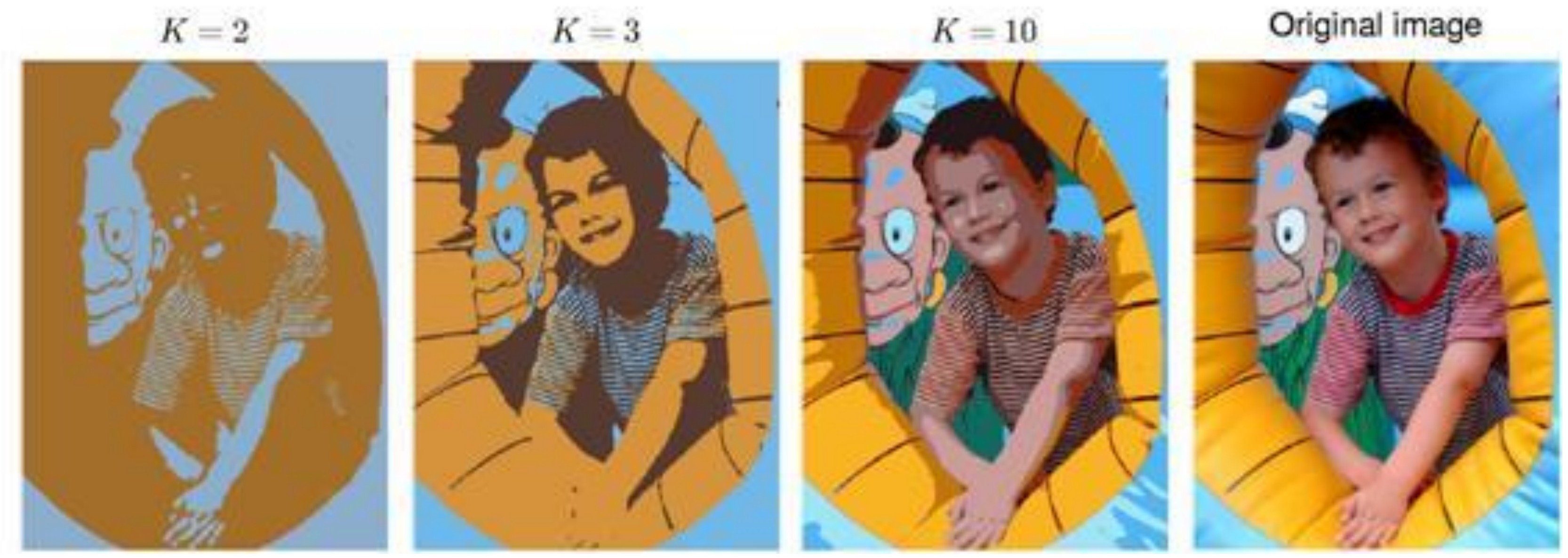
$$u_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

分派样品

重新计算中心

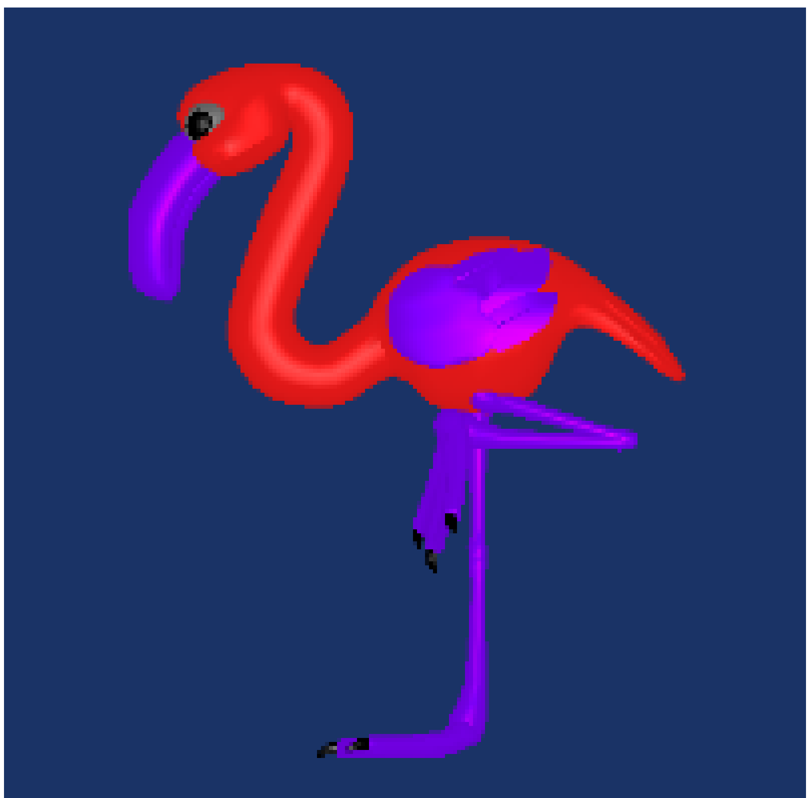
K-均值算法的应用：图像压缩

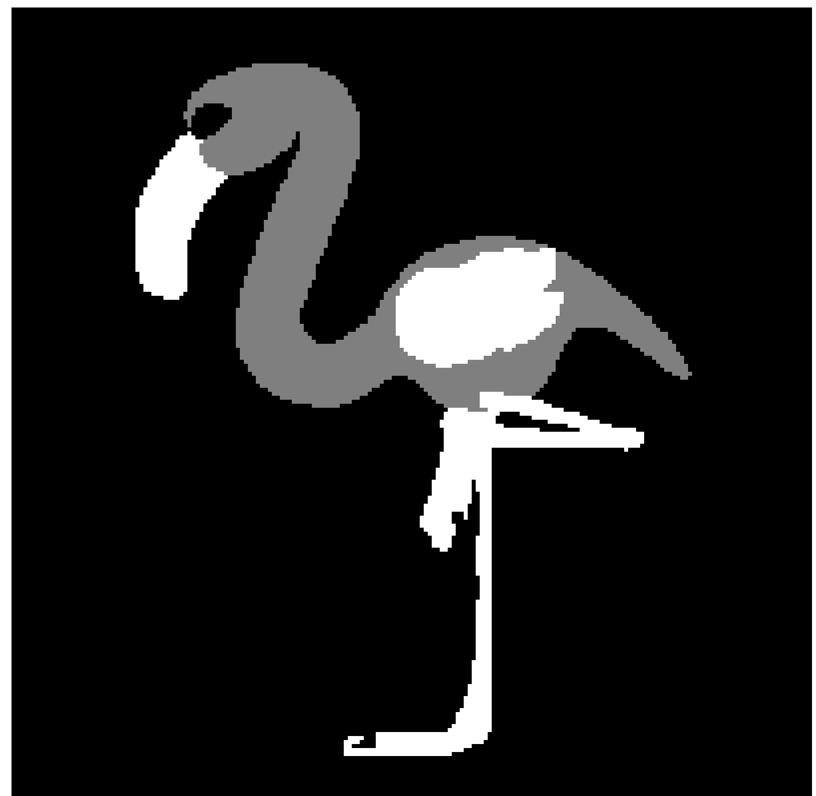
- 群的个数越少，意味着图像被转化成颜色表示数量很少的图像了。



K-均值算法的应用：图像压缩

- 原理和上面人物照片是一致的。





K-均值算法的应用

https://blog.csdn.net/liulingyuan6/article/details/53637812

- 整理了10个天池、DataCastle、DataFountain等中出现的，可使用聚类算法处理的问题场景实例。
- 基于用户位置信息的商业选址

随着信息技术的快速发展，移动设备和移动互联网已经普及到千家万户。在用户使用移动网络时，会自然的留下用户的位置信息。随着近年来GIS地理信息技术的不断完善普及，结合用户位置和GIS地理信息将带来创新应用。如百度与万达进行合作，通过定位用户的位置，结合万达的商户信息，向用户推送位置营销服务，提升商户效益。通过大量移动设备用户的位置信息，为某连锁餐饮机构提供新店选址。

K-均值聚类

存在的问题：

无序问题

在有些实际问题中，要研究的现象与时间的顺序密切相关。

例如我们想要研究，从1949年到2003年以来，国民收入可以划分为几个阶段，阶段的划分必须以年份顺序为依据，总的想法是要将国民收入接近的年份划分到一个段内，要完成类似这样的问题的研究，用k-means的方法显然是不行了。

过程测量数据----时间指标

有序样品的聚类

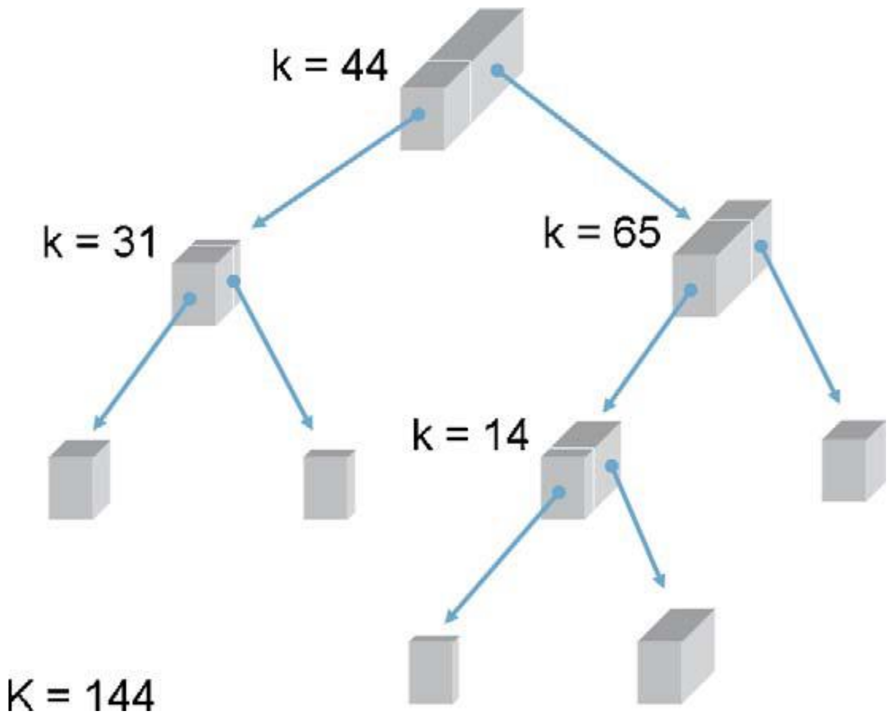
假设用$x_1, x_2, \cdots x_n$表示 n 个有顺序的样品，有序样品的分类结果要求每一类必须呈：$\{x_i, x_{i+1}, \cdots x_{i+j}\}$，$i \geq 1, j \geq 0$, 由于增加了有序这个约束条件，对分类带来哪些影响？

对于这类有序样品的分类，实质上是需要找出一些分点，将它们划分成几个分段，每个分段看作一类，称这种分类为分割。

显然，分点在不同位置可以得到不同的分割。这样就存在一个如何决定分点，使达到所谓最优分割的问题。即要求一个分割能使各段内部样品间的差异最小，而各段之间样品的差异最大。这就是决定分割点的依据。

有序样品的聚类

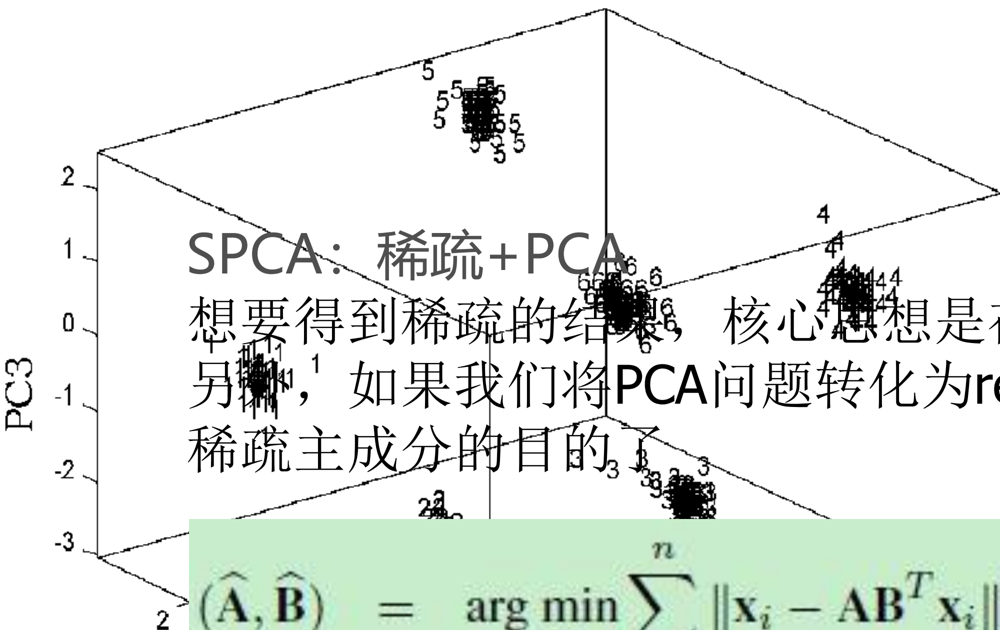
有序聚类



度量指标：
STD
RMSE

知识拓展：降维+聚类

利用特征提取方法实现降维，针对降维后的数据进行聚类



主成分的物理意义？
可解释性欠缺

SPCA：稀疏+PCA

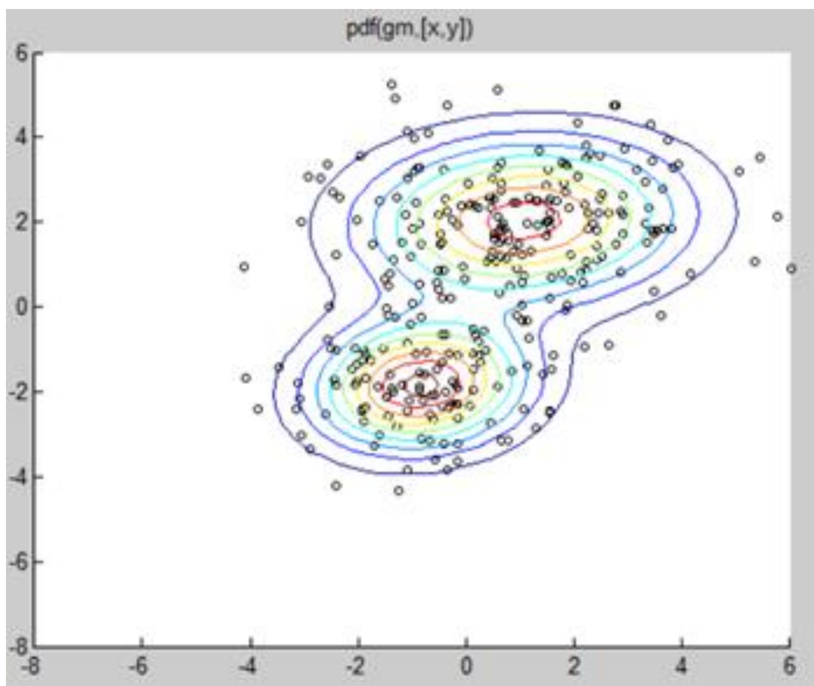
想要得到稀疏的结果，核心思想是在优化参数时加入 L1 penalty.
另外，如果我们将PCA问题转化为regression问题，那么就达到了求解稀疏主成分的目的了

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_j\|_1$$

$$\text{subject to} \quad \mathbf{A}^T\mathbf{A} = I_{k\times k}.$$

知识拓展：高斯混合模型（GMM）聚类

GMM是将若干个概率分布为高斯分布的模型混合在一起的模型。简单地说，k-means 的结果是每个数据点被 assign 到其中某一个 cluster 了，而 GMM 则给出这些数据点被 assign 到每个 cluster 的概率，又称作 soft assignment 。



知识拓展：高斯混合模型（GMM）聚类

高斯模型混合模型理论上可以拟合任意形状的概率分布；将一个事物分解为若干的基于高斯概率密度函数（正态分布曲线）形成的模型。