聚类分析

翟祥

目录

* 聚类分析概述
* 聚类分析数据准备
* 聚类分析技术
* 聚类结果探索
* 聚类结果部署

何为聚类分析

* 聚类分析又称群分析，它是研究对样品或指标进行分类的一种多元统计方法。
* 所谓的“类”，通俗地说就是相似元素的集合.聚类分析是按照观测样品取值的相似程度，对观测样品进行分类，使在同一类内的观测样品是相似的，不同类间的观测是不相似的。
* 什么是分类?它只不过是将一个观测对象指定到某一类(组)。

何为聚类分析

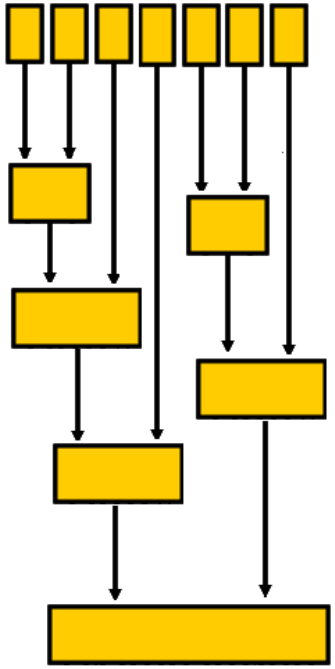
分类的问题可以分成两种：

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).

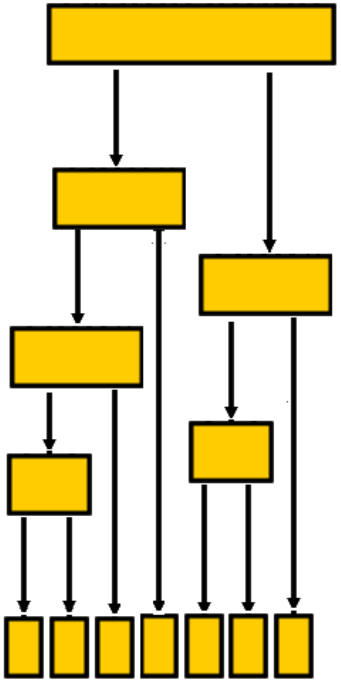
另一种是事先不知道研究的问题应分为几类，更不知道观测到的个体的具体分类情况，我们的目的正是需要通过对观测数据所进行的分析处理，选定一种度量个体接近程度的量，确定分类数目，建立一种分类方法，并按亲近程度对观测对象给出合理的分类. 这种问题在实际中大量存在，它正是聚类分析所要解决的问题.

聚类分析

Agglomerative



Divisive



聚类分析流程

1、聚类分析的数据准备

变量和观测选择

分布分析

量纲剔除

2、聚类分析过程

3、聚类后处理

类数确认

标签确定

4、模型部署

聚类分析

聚类分析根据一批样品的许多观测指标，按照一定的数学公式具体地计算一些样品或一些参数(指标)的相似程度，把相似的样品或指标归为一类，把不相似的归为一类。

例如对上市公司的经营业绩进行分类；据经济信息和市场行情，客观地对不同商品、不同用户及时地进行分类。又例如当我们对企业的经济效益进行评价时，建立了一个由多个指标组成的指标体系，由于信息的重叠，一些指标之间存在很强的相关性，所以需要将相似的指标聚为一类，从而达到简化指标体系的目的。

定义距离的准则

定义距离要求满足第i个和第j个样品之间的距离如下四个条件（距离可以自己定义，只要满足距离的条件）

$d_{ij} \geq 0$对一切的i和j成立;

$d_{ij} = 0$当且仅当$i = j$成立;

$d_{ij} = d_{ji}0$对一切的i和j成立;

$d_{ij} \leq d_{ik} + d_{kj}$对于一切的$i$和$j$成立.

距离矩阵

至此，我们已经可以根据所选择的距离构成样本点间的距离表,样本点之间被连接起来。

G_q \ G_p	G_1	G_2	…	G_n
G_1	0	d_{12}	…	d_{1n}
G_2	d_{21}	0		d_{2n}
⋮	⋮	⋮		⋮
G_n	d_{n1}	d_{n2}	…	0

常见距离

(1) 明氏距离测度

设 $\mathbf{x_i}=(x_{i1},\ x_{i2},\ \cdots,\ x_{ip})'$ 和 $\mathbf{x_j}=(x_{j1},x_{j2},\cdots,x_{jp})'$ 是第i和 j 个样品的观测值，则二者之间的距离为：

明氏距离 $$d_{ij}=\left(\sum_{k=1}^{p}|x_{ik}-x_{jk}|^{g}\right)^{\frac{1}{g}}$$

特别，欧氏距离 $$d_{ij}=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^{2}}$$

常见距离

(2)杰氏距离

这是杰斐瑞和马突斯塔(Jffreys & Matusita)所定义的一种距离，其计算公式为：

$$d_{ij}(J)=\left[\sum_{k=1}^{p}(\sqrt{x_{ik}}-\sqrt{x_{jk}})^2\right]^{1/2}$$

常见距离

(3)兰氏距离

这是兰思和维廉姆斯(Lance & Williams)所给定的一种距离，其计算公式为：

$$d_{ij}(L)=\sum_{k=1}^{p}\frac{\left|x_{ik}-x_{jk}\right|}{x_{ik}+x_{jk}}$$

这是一个自身标准化的量，由于它对大的奇异值不敏感，这样使得它特别适合于高度偏倚的数据。虽然这个距离有助于克服明氏距离的第一个缺点，但它也没有考虑指标之间的相关性。

常见距离

(4)马氏距离

这是印度著名统计学家马哈拉诺比斯(P．C．Mahalanobis)所定义的一种距离，其计算公式为：

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$

分别表示第i个样品和第j样品的p指标观测值所组成的列向量，即样本数据矩阵中第i个和第j个行向量的转置，Σ表示观测变量之间的协方差短阵。在实践应用中，若总体协方差矩阵Σ未知，则可用样本协方差矩阵作为估计代替计算。

常见距离

(5) 斜交空间距离

由于各变量之间往往存在着不同的相关关系，用正交空间的距离来计算样本间的距离易变形，所以可以采用斜交空间距离。

$$d_{ij}=\left[\frac{1}{p^2}\sum_{h=1}^{p}\sum_{k=1}^{p}(x_{ih}-x_{jh})(x_{ik}-x_{jk})\gamma_{hk}\right]^{1/2}$$

当各变量之间不相关时，斜交空间退化为欧氏距离。

常见相似程度

（1）相似系数

设 $\mathbf{x_i}=\left(x_{i1},\ x_{i2},\ \cdots,\ x_{ip}\right)'$ 和 $\mathbf{x_j}=(x_{j1},x_{j2},\cdots,x_{jp})'$ 是第 i 和 j 个样品的观测值，则二者之间的相似测度为:

其中 $$\gamma_{ij}=\frac{\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{[\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)^2][\sum_{k=1}^{p}(x_{jk}-\bar{x}_j)^2]}}$$

常见相似程度

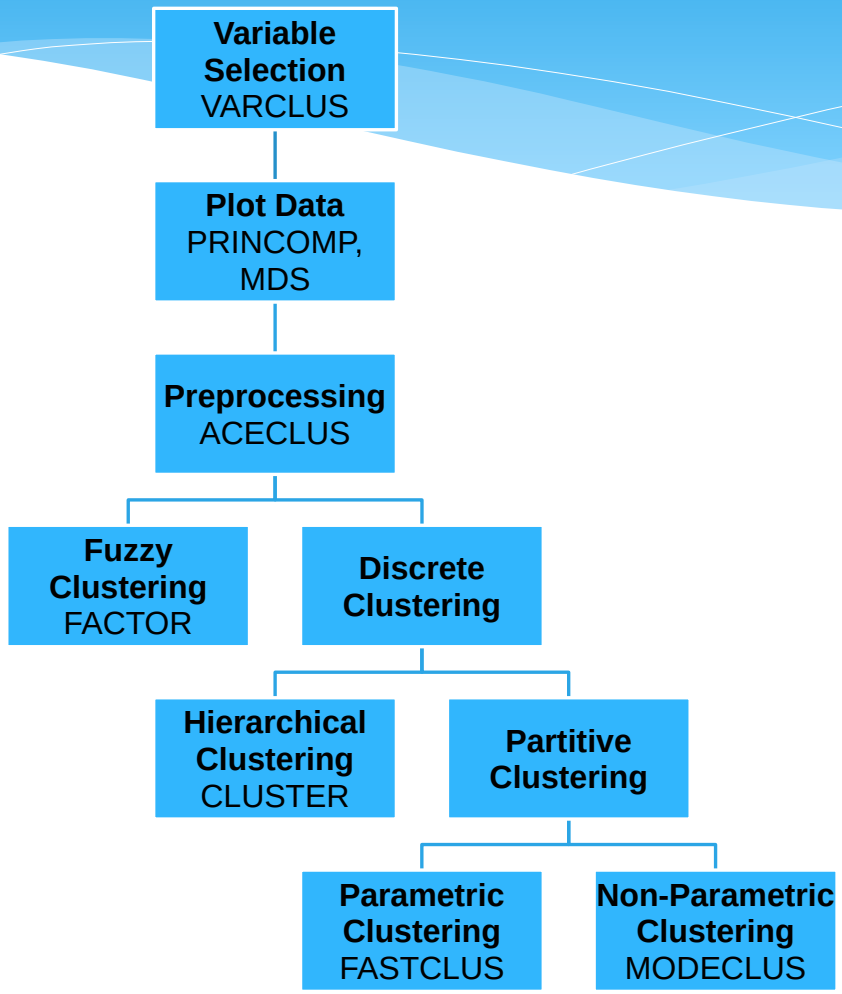
（2）夹角余弦

夹角余弦时从向量集合的角度所定义的一种测度变量之间亲疏程度的相似系数。设在n维空间的向量

$$\mathbf{x}_i=(x_{1i},x_{2i},\cdots,x_{ni})' \qquad \mathbf{x}_j=(x_{1j},x_{2j},\cdots,x_{nj})'$$

$$c_{ij}=\cos\alpha_{ij}=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{n}x_{ki}^2\sum_{k=1}^{n}x_{kj}^2}} \qquad d_{ij}^2=1-C_{ij}^2$$

聚类分析概况


Variable Selection
VARCLUS
Plot Data
PRINCOMP, MDS
Preprocessing
ACECLUS
Fuzzy Clustering
FACTOR
Discrete Clustering
Hierarchical Clustering
CLUSTER
Partitive Clustering
Parametric Clustering
FASTCLUS
Non-Parametric Clustering
MODECLUS

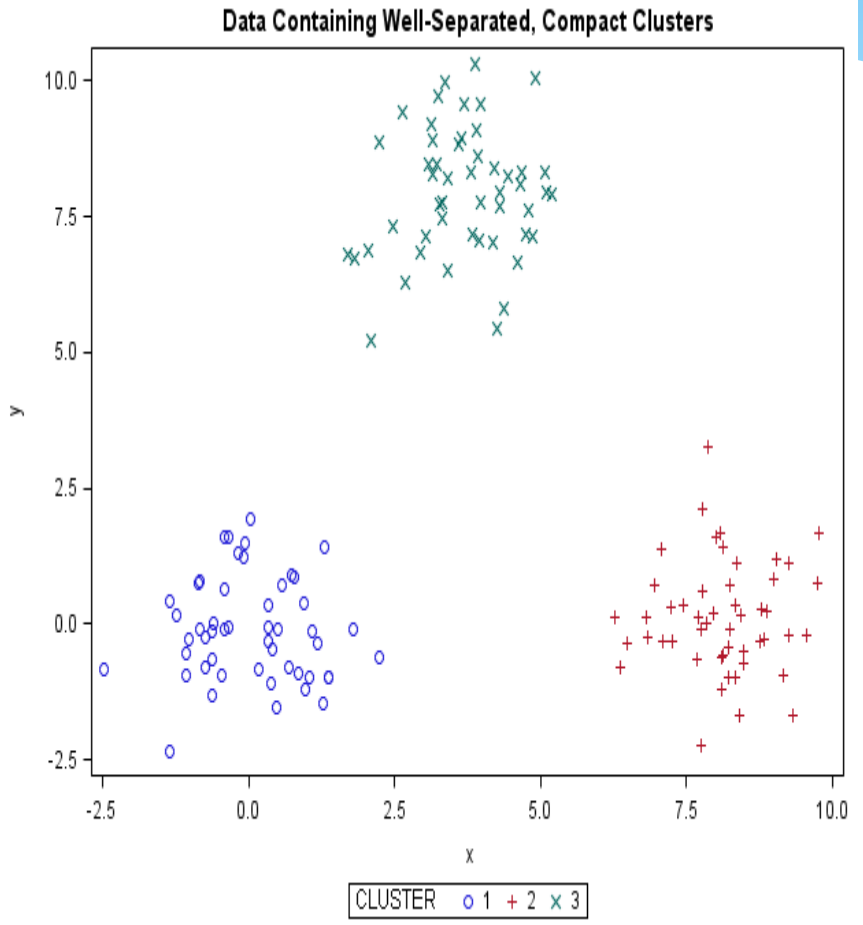
目录

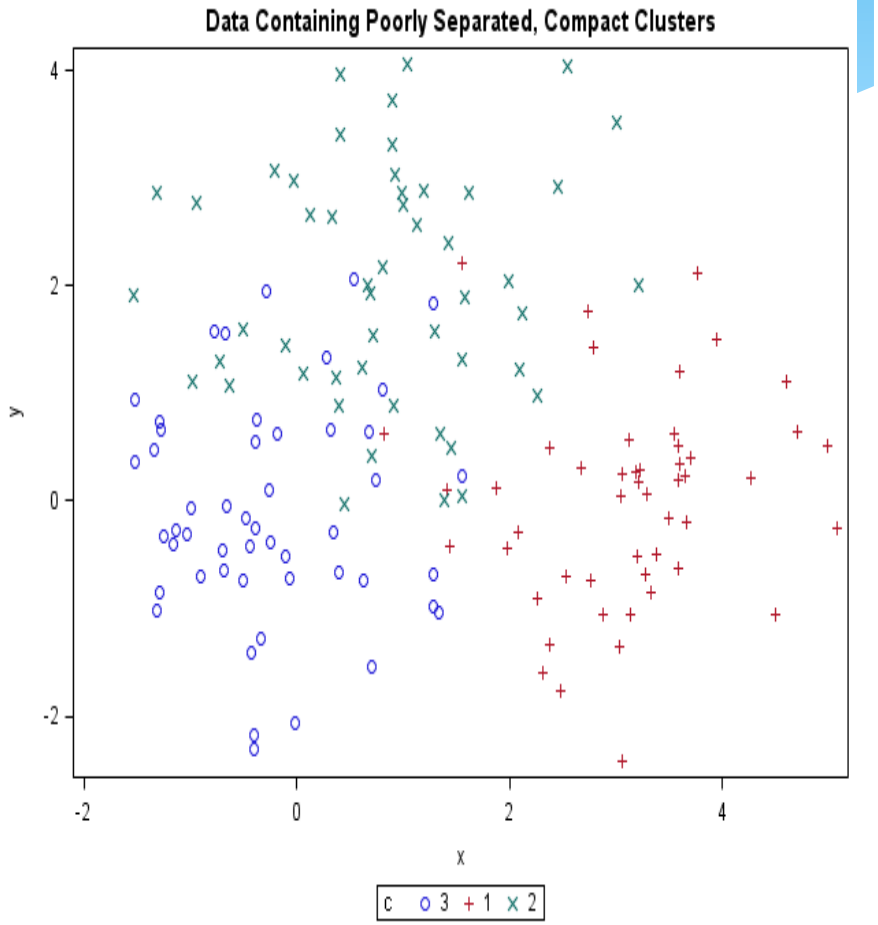
* 聚类分析概述
* 聚类分析数据准备
* 聚类分析技术
* 聚类结果探索
* 聚类结果部署

聚类分析数据准备

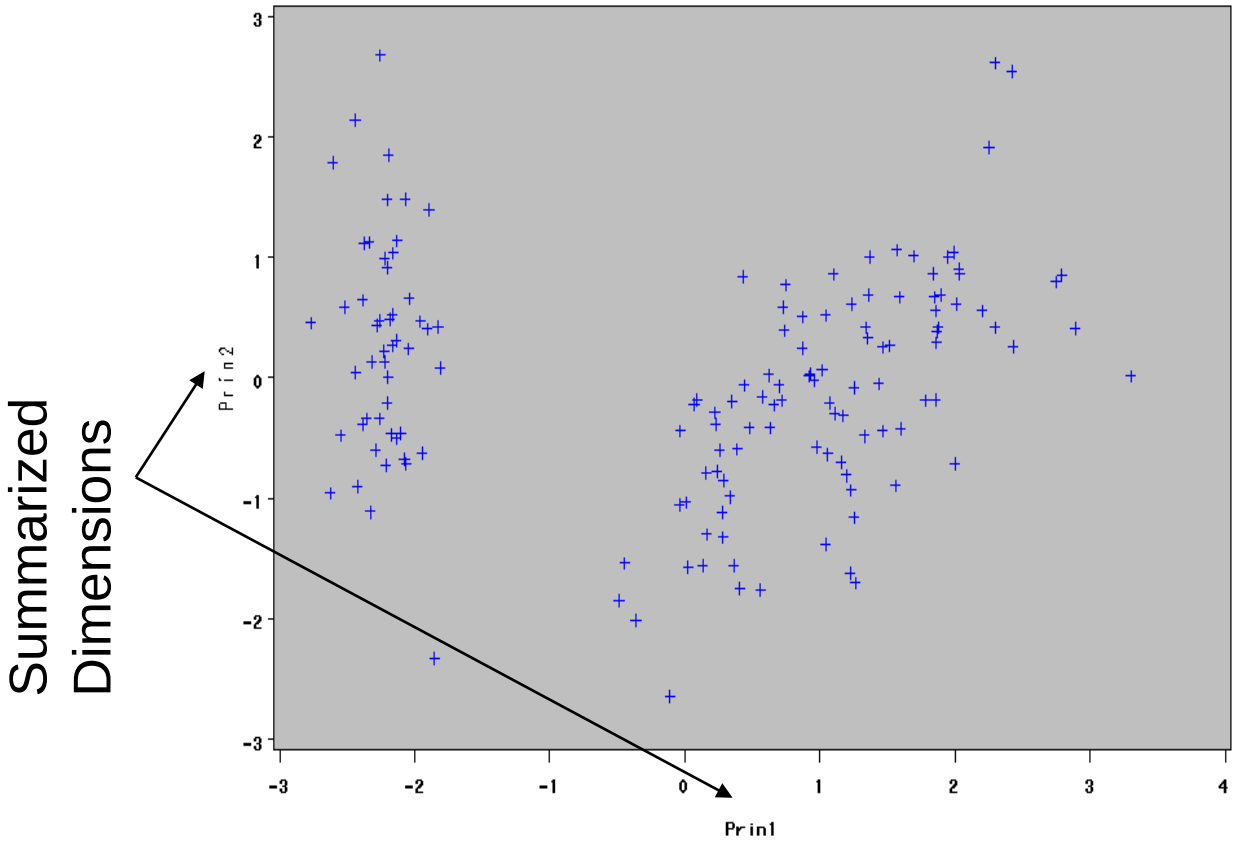
1. 数据和样本选择(*Who am I clustering?*)
2. 变量选择(*What characteristics matter?*)
3. 数据探索(*What shape/how many clusters?*)
4. 标准化(*Are variable scales comparable?*)
5. 数据变换(*Are variables correlated? Are clusters elongated?*)

数据探索（降维）


Data Containing Well-Separated, Compact Clusters
10.0
7.5
5.0
2.5
0.0
-2.5
y
-2.5
0.0
2.5
5.0
7.5
10.0
x
CLUSTER 1 2 3


Data Containing Poorly Separated, Compact Clusters
4
2
0
-2
y
-2
0
2
4
x
c 3 1 2

主成分分析或者多维标度分析（降维）



标准化

1. 中心化变换 $x_{ij}^{*}=x_{ij}-\bar{x}_j \quad (i=1,2,\cdots,n;\ j=1,\cdots,m)$

变换后数据的均值为0，而协差阵不变.

2. 标准化变换

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{s_j} \quad \begin{pmatrix} i=1,2,\cdots,n \\ j=1,2,\cdots,m \end{pmatrix}$$

变换后的数据,每个变量的样本均值为0,标准差为1,而且标准化变换后的数据$\{x^{*}_{ij}\}$与变量的量纲无关.

3. 极差标准化变换

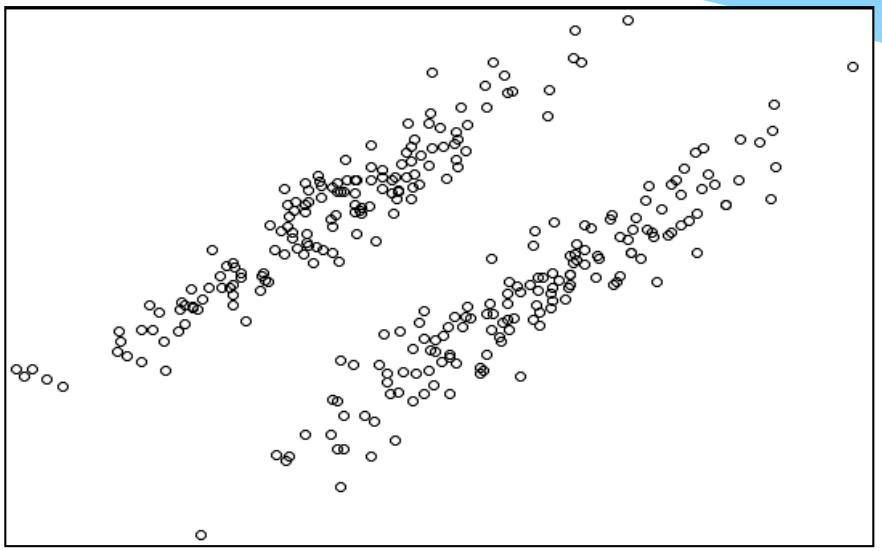
$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{R_j} \quad \begin{pmatrix} i=1,2,\cdots,n \\ j=1,2,\cdots,m \end{pmatrix}$$

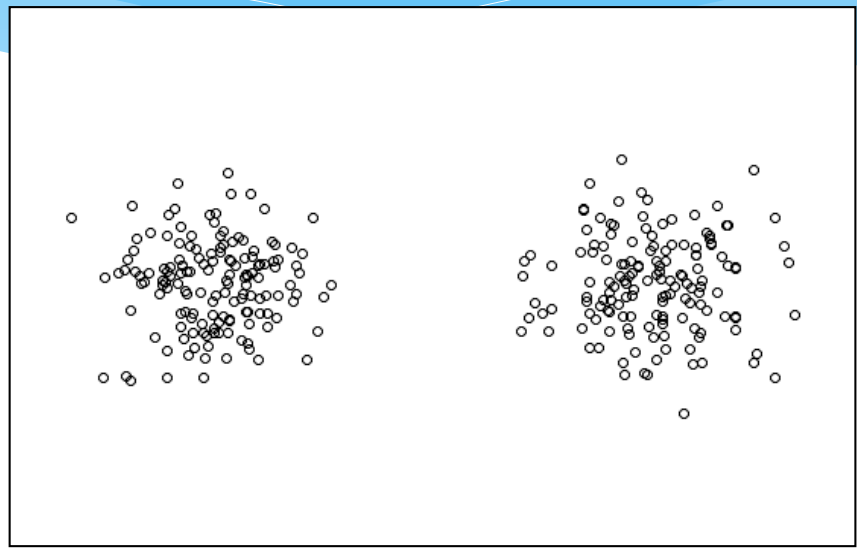
变换后的数据,每个变量的样本均值为0,极差为1,变换后的数据也是无量纲的量.

标准化方法

METHOD	LOCATION	SCALE
MEAN	mean	1
MEDIAN	median	1
SUM	0	sum
EUCLEN	0	Euclidean Length
USTD	0	standard deviation about origin
STD	mean	standard deviation
RANGE	minimum	range
MIDRANGE	midrange	range/2
MAXABS	0	maximum absolute value
IQR	median	interquartile range
MAD	median	median absolute deviation from median
ABW(c)	biweight 1-step M-estimate	biweight A-estimate
AHUBER(c)	Huber 1-step M-estimate	Huber A-estimate
AWAVE(c)	Wave 1-step M-estimate	Wave A-estimate
AGK(p)	mean	AGK estimate (ACECLUS)
SPACING(p)	mid minimum-spacing	minimum spacing
L(p)	L(p)	L(p) (Minkowski distances)
IN(ds)	read from data set	read settings from data set "ds"

数据变换（处理变量间的相关）





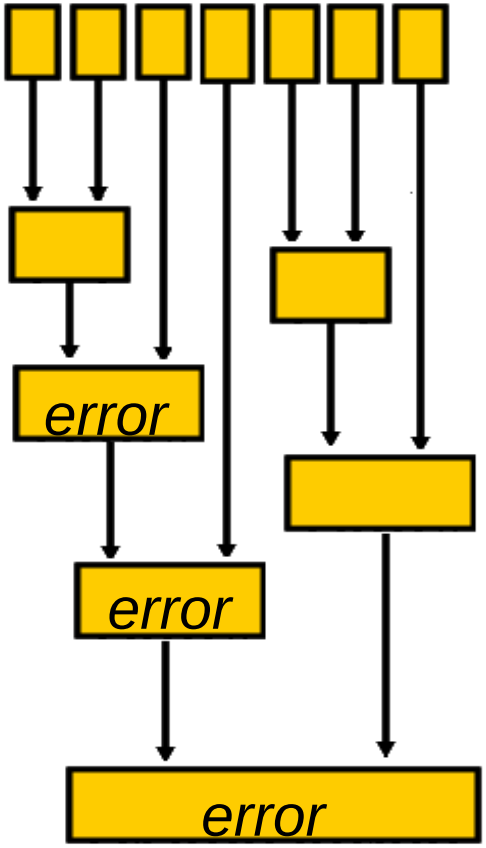
* Before ACECLUS

After ACECLUS

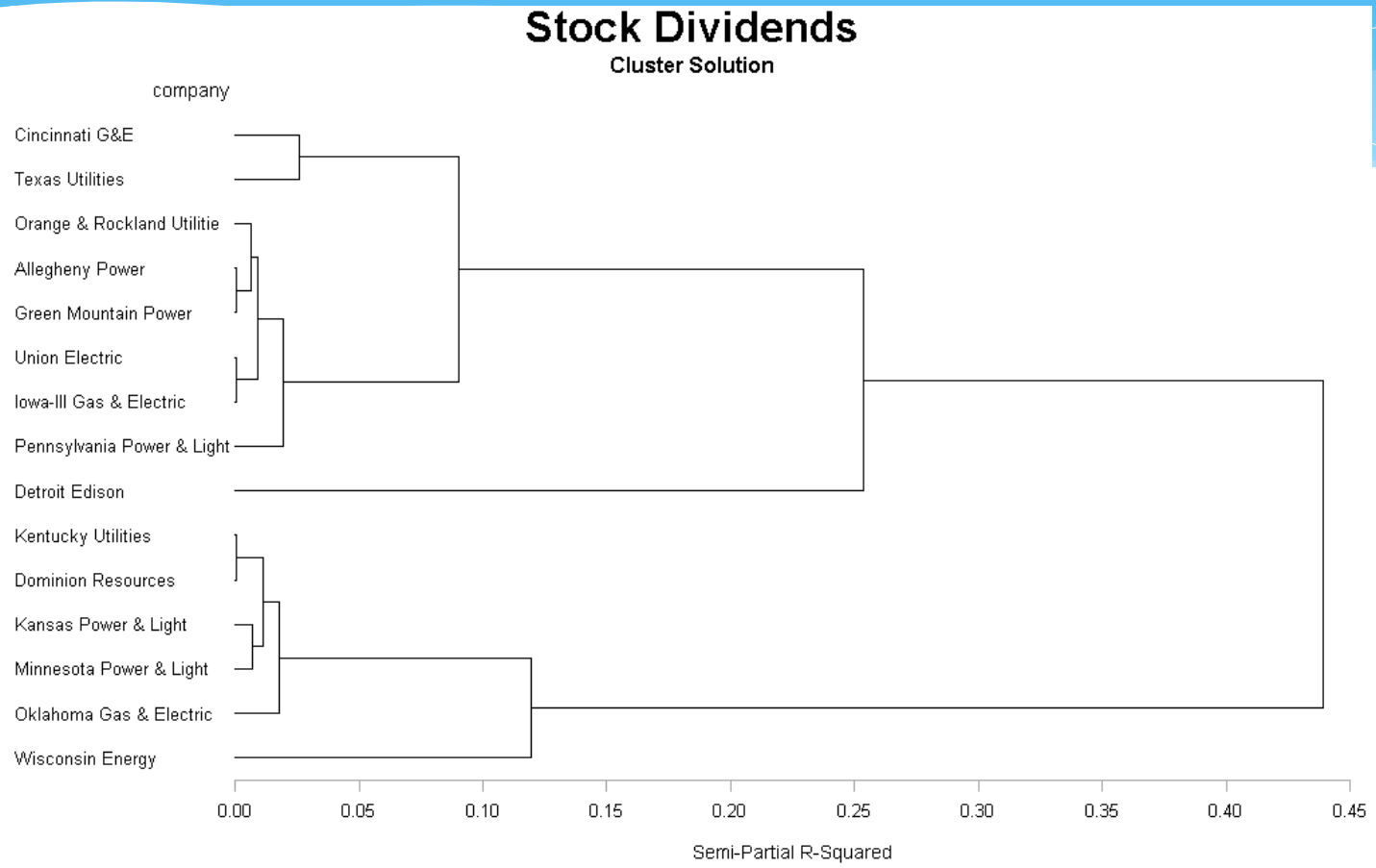
目录

* 聚类分析概述
* 聚类分析数据准备
* 聚类分析技术
* 聚类结果探索
* 聚类结果部署

分层聚类

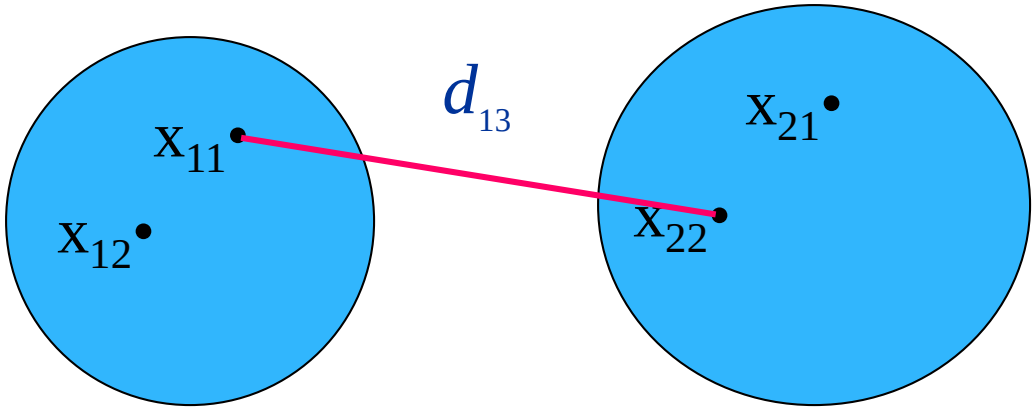


分层聚类


Stock Dividends
Cluster Solution
company
Cincinnati G&E
Texas Utilities
Orange & Rockland Utilitie
Allegheny Power
Green Mountain Power
Union Electric
Iowa-Ill Gas & Electric
Pennsylvania Power & Light
Detroit Edison
Kentucky Utilities
Dominion Resources
Kansas Power & Light
Minnesota Power & Light
Oklahoma Gas & Electric
Wisconsin Energy
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
Semi-Partial R-Squared

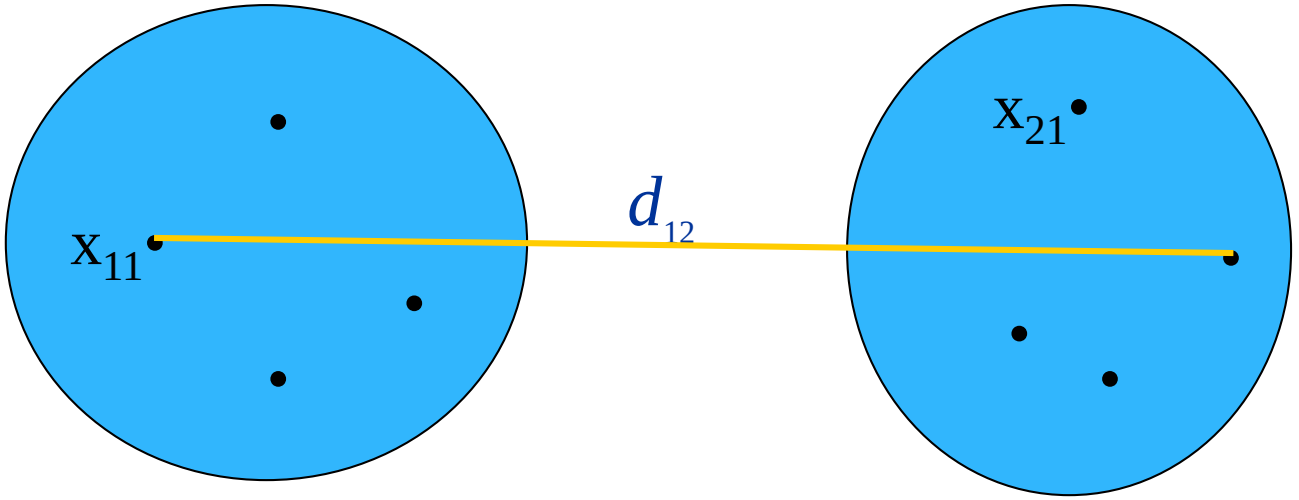
聚类方法

1、 最短距离（Nearest Neighbor)



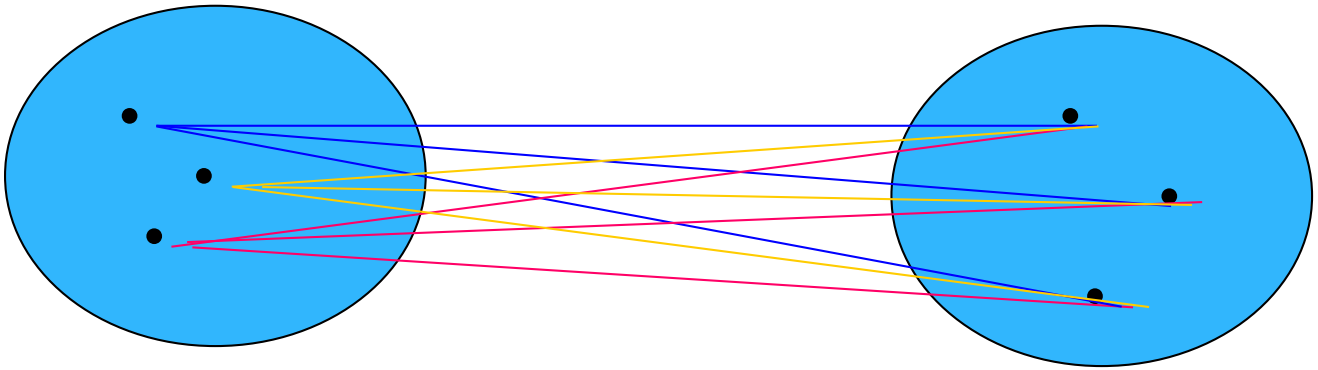
聚类方法

最长距离（Furthest Neighbor）



聚类方法

组间平均连接（Between-group Linkage)

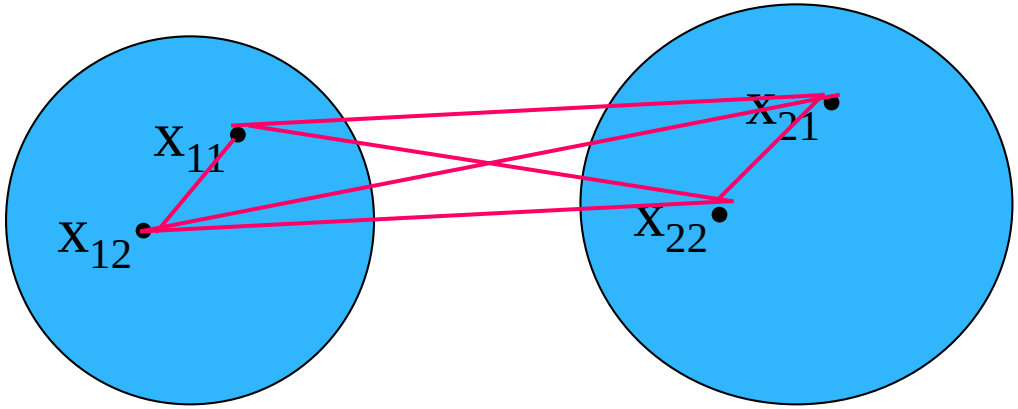


$$\frac{d_1 + \cdots + d_9}{9}$$

聚类方法

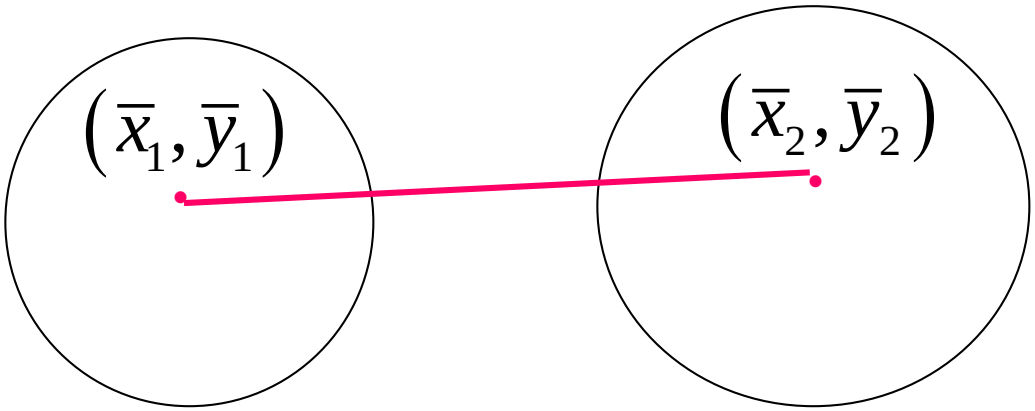
组内平均连接法（Within-group Linkage）

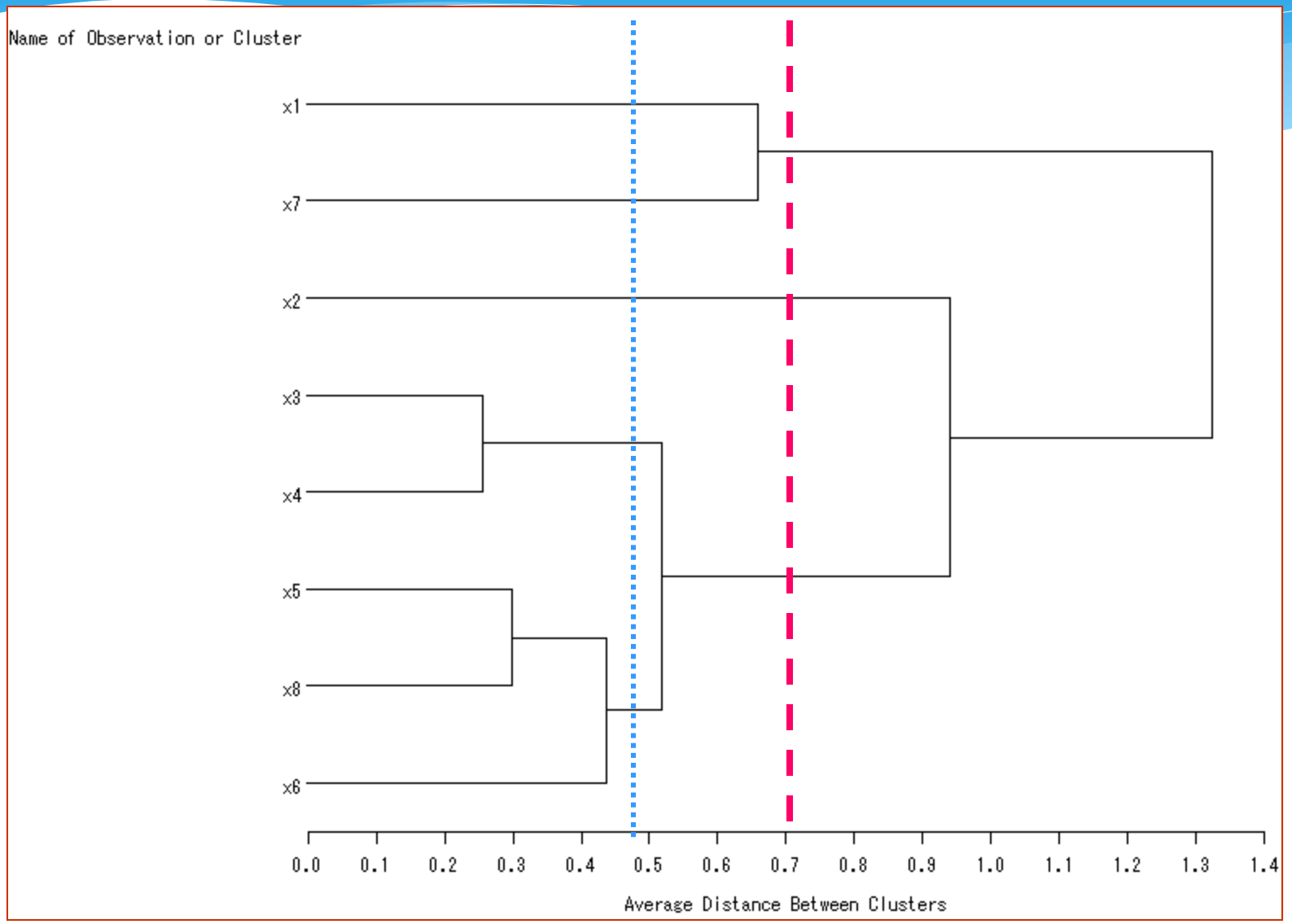
$$\frac{d_1 + d_2 + d_3 + d_4 + d_5 + d_6}{6}$$



聚类方法

重心法（Centroid clustering):均值点的距离




Name of Observation or Cluster
x1
x7
x2
x3
x4
x5
x8
x6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
Average Distance Between Clusters

层次聚类示例

设抽取5个样品，每个样品观察2个指标，

x_1：您每月大约喝多少瓶啤酒，

x_2：您对“饮酒是人生的快乐”这句话的看法如何？观察数据如下，对这5个样品分类。

	x_1	x_2
1	20	7
2	18	10
3	10	5
4	4	5
5	4	3

1.计算5个样品两两之间的距离 d_{ij}（采用欧氏距离），

记为距离矩阵

	②	③	④	⑤
①	3.6	10.2	16.12	16.49
②		9.43	14.87	15.65
③			6	6.32
④				**2**

2. 合并距离最小的两类为新类，按顺序定为第 6 类。

$d_{45}=2$为最小，　⑥$=\{4,5\}$

3、计算新类⑥与各当前类的距离，

$$d_{61}=\min\{d_{41},d_{51}\}=\min\{16.12,16.49\}=16.12$$

$$d_{62}=\min\{d_{42},d_{52}\}=\min\{14.87,15.65\}=14.87$$

$$d_{63}=\min\{d_{43},d_{53}\}=6$$

得距离矩阵如下：

	②	③	⑥
①	**3.6**	10.2	16.12
②		9.43	14.87
③			6

4、重复步骤2、3，合并距离最近的两类为新类，直到所有的类并为一类为止。

$d_{12}=3.6$ 为最小，⑦= $\{1,2\}$

$d_{73}=\min\{d_{13},d_{23}\}=9.43$　　$d_{76}=\min\{d_{16},d_{26}\}=14.87$

	⑥	⑦
③	**6**	9 .43
⑥		14.87

5、$d_{36}=6$ 为最小，⑧= $\{3,6\}$

$d_{87}=\min\{d_{37},d_{67}\}=9.43$

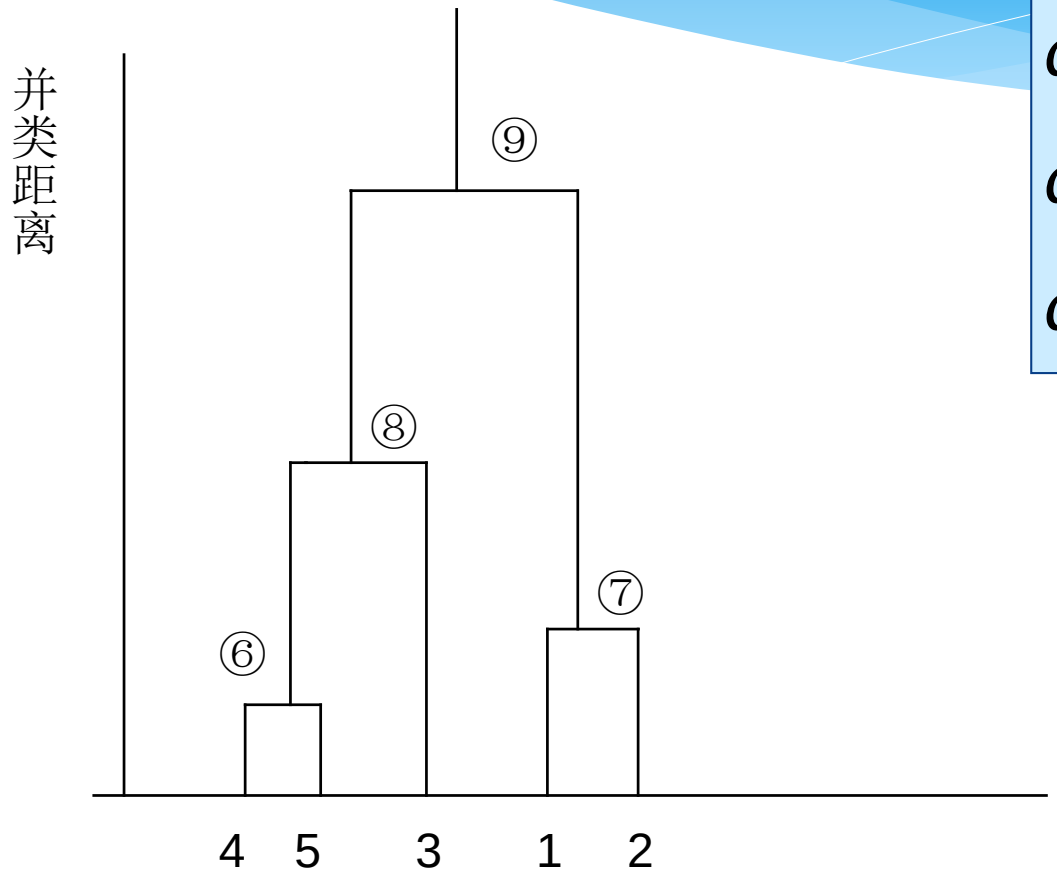
6、按聚类的过程画聚类谱系图

$$d_{4,5}=2$$

$$d_{1,2}=3.6$$

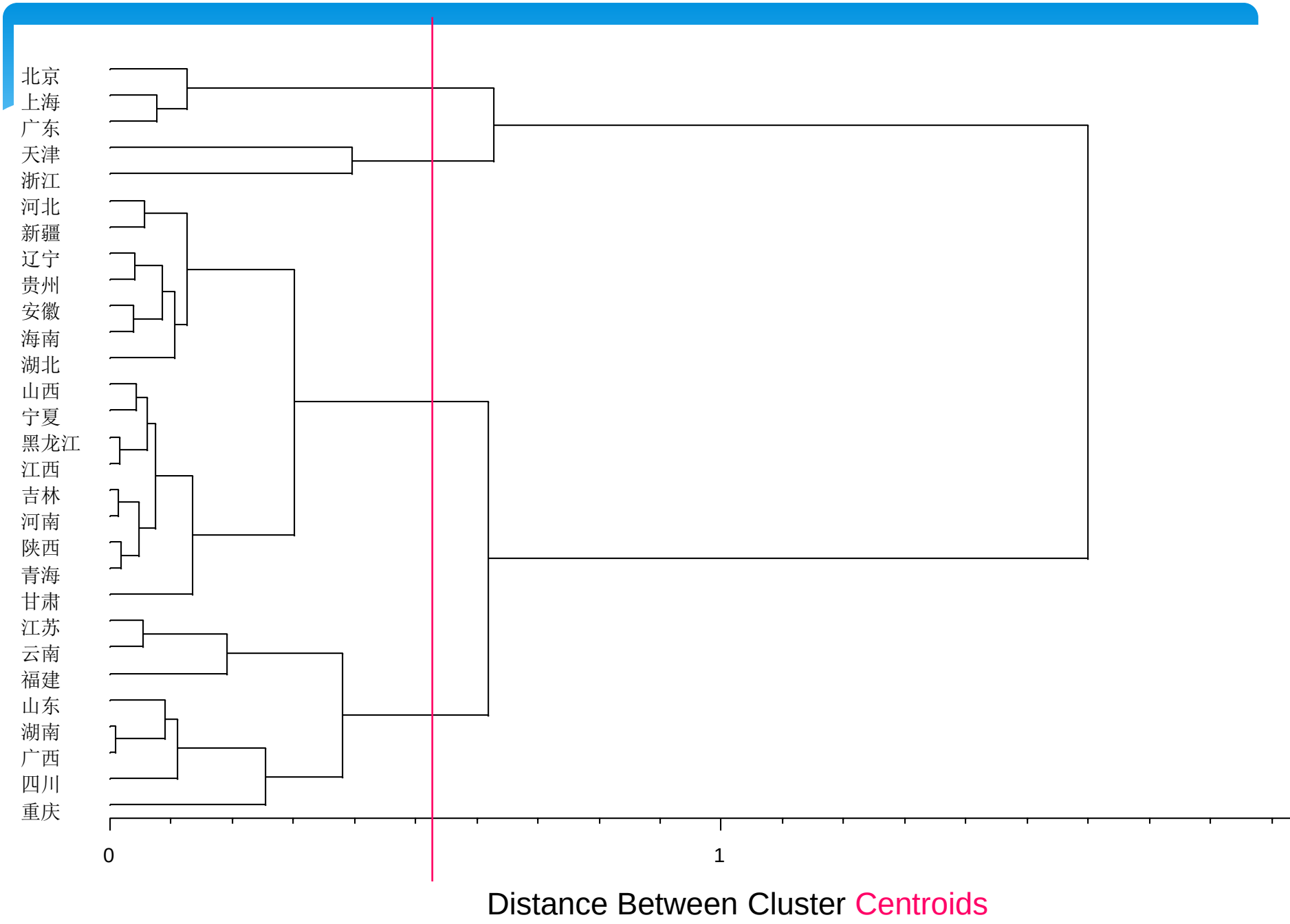
$$d_{3,6}=6$$

$$d_{7,8}=9.43$$

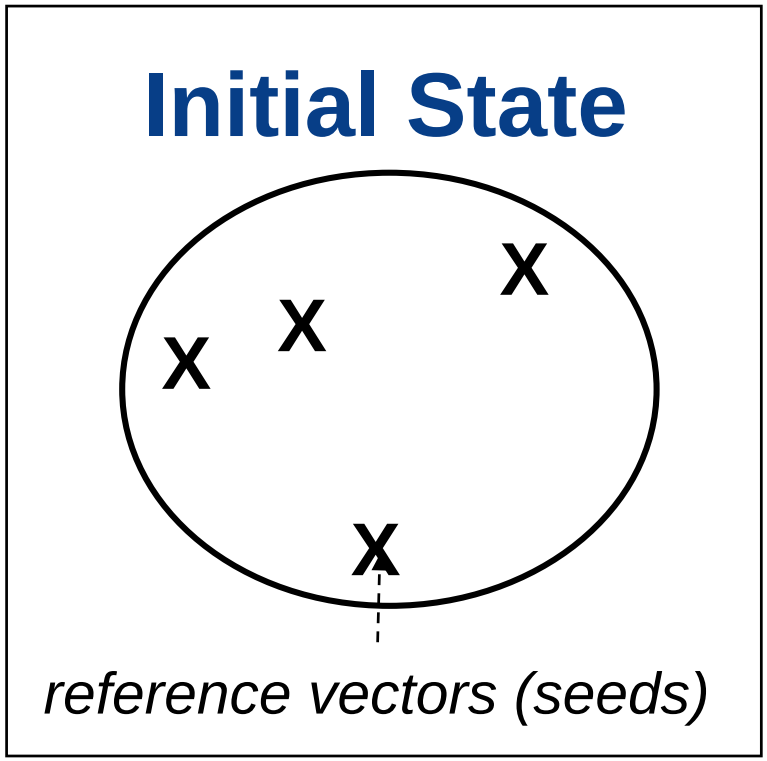


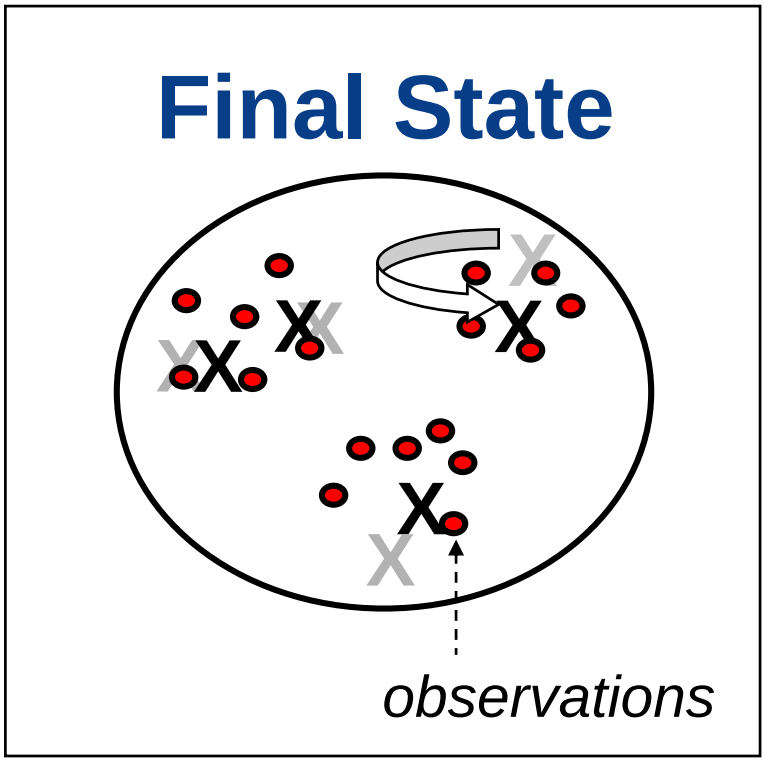
7、决定类的个数与类。

观察此图，我们可以把5个样品分为3类，$\{1,2\}$、$\{3\}$、$\{4,5\}$。

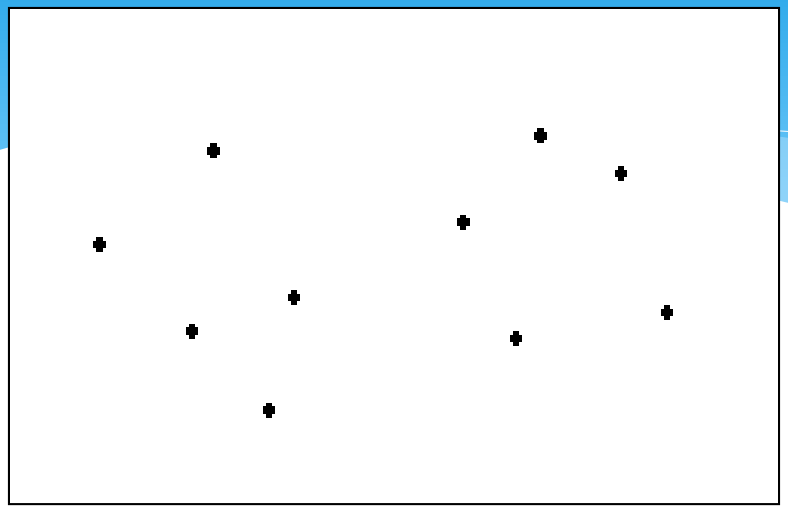

北京
上海
广东
天津
浙江
河北
新疆
辽宁
贵州
安徽
海南
湖北
山西
宁夏
黑龙江
江西
吉林
河南
陕西
青海
甘肃
江苏
云南
福建
山东
湖南
广西
四川
重庆
0
1
Distance Between Cluster Centroids

划分聚类

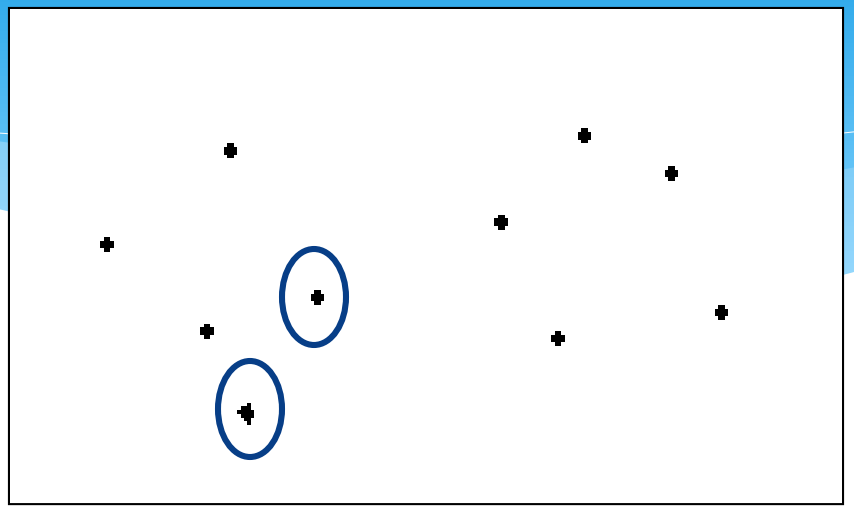

Initial State
reference vectors (seeds)


Final State
observations

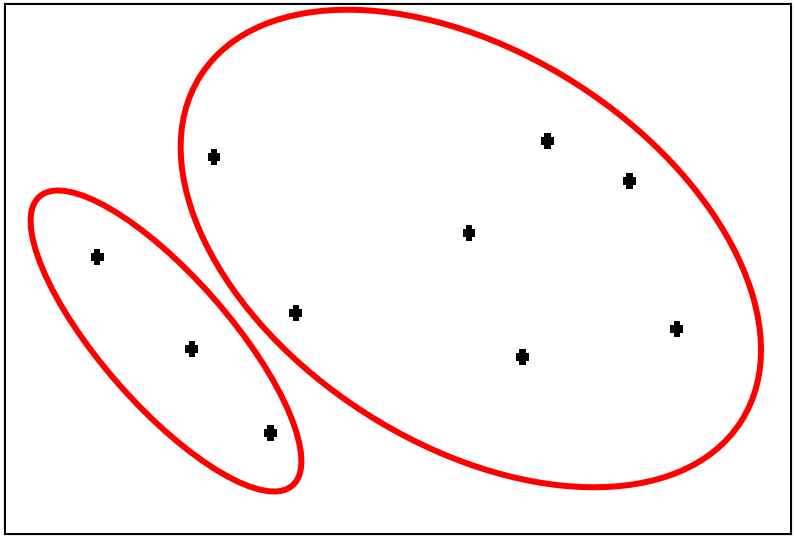
（a）空间的群点



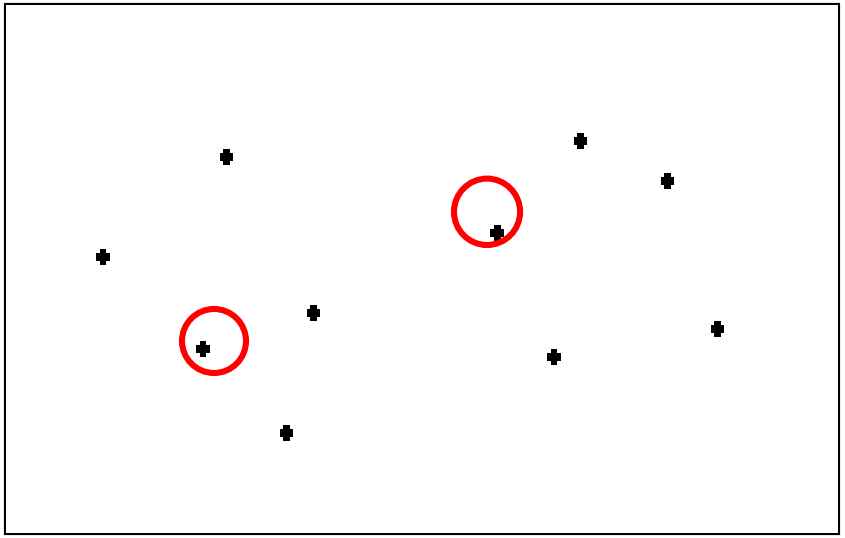
(b) 任取两个凝聚点



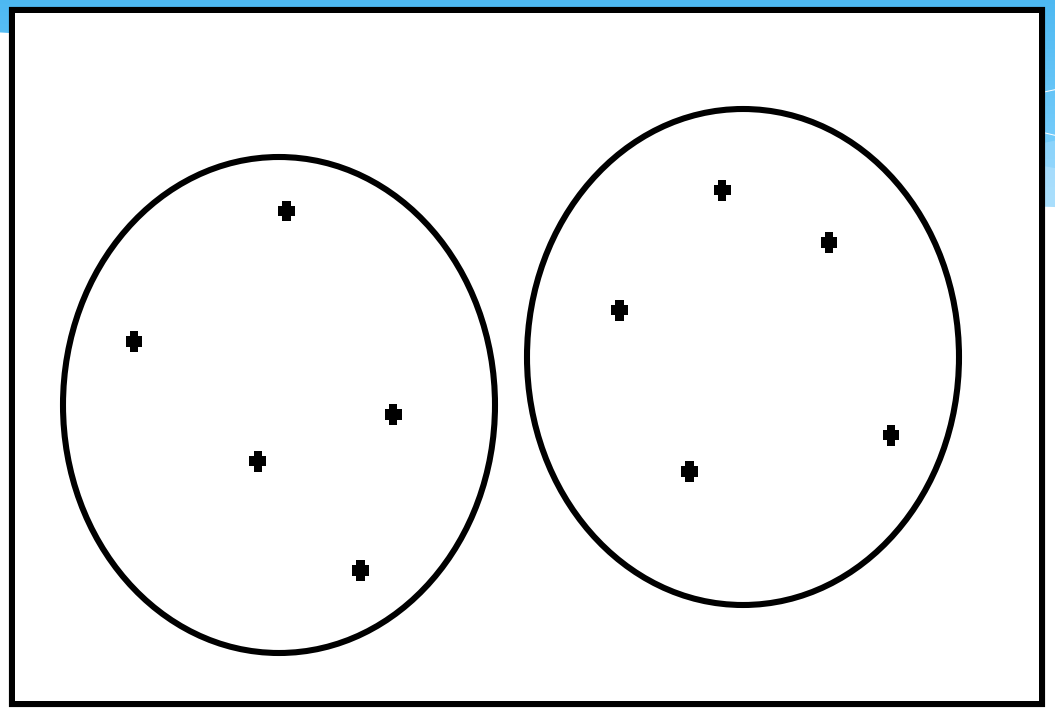
(c) 第一次分类



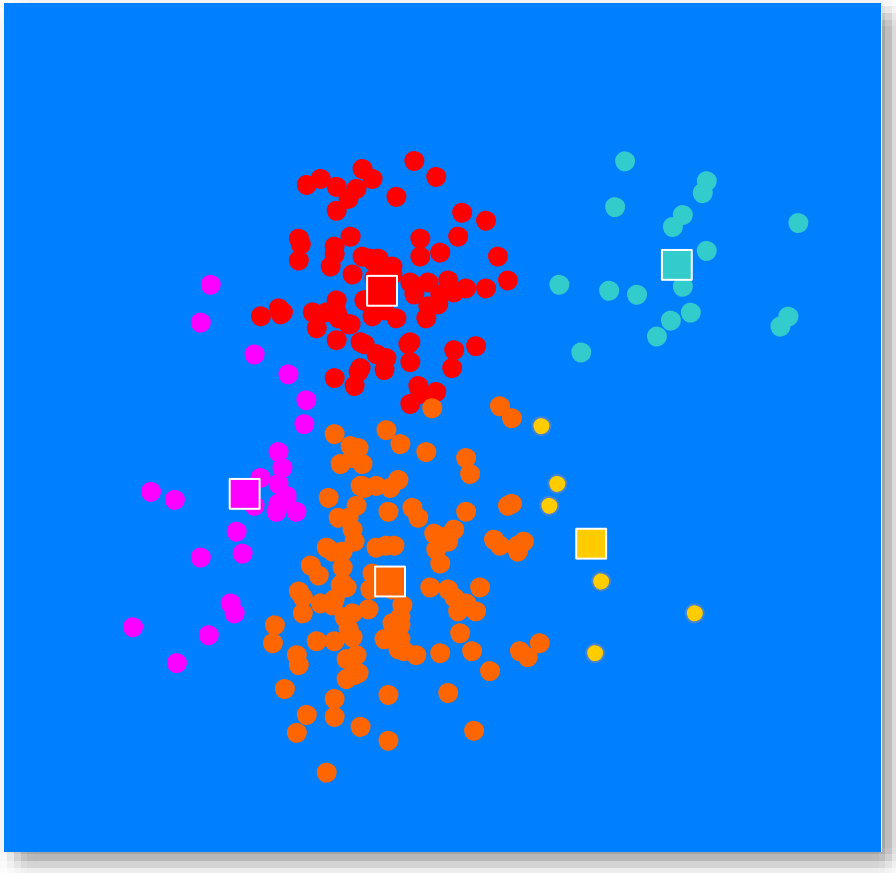
(d) 求各类中心



(e) 第二次分类



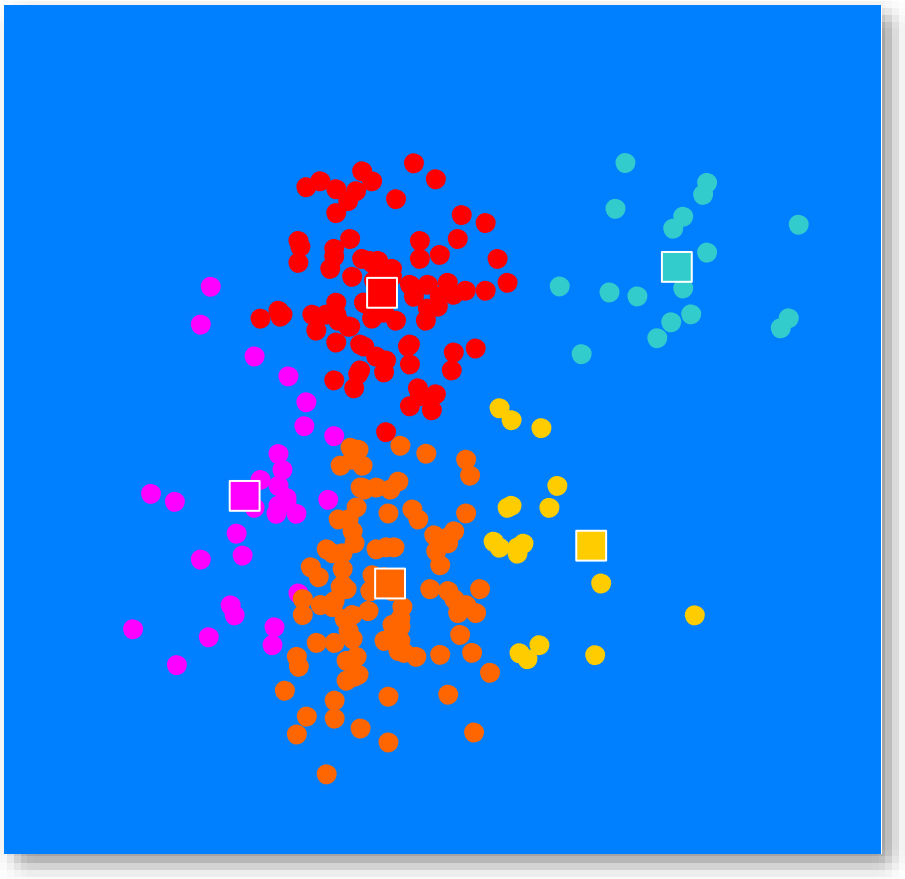
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. Repeat steps 4 and 5 until convergence.

...

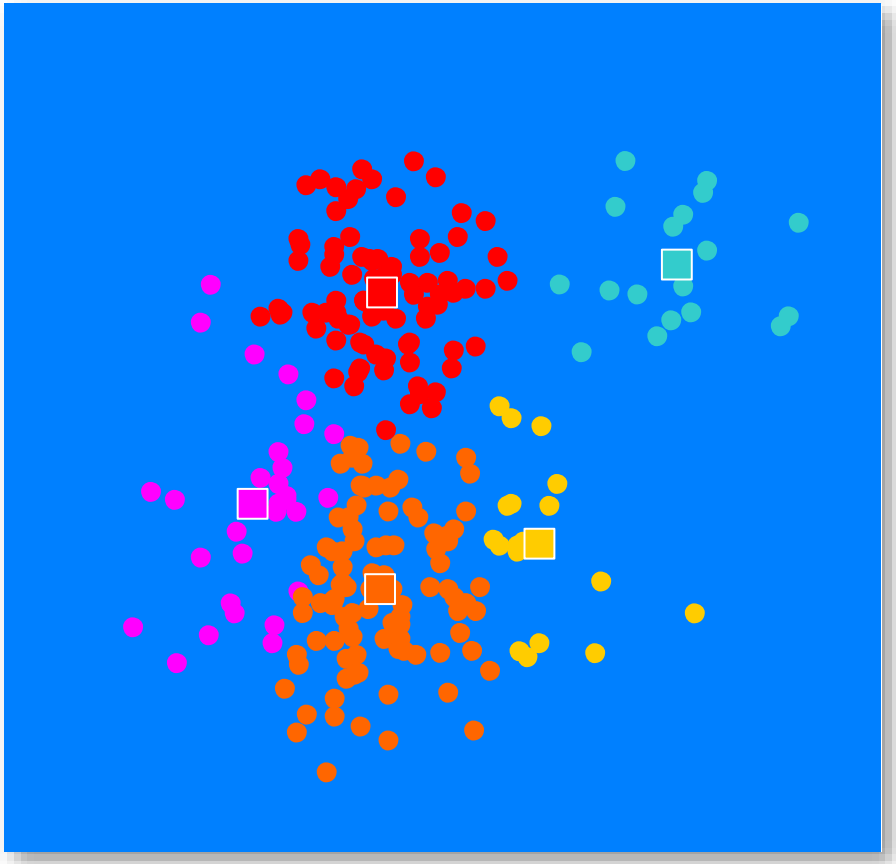
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. Repeat steps 4 and 5 until convergence.

...

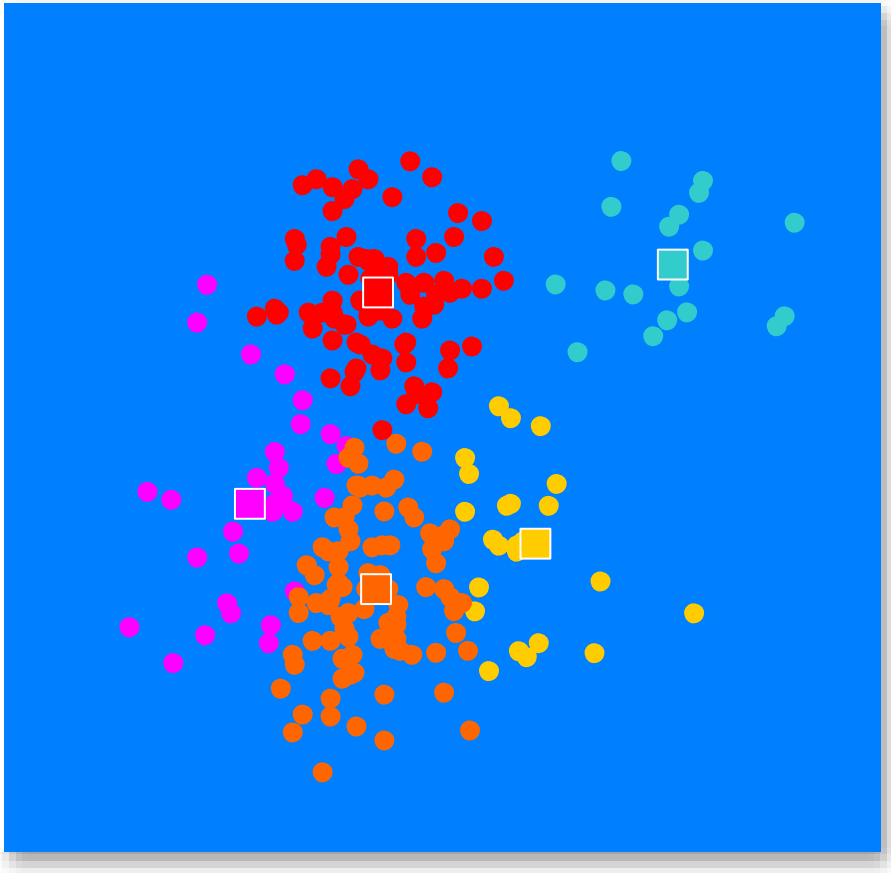
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

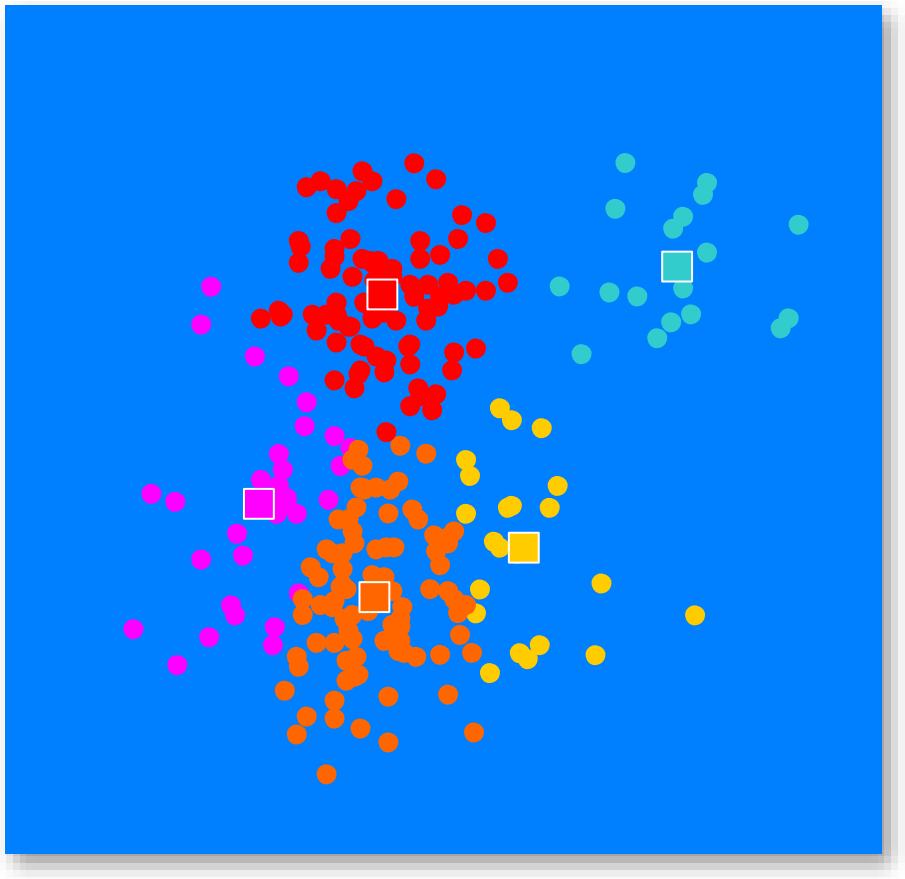
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. **Repeat steps 4 and 5 until convergence.**

...

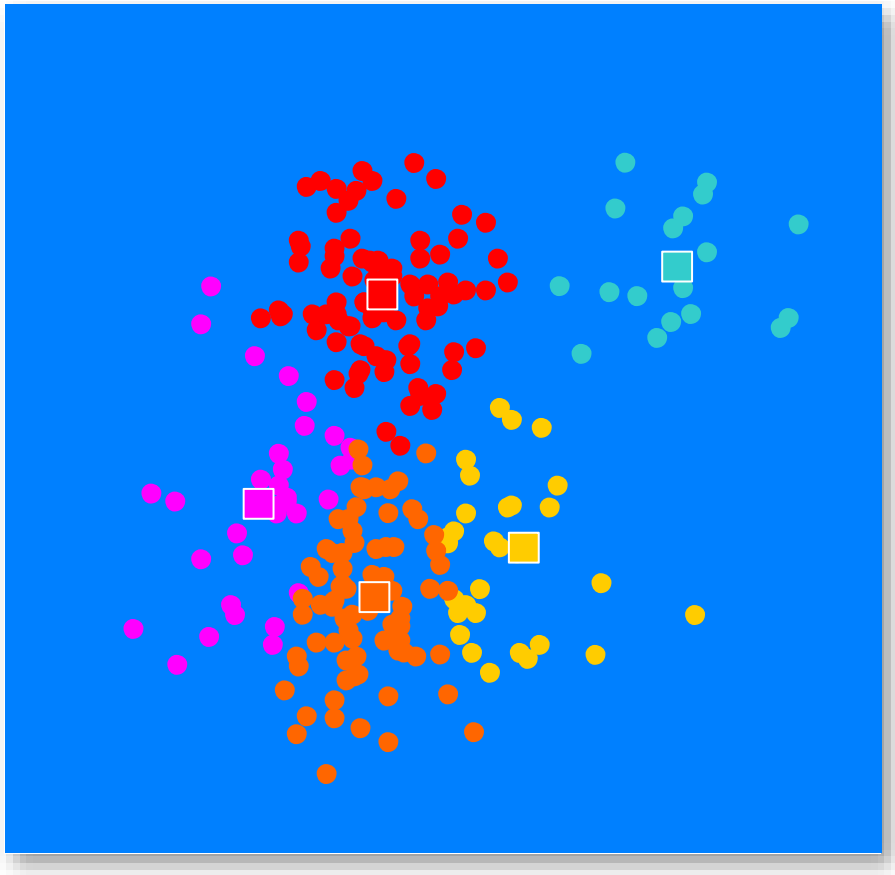
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

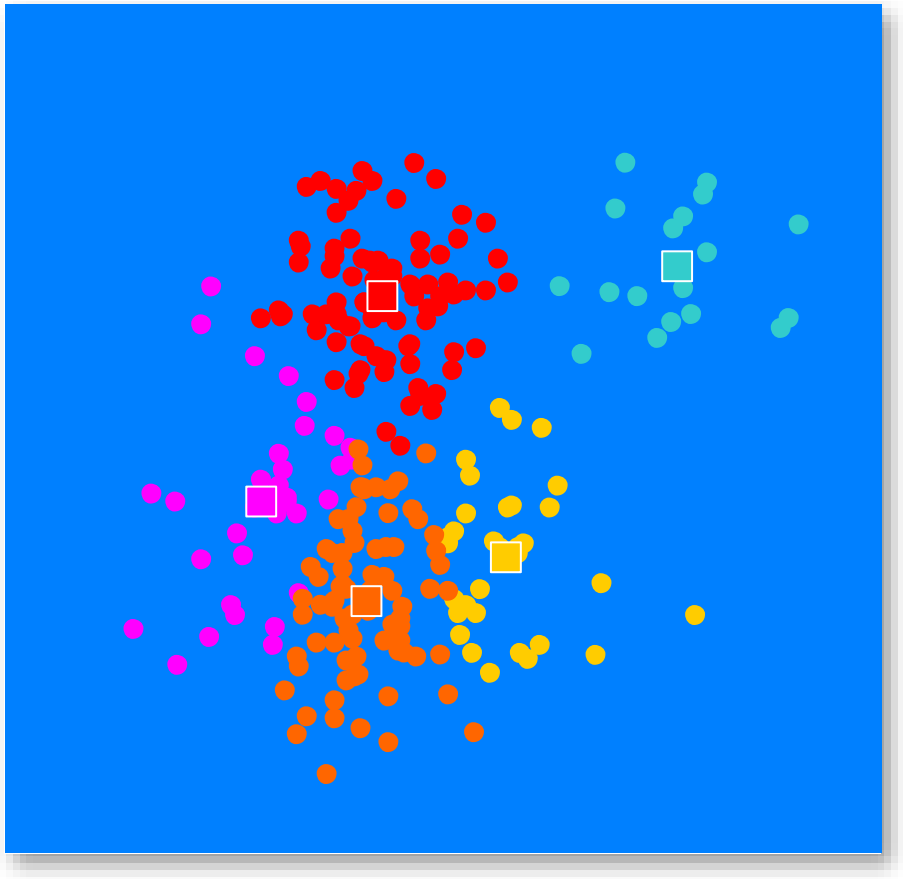
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. **Repeat steps 4 and 5 until convergence.**

...

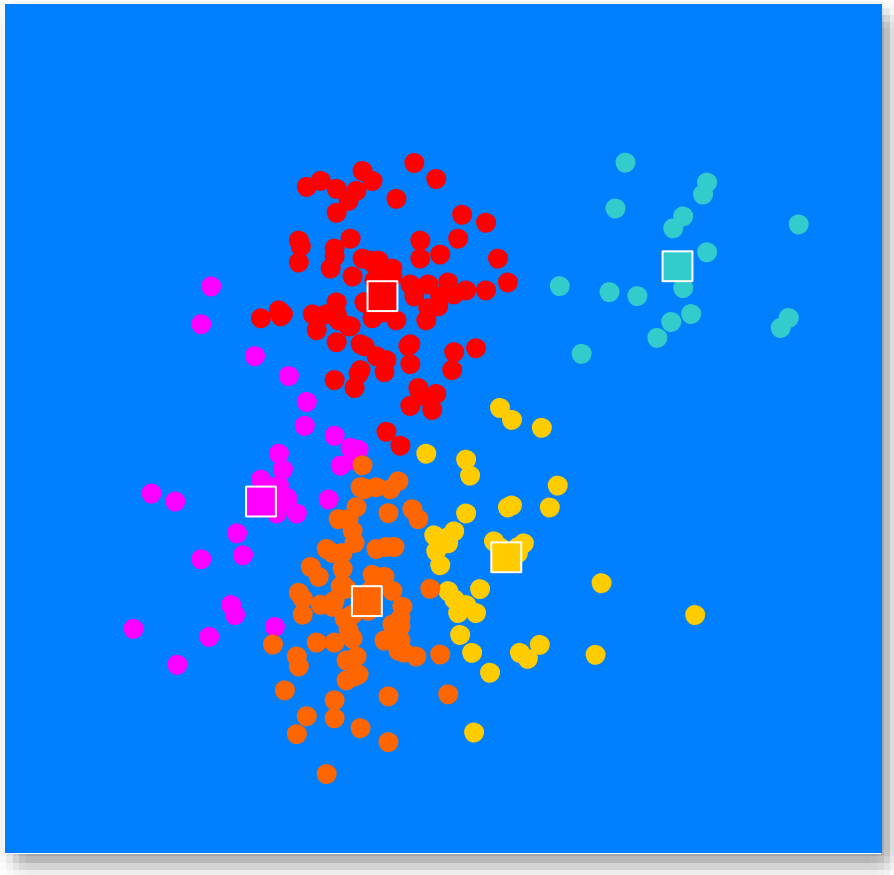
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

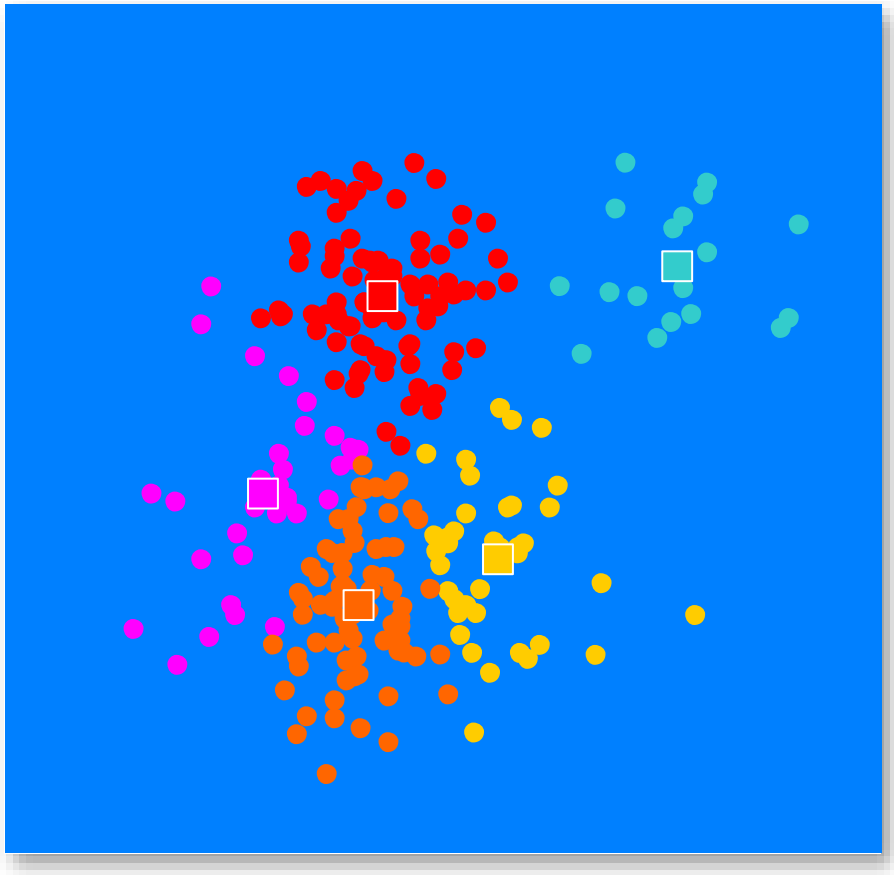
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. **Reassign cases.**
6. **Repeat steps 4 and 5 until convergence.**

...

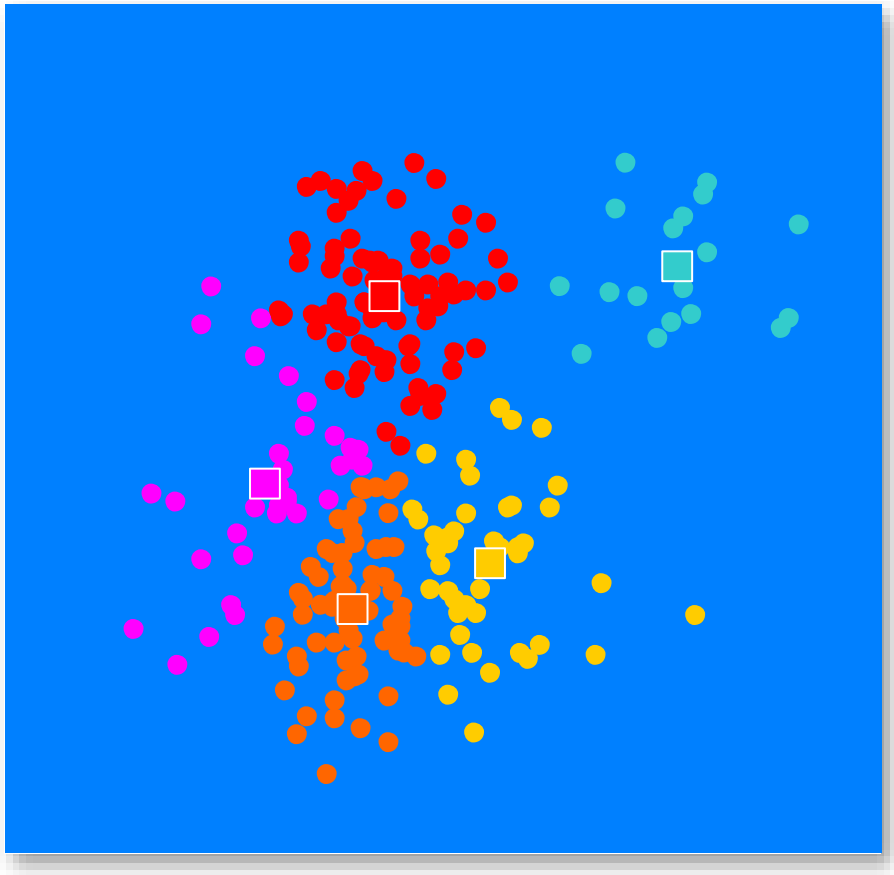
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. **Update cluster centers.**
5. Reassign cases.
6. **Repeat steps 4 and 5 until convergence.**

...

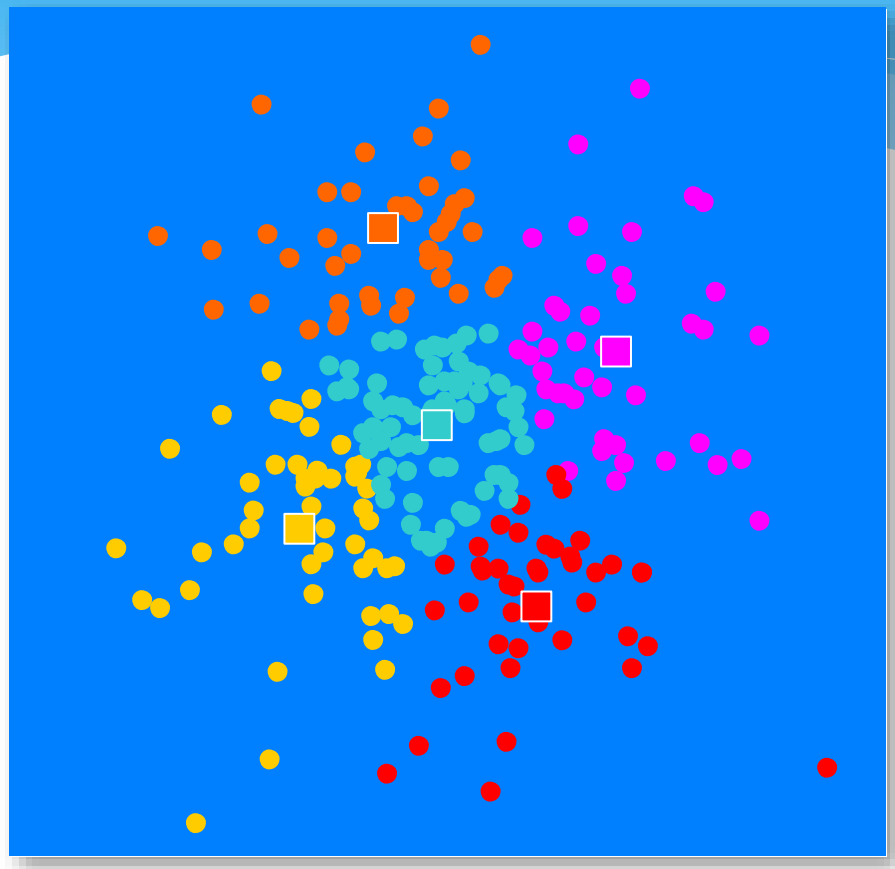
K均值聚类



1. Select inputs.
2. Select *k* cluster centers.
3. Assign cases to closest center.
4. Update cluster centers.
5. Reassign cases.
6. Repeat steps 4 and 5 until **convergence.**

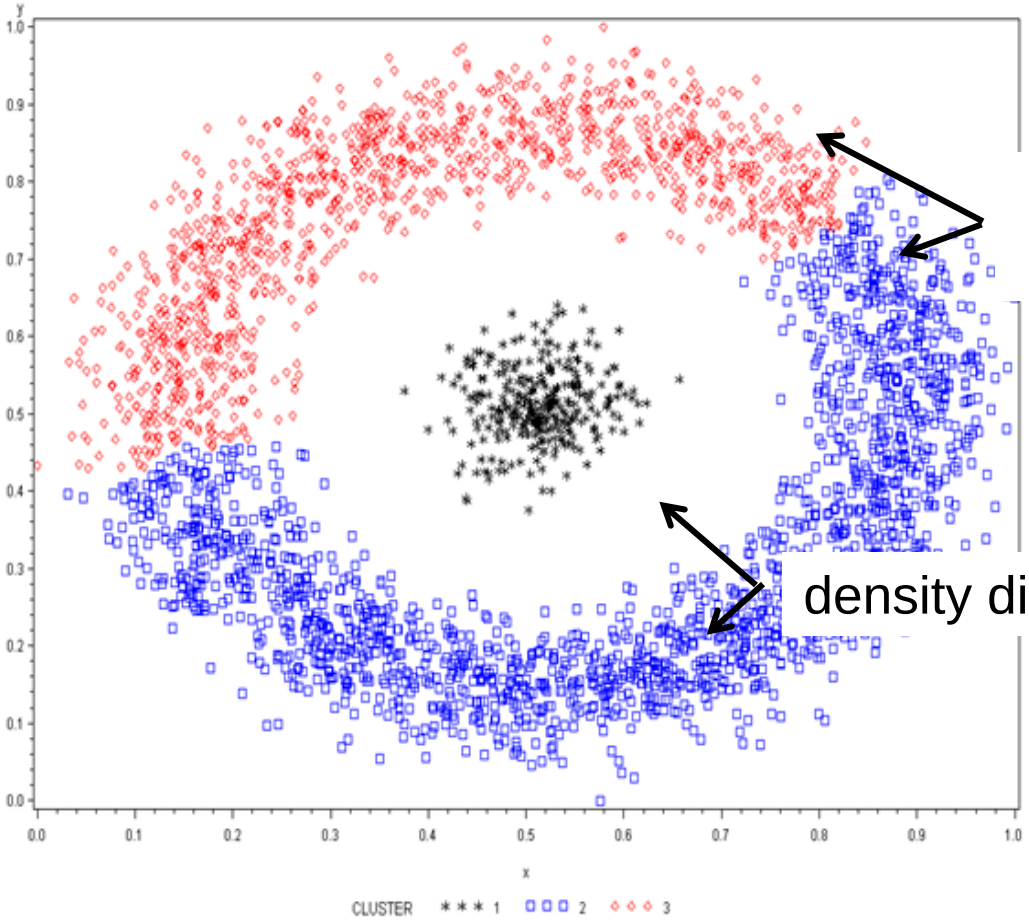
...

K均值聚类



When no clusters exist, use the *k*-means algorithm to partition cases into contiguous groups.

非参聚类



no density difference

density difference

目录

* 聚类分析概述
* 聚类分析数据准备
* 聚类分析技术
* 聚类结果探索
* 聚类结果部署

类别轮廓分析

1. For the cluster being profiled (cluster *k*), classify each observation as being a member of cluster *k* (with a value of 1) or not a member of cluster *k* (with a value of 0).
2. Use logistic regression analysis to rank order the input variables in their ability to distinguish cluster *k* from the others.
3. Generate a comparative plot of cluster *k* and the rest of the data.

示例：客户聚类分析

* 在众多客户当中，如果不进行分类，通常无法进行行为和习惯的分析.
  * Bargain hunter
  * Man/woman on a mission
  * Impulse shopper
  * Weary parent
  * DINK (dual income, no kids)

示例：商店聚类分析

* You want to open new grocery stores in the U.S. based on demographics. Where should you locate the following types of new stores?
  * low-end budget grocery stores
  * small boutique grocery stores
  * large full-service supermarkets

Example: 时尚分类

* Based on the four styles of pants that your customers can purchase, can you identify stores as serving similar fashion types?

* country-club dresser
* fashion trendsetter
* comfort kick-back dresser

16组中呈现出明显的优势、弱势特征

组号	优势特征	弱势特征	描述性名称
#1	语音每次呼叫时间、香港（澳门）呼叫、非繁忙时段呼叫	繁忙时呼叫、IP呼叫、短信、转移	业余活跃组
#2	繁忙时段月均呼叫次数、漫游地区呼叫、香港呼叫次数	转移呼叫、短信、转移	业务繁忙组
#4	IP呼叫、转移呼叫		贵中求惠组
#6	IP呼叫	短信、转移	IP手机组
#9	IP呼叫、短信	非繁忙时段呼叫	新生潜力组
#12	非繁忙时段呼叫	漫游地区呼叫、转移、短信	夜间积极组
#14	繁忙时段月均呼叫次数	漫游呼叫、非繁忙呼叫、转移	本地繁忙组
#16	繁忙时段月均呼叫次数、转移呼叫、香港（澳门）呼叫	IP呼叫	繁忙大客户组
#8	短信	转移呼叫、IP	短信专家组
#11	转移呼叫	繁忙时段月均呼叫次数、短信	热衷转移组
#15	漫游地区呼叫	短信、繁忙呼叫次数	频繁出差组
#3	语音每次呼叫时间	繁忙时段次数、短信	情深语长组
#5		繁忙时段次数、每次呼叫时间、短信	消极等待组
#7	呼入/呼出比	短信	等待接听组
#10		繁忙时段次数、呼入/呼出比、每次呼叫时间	休眠组
#13		繁忙时段月均呼叫次数	寂寞无声组

聚类结果分析

根据每个类中变量的取值情况，为每个类取名称

聚类均值						
聚类	入会之日算	最近一次消费距今时间长度	代表客户在一定时间内的消费频率	在一定时间内的升级里程	在一定时间内所乘航班的平均舱位折扣系	客户类型
1	-0.60	-0.92	1.31	1.33	0.34	重要价值
2	-0.45	-0.12	-0.05	-0.35	-1.19	重要发展
3	1.19	-1.34	1.74	1.75	0.63	重要价值
4	-1.23	0.36	-0.37	-0.21	0.49	一般发展
5	0.19	1.27	-1.29	-1.18	1.05	一般挽留
6	1.26	-0.12	0.26	0.25	-0.18	重要保持
7	-0.22	1.34	-1.37	-1.35	-0.88	一般保持
8	0.05	-0.49	0.25	0.33	0.39	一般价值

目录

* 聚类分析概述
* 聚类分析数据准备
* 聚类分析技术
* 聚类结果探索
* 聚类结果部署

* 谢谢各位

* Q&A