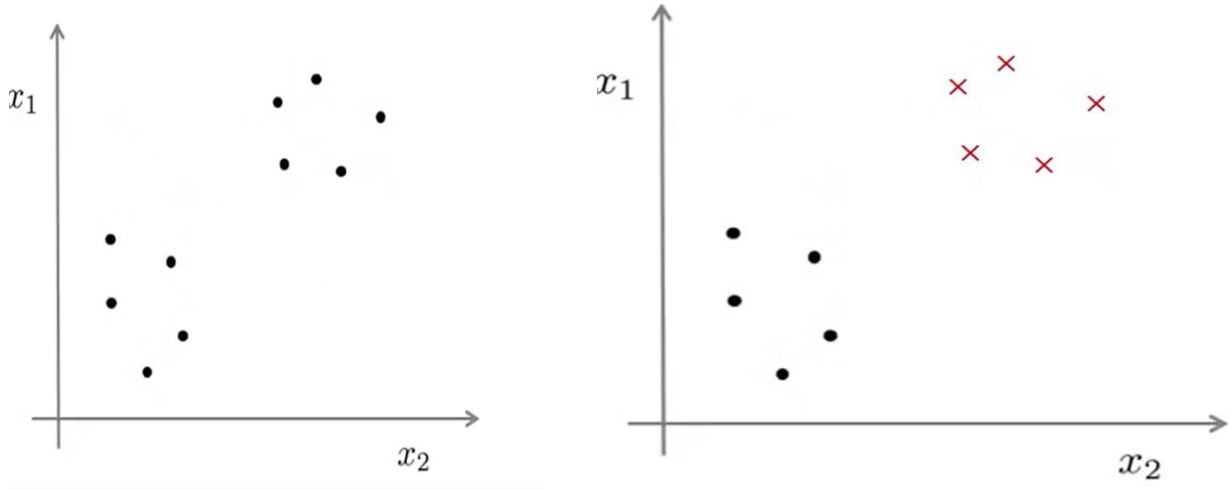
ZIXING AI
自兴人工智能

无监督学习
(Unsupervised learning-)

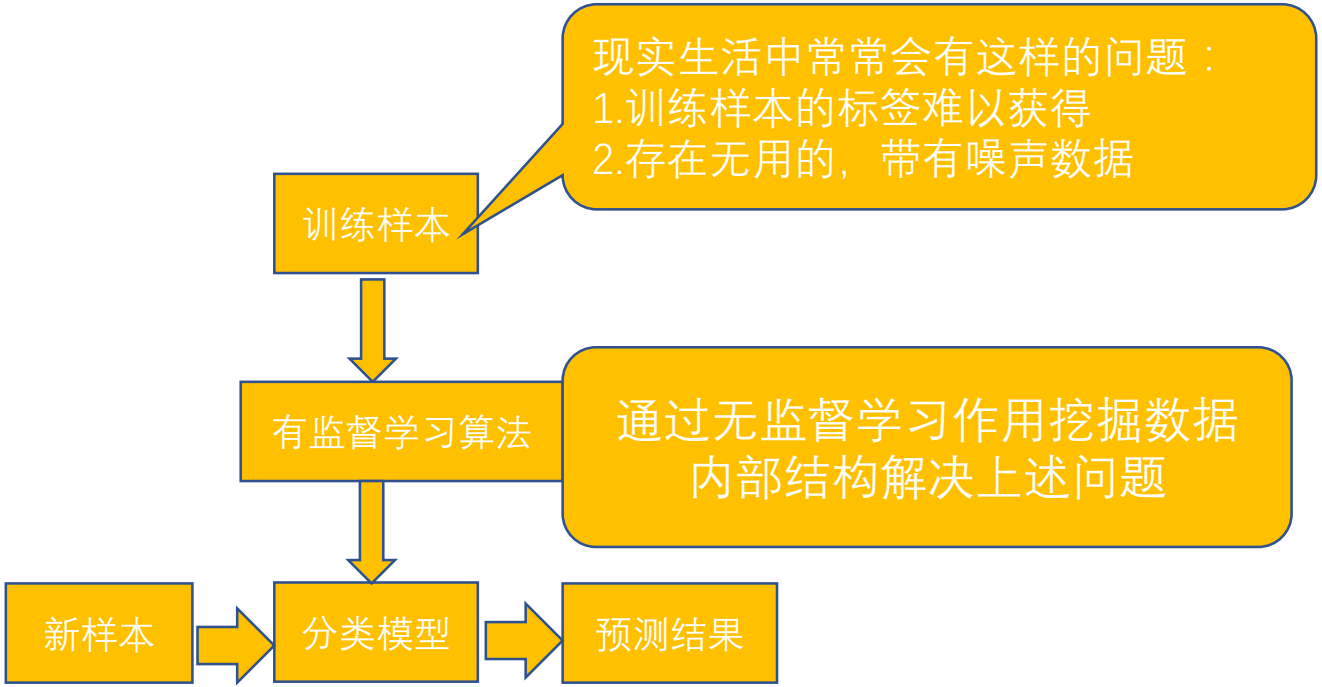
高琰

无监督学习



只给计算机训练数据，不给结果（标签），因此计算机无法准确地知道哪些数据具有哪些标签，只能凭借强大的计算能力分析数据的特征，发现数据本身的内部结构特点。

为什么要无监督学习？


现实生活中常常会有这样的问题：
1.训练样本的标签难以获得
2.存在无用的，带有噪声数据
训练样本
有监督学习算法
通过无监督学习作用挖掘数据
内部结构解决上述问题
新样本
分类模型
预测结果

无监督学习的应用背景

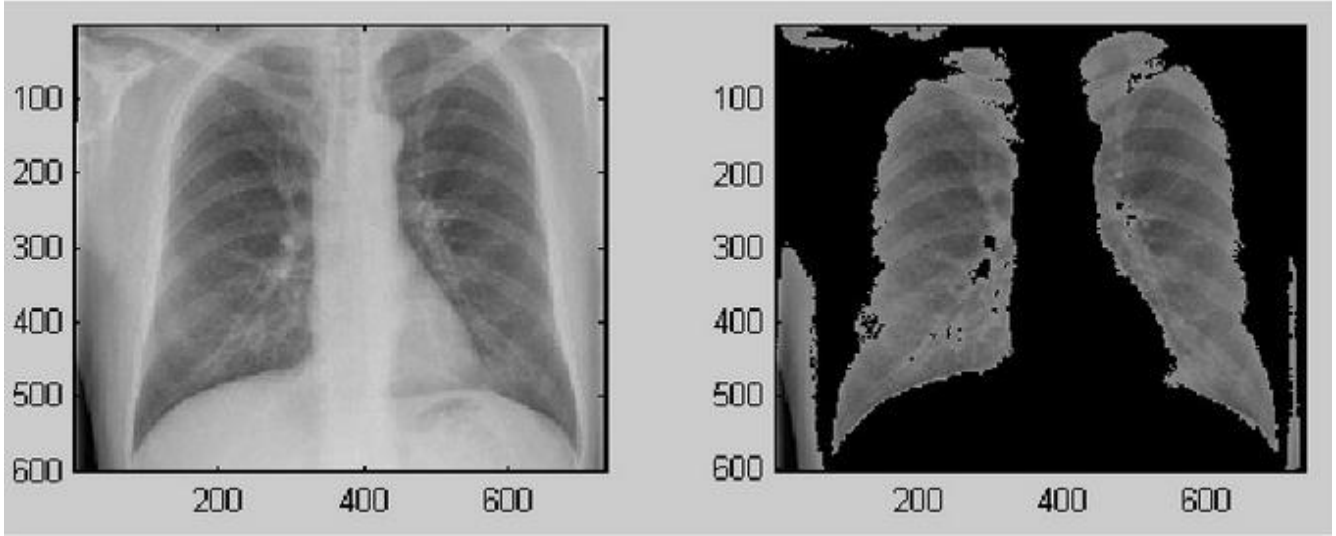
聚类

- 常见的应用背景包括：
  - 从庞大的样本集合中选出一些具有代表性的加以标注用于分类器的训练。
  - 先将所有样本自动分为不同的类别，再由人类对这些类别进行标注。
  - 在无类别信息情况下，寻找好的特征。

特征分析

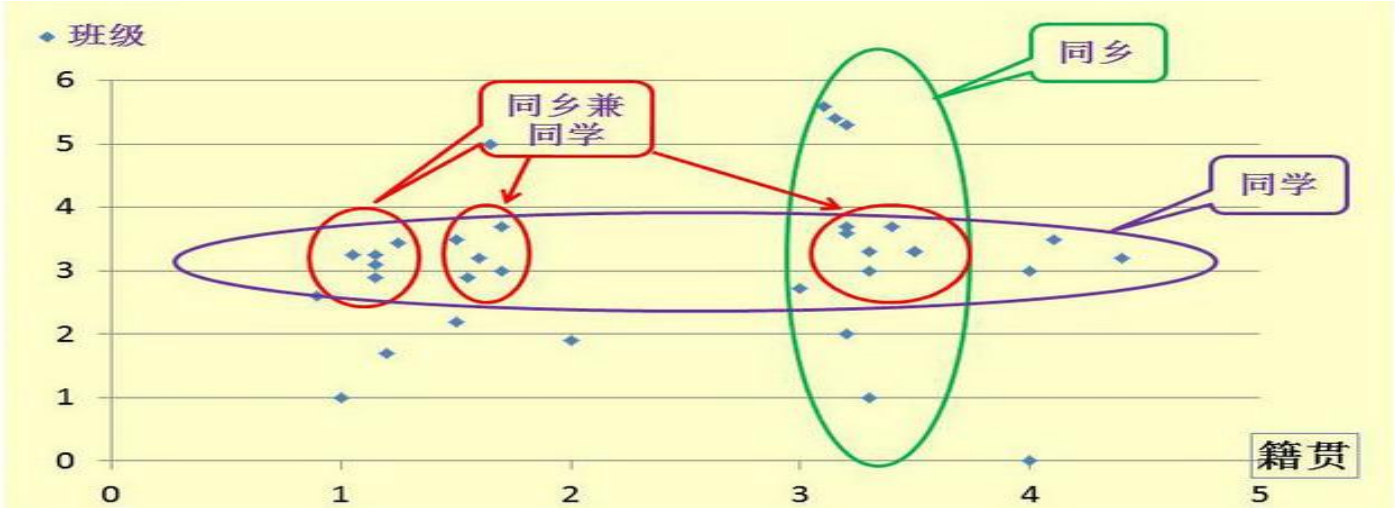
无监督学习的例子

- 尘肺分期自动判读中的肺野分割

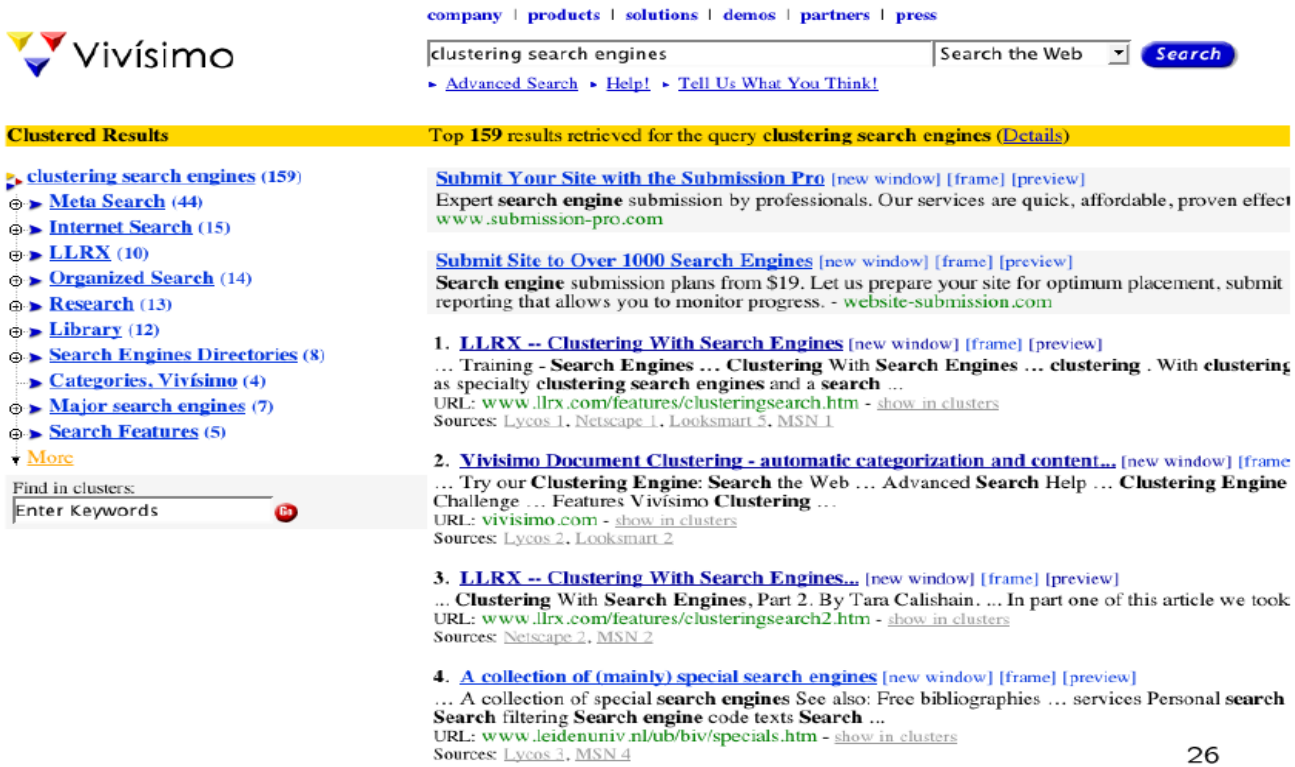


无监督学习的例子

- 客户分割（segmentation）是一种发现用户特性的方法。基于数据内部结构的分割将自然客户分组，从而给你一个客户信息的概况，这可以直接转化为增加客户的经营策略。



无监督学习的例子


Vivísimo
company | products | solutions | demos | partners | press
clustering search engines
Search the Web
Search
Advanced Search | Help! | Tell Us What You Think!
Clustered Results
Top 159 results retrieved for the query clustering search engines (Details)
clustering search engines (159)
Meta Search (44)
Internet Search (15)
LLRX (10)
Organized Search (14)
Research (13)
Library (12)
Search Engines Directories (8)
Categories, Vivísimo (4)
Major search engines (7)
Search Features (5)
More
Find in clusters:
Enter Keywords
Submit Your Site with the Submission Pro [new window] [frame] [preview]
Expert search engine submission by professionals. Our services are quick, affordable, proven effect
www.submission-pro.com
Submit Site to Over 1000 Search Engines [new window] [frame] [preview]
Search engine submission plans from $19. Let us prepare your site for optimum placement, submit reporting that allows you to monitor progress. - website-submission.com
1. LLRX -- Clustering With Search Engines [new window] [frame] [preview]
... Training - Search Engines ... Clustering With Search Engines ... clustering . With clustering as specialty clustering search engines and a search ...
URL: www.llrx.com/features/clusteringsearch.htm - show in clusters
Sources: Lycos 1, Netscape 1, Looksmart 5, MSN 1
2. Vivisimo Document Clustering - automatic categorization and content... [new window] [frame
... Try our Clustering Engine: Search the Web ... Advanced Search Help ... Clustering Engine Challenge ... Features Vivísimo Clustering ...
URL: vivisimo.com - show in clusters
Sources: Lycos 2, Looksmart 2
3. LLRX -- Clustering With Search Engines... [new window] [frame] [preview]
... Clustering With Search Engines, Part 2. By Tara Calishain. ... In part one of this article we took
URL: www.llrx.com/features/clusteringsearch2.htm - show in clusters
Sources: Netscape 2, MSN 2
4. A collection of (mainly) special search engines [new window] [frame] [preview]
... A collection of special search engines See also: Free bibliographies ... services Personal search Search filtering Search engine code texts Search ...
URL: www.leidenuniv.nl/ub/biv/specials.htm - show in clusters
Sources: Lycos 3, MSN 4

无监督学习

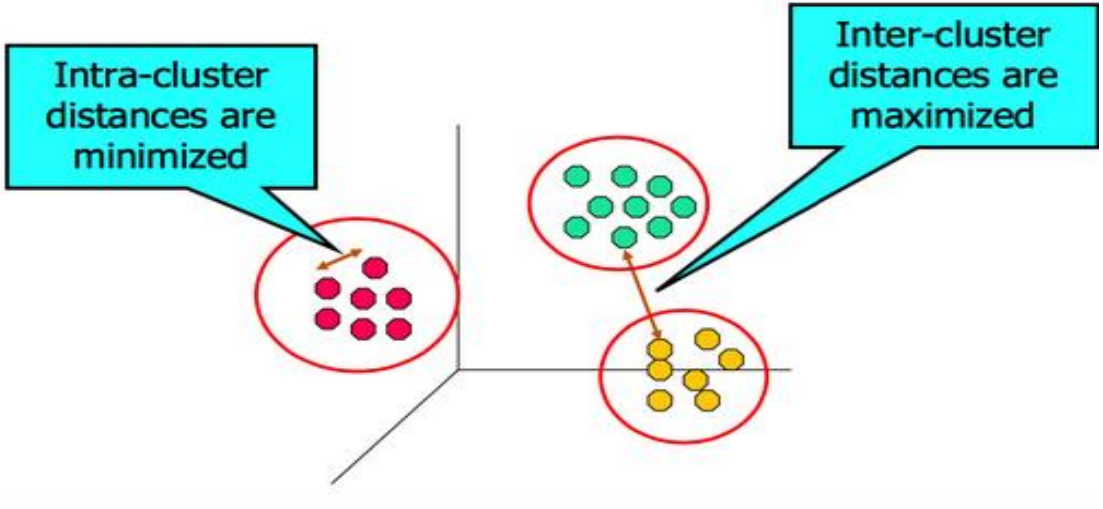
- 下面例子，哪个使用的是无监督算法：
  - 给定邮件的标签（垃圾邮件或者是非垃圾邮件），学习一个垃圾邮件过滤器
  - 给定在网络上的一组新闻，将他们划分成相同的故事的新闻集合
  - 给定一组用户数据，自动地发现市场分割并且将顾客根据市场分割进行划分
  - 给定一组病人（标记了中风或是不中风）集合，学习预测新病人是否中风？

聚类（Clustering）

- 聚类定义：
  - 在一堆的数据中寻找一种”自然分组”(k组)。聚类中的组叫做簇（Cluster）希望同簇的样本较为相似，而不同簇的样本间有明显不同。

什么是一个好的聚类方法?

- 一个好的聚类要具备以下两个特点：
  - 高的簇内相似性（High intra-cluster similarity）
  - 低的簇间相似性 （Low inner-cluster similarity）



相似性(similarity)

相似性通常很难去定义，只能凭主观确定

差异性表示

- 与**相似性**相对应的就是**差异性**(dissimilarity或者说distance)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

数据矩阵

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

差异矩阵

数据类型

- 二元变量（Binary variables）
- 区间标度变量（Interval-scaled variables ）
- 标称型，序数型和比例型变量（Nominal, ordinal and ratio variables ）
- 混合类型变量（Variables of mixed types）

例

对象	Test-1	Test-2	Test-3
1	A	优秀	45
2	B	一般	22
3	C	好	64
4	A	优秀	28

Test-1, Test-2, Test-3都是什么数据类型？？？？

区间标度变量差异度计算

- 数据标准化
  - 计算绝对偏差的平均值:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

其中 $$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf})$$

  - 计算标准度量值 (*z-score*) $$z_{if} = \frac{x_{if} - m_f}{s_f}$$

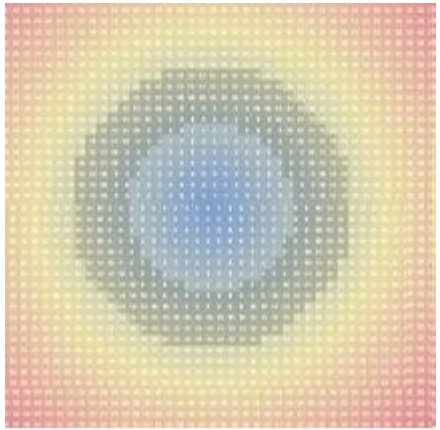
- 计算距离

距离计算公式

- **欧几里德距离**:

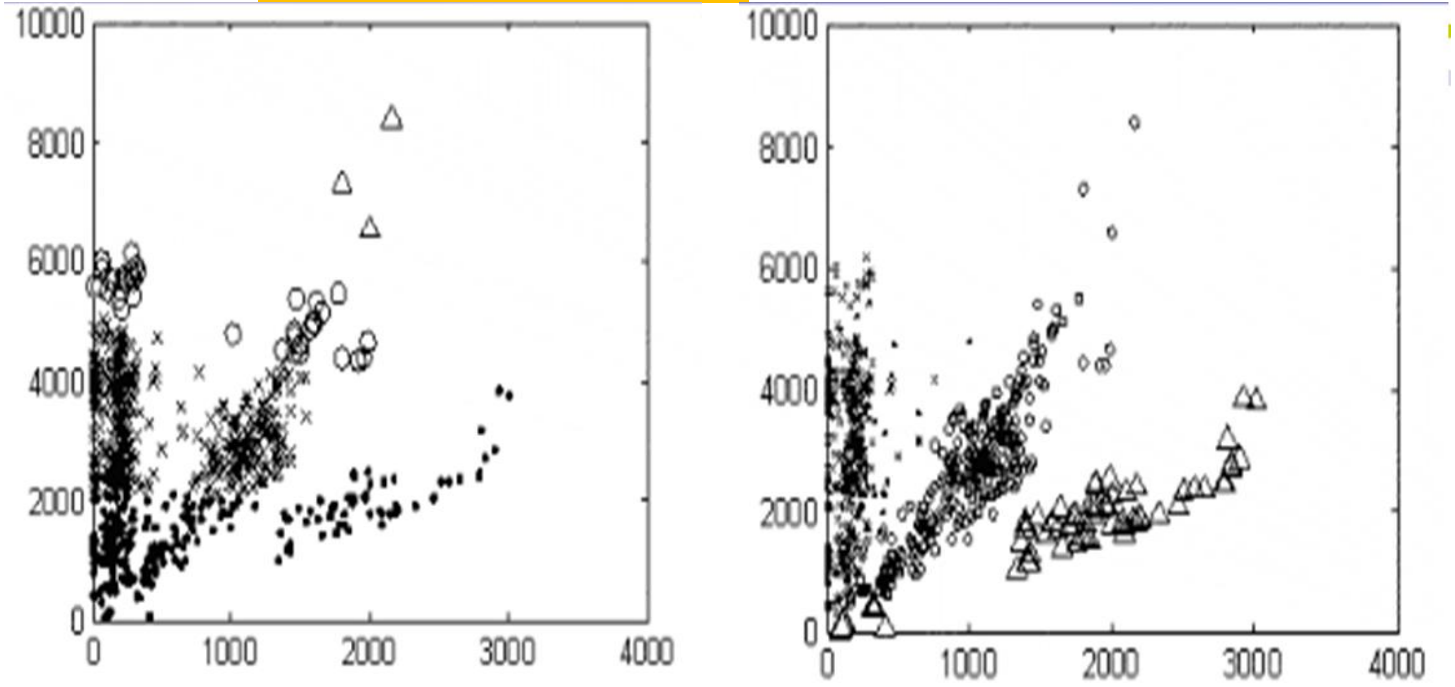
$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

  - 距离函数有如下特性：
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- 可以根据每个变量的重要性赋予一个权重



区间标度变量余弦相似度

$$\mathrm{sim}(x,y)=\frac{xy}{\|x\|\|y\|}$$



余弦相似度

文档	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
文档1	5	0	3	0	2	0	0	2	0	0
文档2	3	0	2	0	1	1	0	1	0	1
文档3	0	7	0	2	1	0	0	3	0	0
文档4	0	1	0	0	1	2	2	0	3	0

$$x \bullet y = 5\times3+0\times0+3\times2+0\times0+2\times1+0\times1+2\times1+0\times0+0\times1=25$$

$$\| x \| = \sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2} = 6.48$$

$$\| y \| = \sqrt{3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2} = 4.12$$

$$sim(x, y) = 0.94$$

二元变量

- 二元变量的可能性表

object i	object j			
		1	0	sum
	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	p

其中每个对象有p个变量，且

$$p=a+b+c+d$$

二元变量

- 对称的

  对称的二元变量，采用**简单匹配系数**来评价两个对象之间的相异度

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

非对称二元变量

- 非对称的

  如果变量的两个状态不是同样重要的，则称该变量是不对称的。

  根据惯例，将比较重要通常也是出现概率比较小的状态编码为1，将另一种状态编码为0。

  对于非对称的二员变量，采用**Jaccard系数**来评价两个对象之间的相异度

$$d(i,j)=\frac{b+c}{a+b+c}$$

非对称二元变量相异度

- 例

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- gender 是一个对称的二元变量，其它的都是非对称的二元变量
- 将值 Y和 P 编码为1, 值 N 编码为 0，非对称二元变量：

	a	b+c	d
(Jack,M)	?	?	?
(J，J)	?	?	?
(M,Jim)	?	?	?

标称变量（Nominal Variables）

- 标称变量是二元变量的推广，它可以具有多于两个的状态，比如 变量map_color可以有 red, yellow, blue, green 四种状态。有两种计算相异度的方法：
- 方法1: **简单匹配方法**
  - m是匹配的数目, *p*是全部变量的数目

$$d(i, j) = \frac{p - m}{p}$$

- 方法2: 使用二元变量
  - 为每一个状态创建一个新的二元变量，可以用非对称的二元变量来编码标称变量。

序数型变量

- 处理的方式与区间标度变量非常相似
  - 用对应的来代替X_{if}
  - 将每个变量的值域映射到[0,1],通过把第f个变量的i个对象用

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

来代替

  - 使用区间标度变量值的方法来计算相异度

思考?？

对象	Test-1	Test-2	Test-3
1	A	优秀	45
2	B	一般	22
3	C	好	64
4	A	优秀	28

优秀：3，好：2，一般：1

$$z_{11} = \frac{3 - 1}{3 - 1} = 1$$

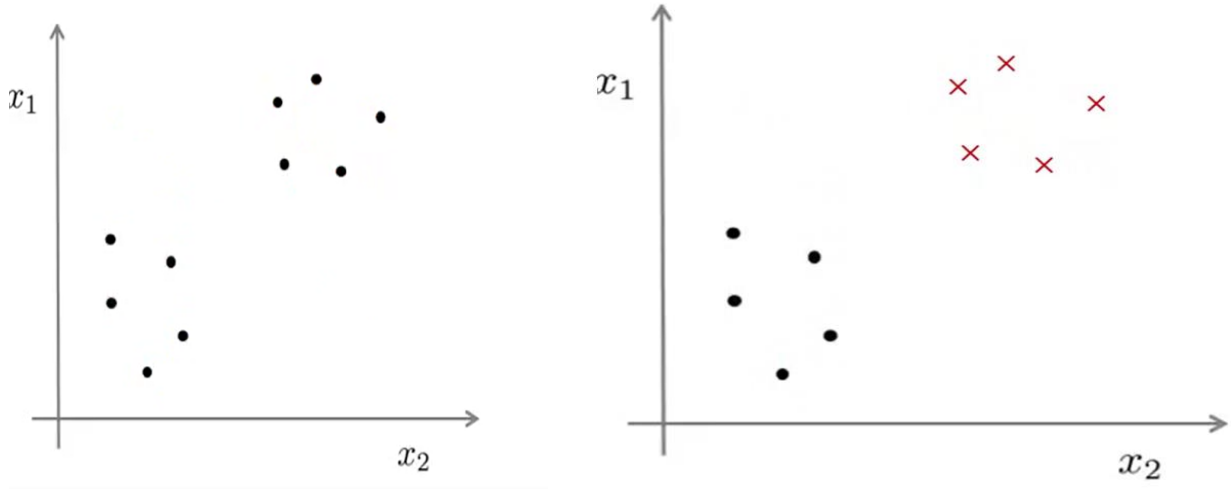
$$z_{21} = \frac{1 - 1}{3 - 1} = 0$$

混合型变量

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}\sum_{ij}^{f}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

- f是二元变量或标称变量
  - $d_{ij}^{(f)}=0$ 如果$x_{if}=x_{jf}$,否则$d_{ij}^{(f)}=1$
- f是数值型变量：使用正常的距离公式
- f是序数型变量
  - 计算秩序r_{if}
  - 并将z_{if}作为数值型变量对待
- 其中$\delta_{ij}^{(f)}=0$,如果x_{if}或x_{jf}缺失（即对象i或j没有属性f的度量值），否则为1

无监督学习



只给计算机训练数据，不给结果（标签），因此计算机无法准确地知道哪些数据具有哪些标签，只能凭借强大的计算能力分析数据的特征，发现数据本身的内部结构特点。

ZIXING AI
自兴人工智能

k-均值(k-means)算法

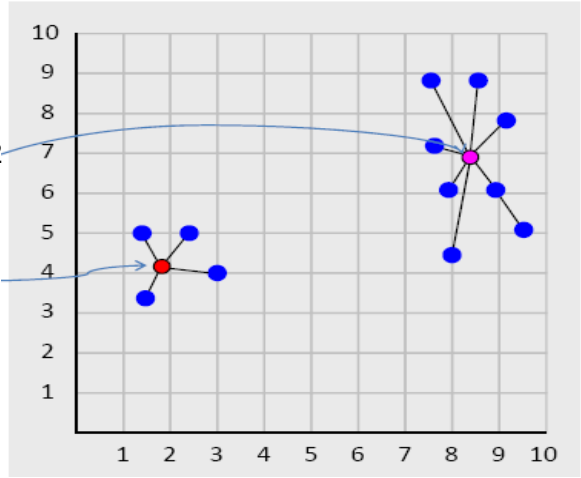
- 1.划分的准则函数
- 2. 算法描述
- 3. 算法的优缺点
- 4. 算法的扩展变形
- 5. 算法的应用

1.划分的准则函数

- **误差平方和(SSE)：**

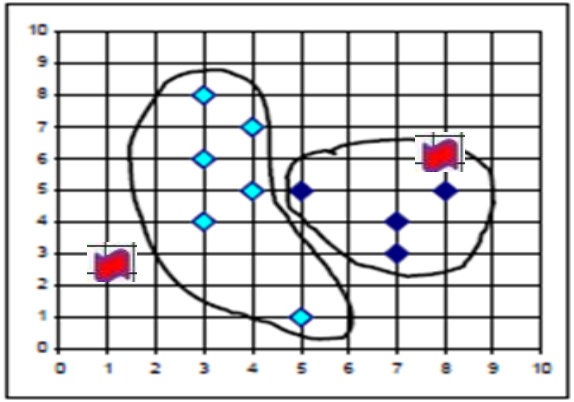
$$J_e = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2$$

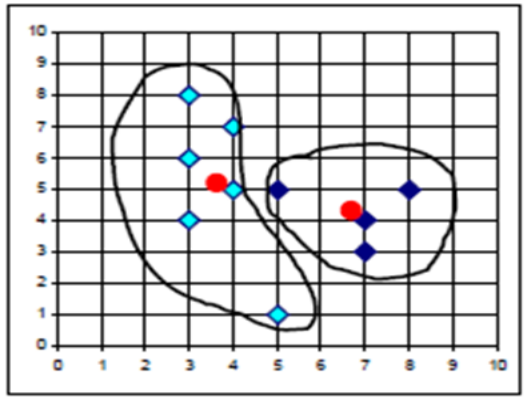
$$m_i = \frac{1}{n_i} \sum_{x \in C_i} x$$

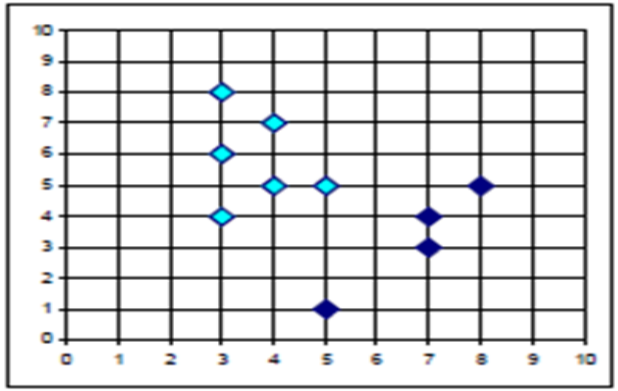


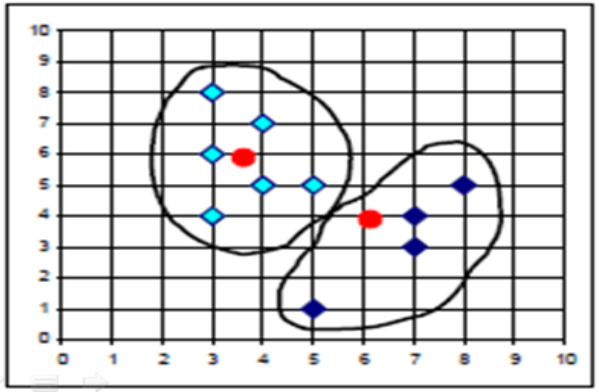
K-均值算法是获得准则函数Je最小的划分

2、算法描述









2.算法描述

- **给定k，算法的处理流程如下:**

  第一步：随机的把所有对象分配到k个非空的簇中；

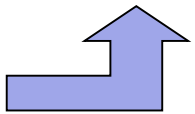
  第二步：计算每个簇的平均值，并用该平均值代表相应的簇中心；

  第三步：将每个对象根据其与各个簇中心的距离，重新分配到与它距离最近的簇中；

  第四步：重复2,3直到k个簇的中心点不再发生变化或准则函数J_e收敛。

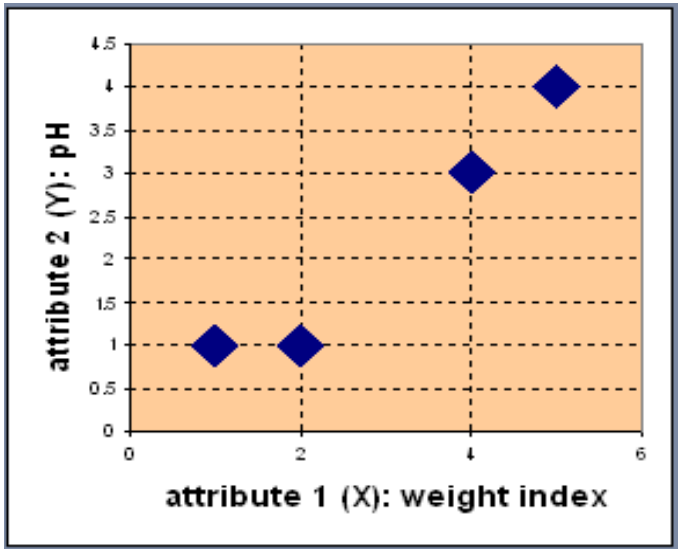
  O (kn)

- **时间复杂度:** O(tkn)



Example :

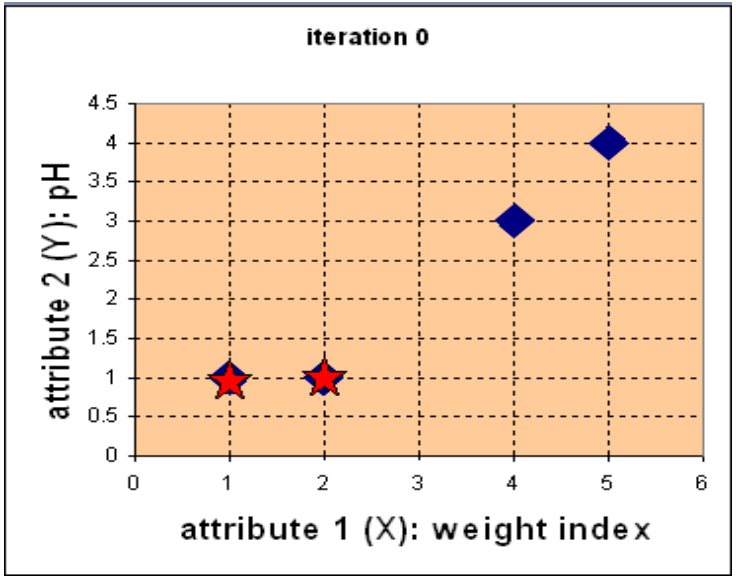
Object	Feature 1 (X): weight index	Feature 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4



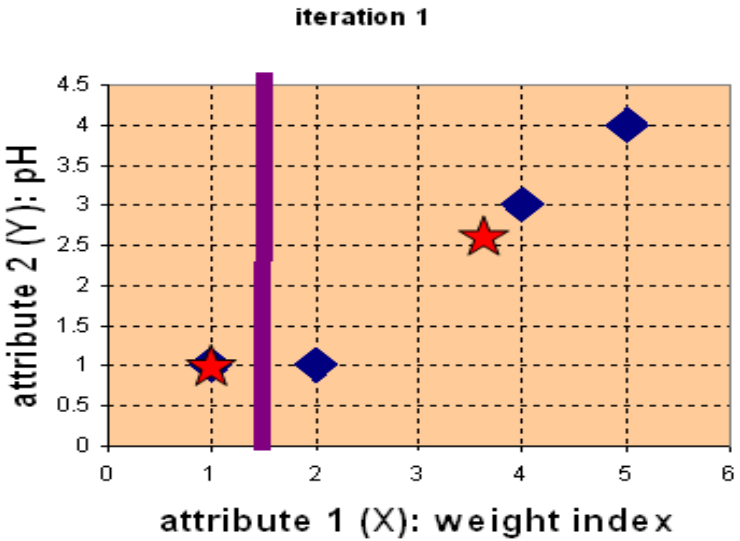
Example

A	B	C	D
1	2	4	5
1	1	3	4

- m_1=(1,1)
- m_2=(2,1)



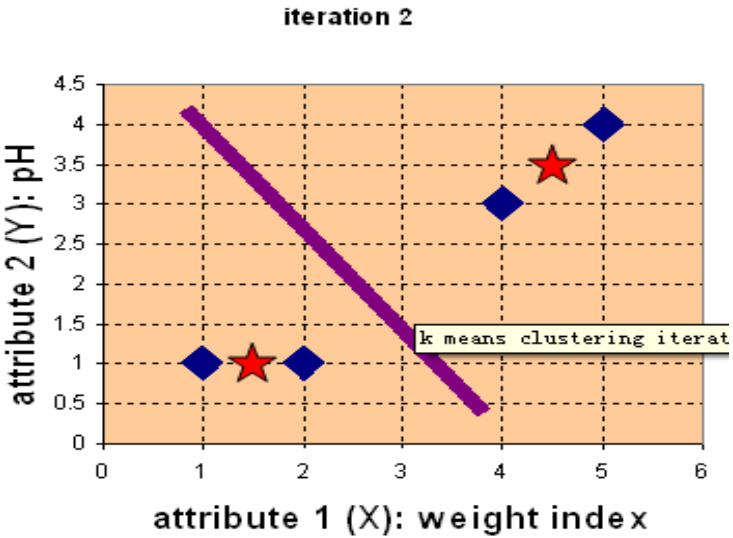
迭代1



$$\begin{array}{cccc} A & B & C & D \\ \begin{pmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{pmatrix} \end{array}$$

$m_1=(1,1)$ $m_2=(11/3,8/3)$

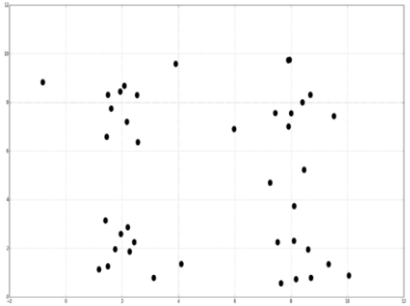
迭代2

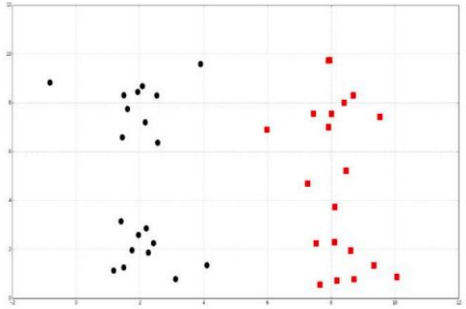
- 

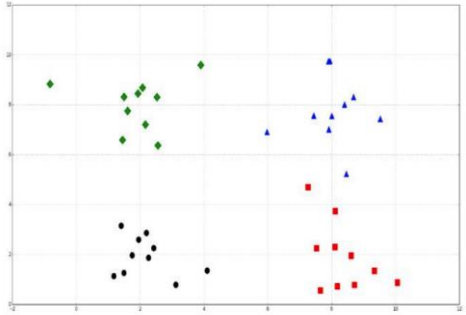
$$\begin{matrix} A & B & C & D \end{matrix}$$
$$\begin{pmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{pmatrix}$$

$m_1=(3/2,1)$ $m_2=(9/2,7/2)$

K值的选择



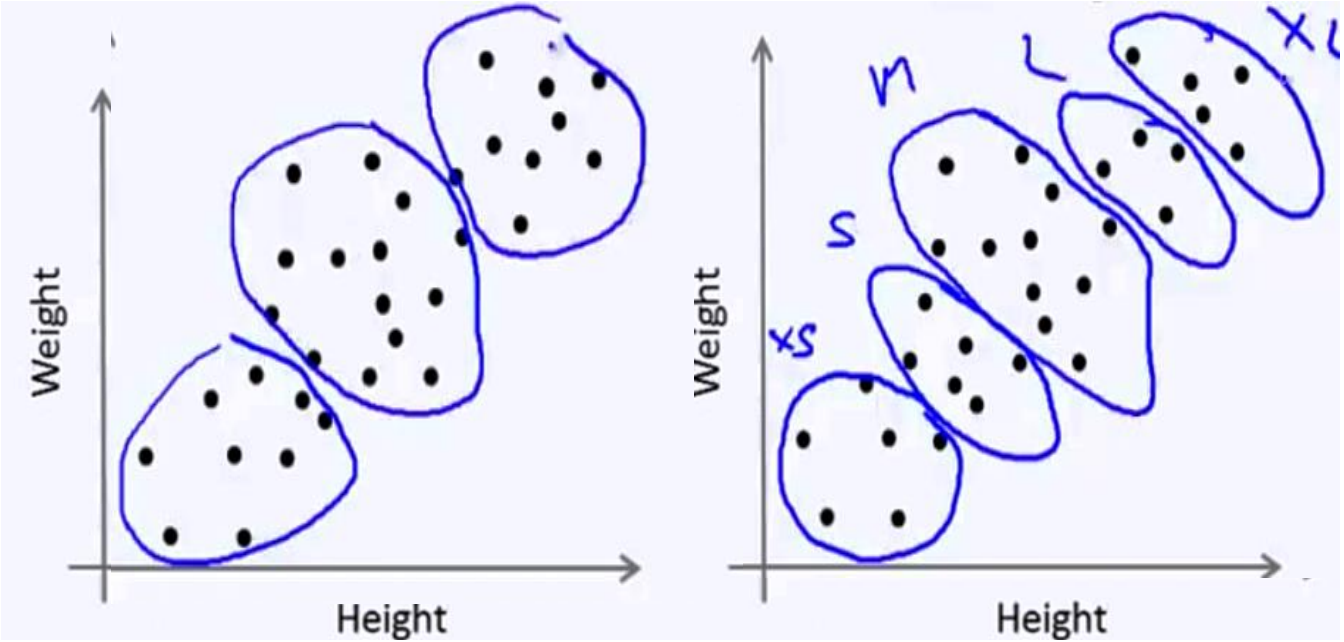




- 根据用需求户定义
- 肘部法则

K值的选择

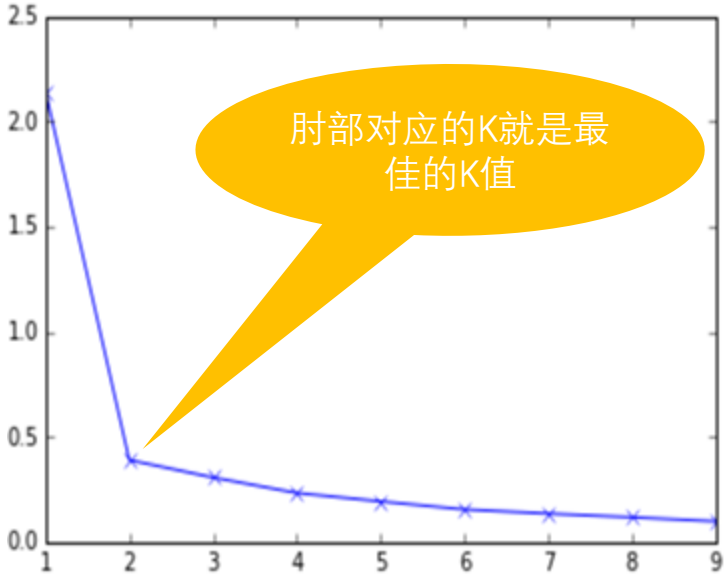
- 根据用需求户定义



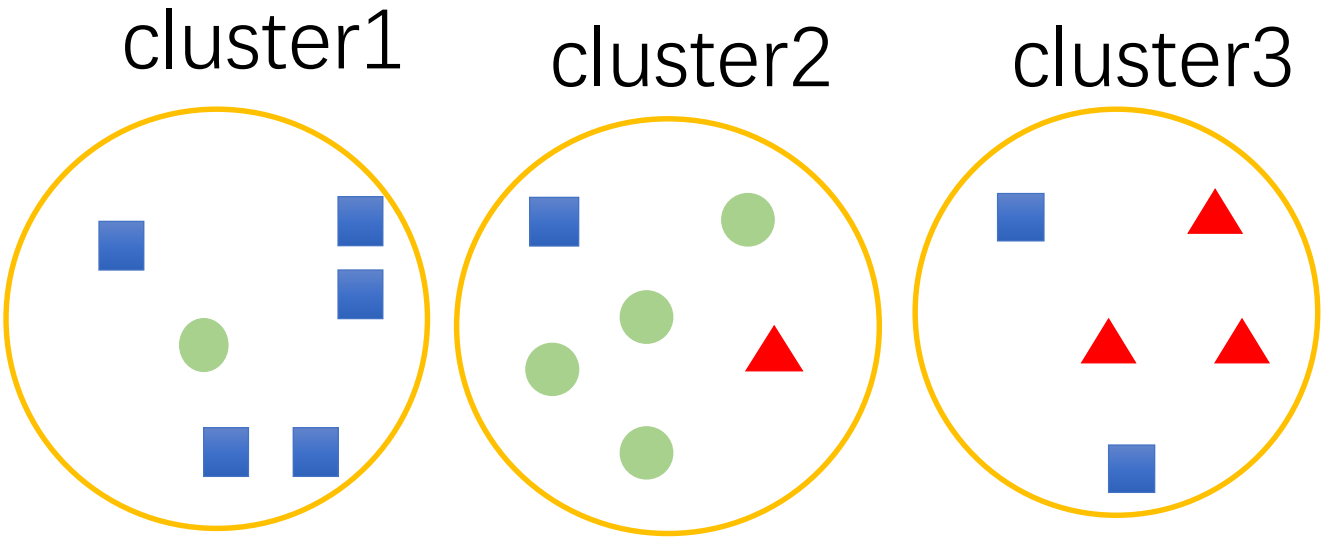
K值的选择

- 肘部法则

把不同K值的成本函数值Je画出来。随着K值的增大，Je会减小；每个类包含的样本数会减少，于是样本离其重心会更近。但是，随着K值继续增大，Je的的变化趋缓。K值增大过程中，Je下降幅度最大的K位置对应的K值就是肘部。



评价指标



评价指标

- Purity

$$\text{purity}(W, C) = \frac{1}{N}\sum_{k}\max_{j}|w_k \cap c_j|$$

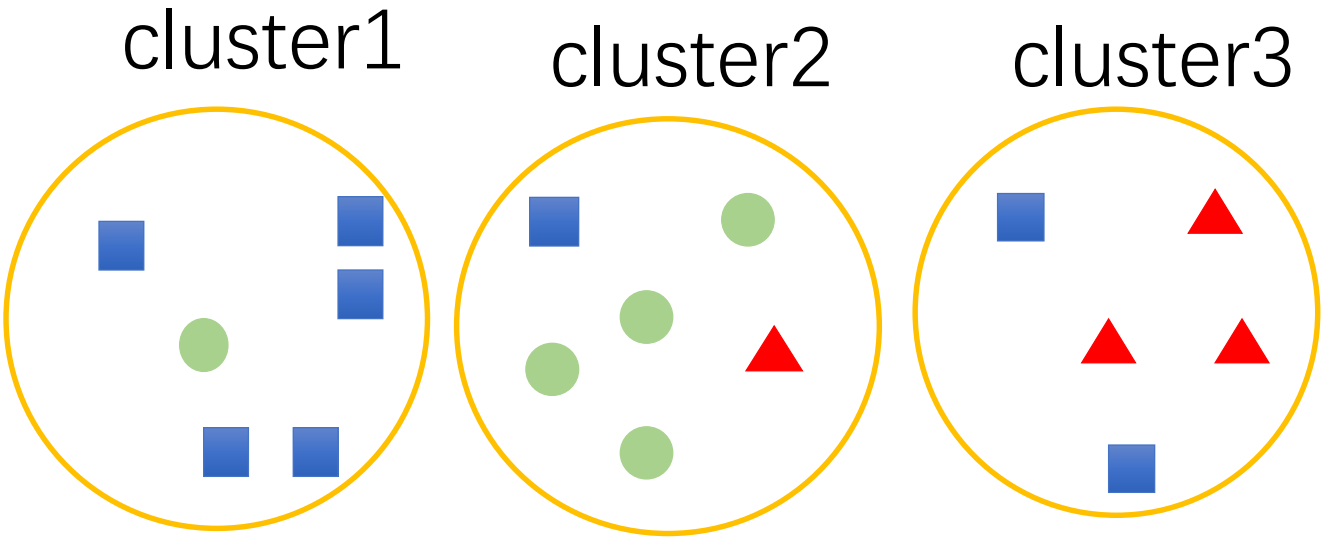
- 优点：方便计算，值在0～1之间，完全错误的聚类方法值为0，完全正确的方法值为1。
- 缺点：无法对退化的聚类方法给出正确的评价，设想如果聚类算法把每篇文档单独聚成一类，那么purity值为？？？

- RI：一种用排列组合原理来对聚类进行评价的手段

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

其中：TP指被聚在一类的文档被正确分类了，TN是不应该聚在一类的文档被正确分开了，FP指不应该放在一类的文档被错误的放在一类，FN只不应该分开的文档被错误地分开了

评价指标



F-measure

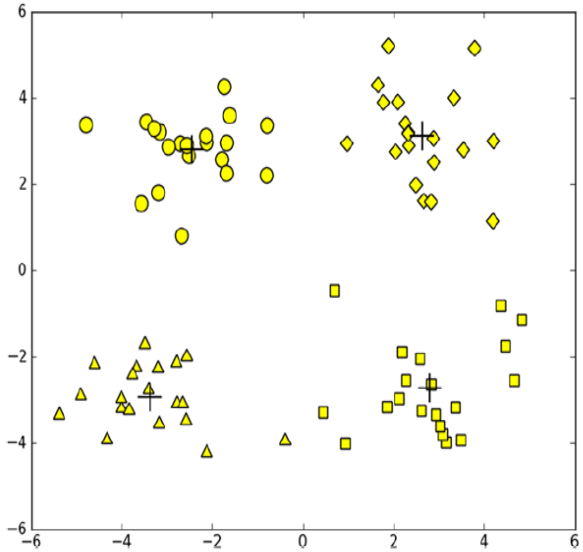
- 基于上述RI方法衍生出的一个方法

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F_\beta = \frac{(\beta^2)\ PR}{\beta^2 P + R}$$

- RI方法有个特点就是把准确率和召回率看得同等重要。有时候我们可能需要某一特性更多一点，这时候就适合F值方法

K均值实战

- 主要子函数
  - 读入数据： loadDataSet(filename)
  - 随机选择簇中心: randCent(dataSet,k)
  - 计算每个点间的距离: distElucd(vecA.vecB)



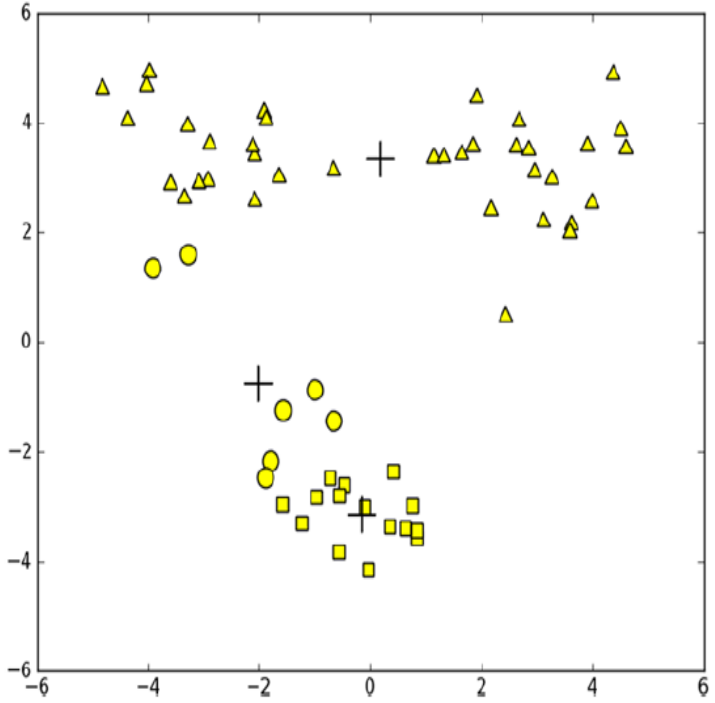
- kmeans函数：
  - myCentroids, clustAssing = kMeans(dataSet,k,distMeans,createCent= randCent)

» datMat=mat(kMeans.loadDataSet(' testSet.txt'))

» myCentroids, clustAssing = kMeans.kMeans(datMat,4)

算法的优缺点

- 优点：容易实现。
- 缺点：
  - 受初始中心点影响，可能收敛局部最优
  - 适用数据类型：数值型数据



k-modes算法

- k-modes算法把k-means算法扩展到可分类数据
- k-modes算法中的中心定义：
  - 根据可分类属性值出现的频率更新聚类中心，聚类中出现频率最高的属性值被选为聚类中心，即modes（类模式）。

{[a,1] [a,2] [b,1] [a,1] [c,3]})类模式为？

k-modes算法

- k-modes算法的距离计算方法：
  - 假设***X***，***Y***是数据集中的两个对象，它们用*m*维属性描述，则这两个对象之间的相异度为：

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j)$$

$$当x_j=y_j,\delta(x_j,y_j)=0;当x_j\neq y_j,\delta(x_j,y_j)=1$$

[a,1] 与[c,3] 的距离d是 ？

实验数据

- 实验数据: UCI的iris数据集
  - iris数据集通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性。

数据集特征	多变量	记录数	150
属性特征	实数	属性数目	4
领域	生活		

实验任务

ZIXING AI
自兴人工智能

- 用传统的kmeans算法对iris数据集进行聚类，将K=3，随机执行3次，输出3次的结果，结果包括：
  - 输出3个簇中心，
  - 统计每个簇内的样本数
  - 比较三次的SSE
- 用传统的kmeans算法对iris数据集进行聚类，将K=2,3,4,5,6,输出结果包括：
  - 输出每次的SSE
  - 将多次的SSE画折线图
  - 选出最佳的K值
- 写出实验报告