机器学习导论

第10章 聚类

谢茂强

南开大学软件学院

本章部分内容来自于吴恩达Coursera网课

目录

目录

聚类：一种无监督学习算法

聚类目标: 对数据集进行聚合,使形成一些簇（类），用以了解数据（分布）的内在性质及规律

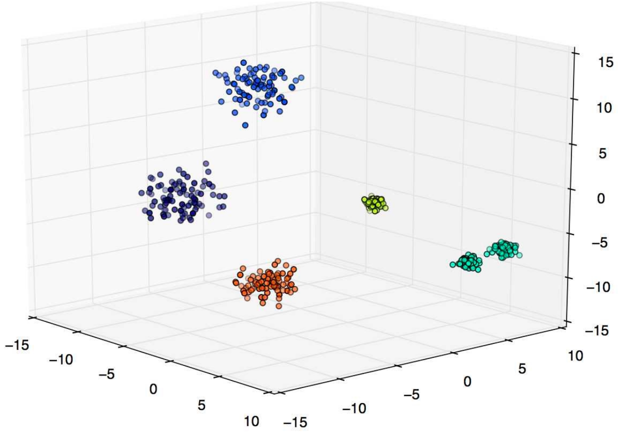
聚类原则: 物以类聚

聚类结果: 相似的归为一类

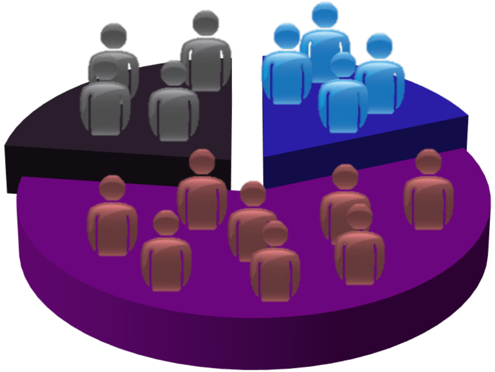
地　　位: 了解数据分布的重要工具，是后续分析的重要基础

示　　例:

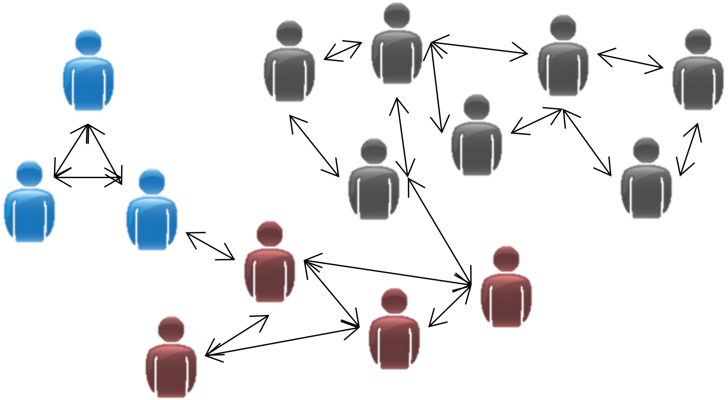
根据客户的特征进行类别划分



聚类的应用

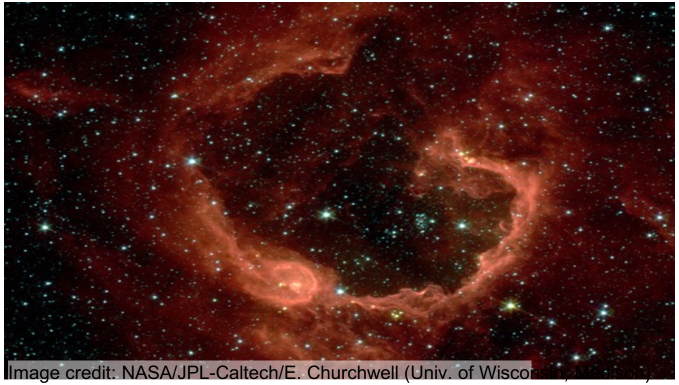


Market segmentation



Social network analysis

Organize computing clusters

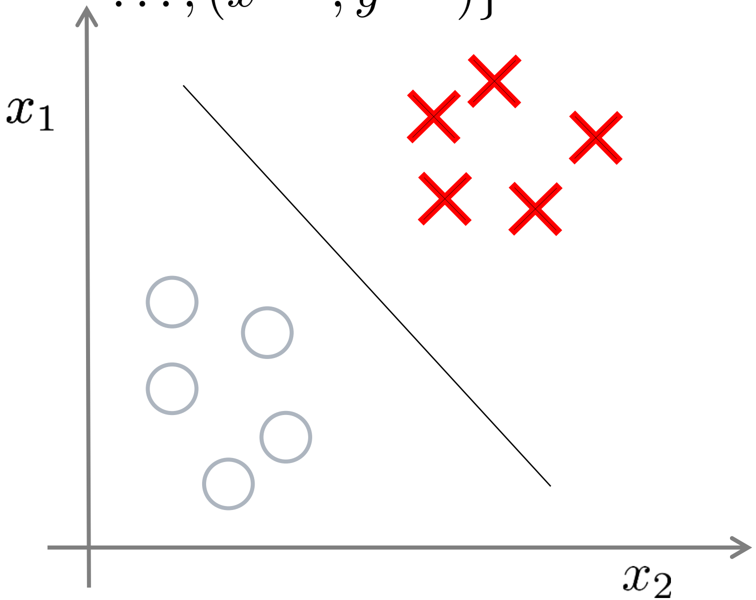


Astronomical data analysis

Supervised learning v.s. Unsupervised learning

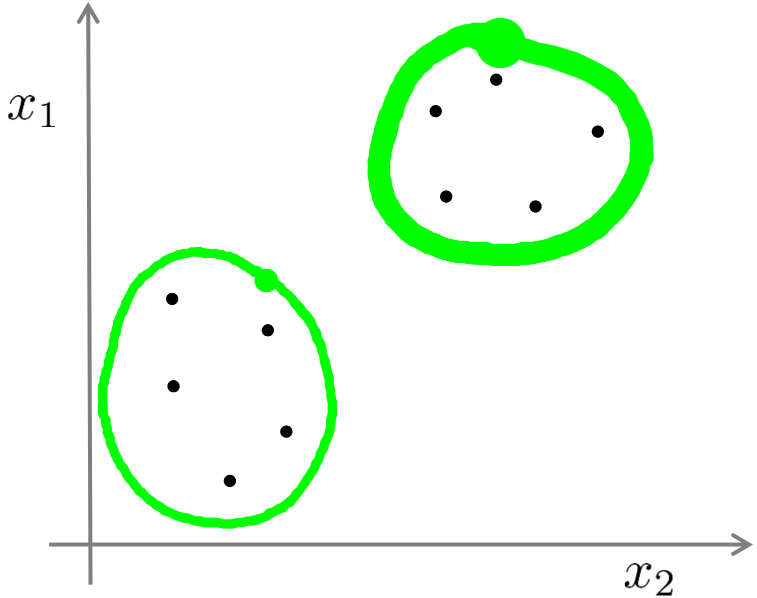
监督学习:

$$\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),(x^{(3)},y^{(3)}),\ldots,(x^{(m)},y^{(m)})\}$$



非监督学习:

$$\{x^{(1)},x^{(2)},x^{(3)},\ldots,x^{(m)}\}$$



1. 聚类的输入输出定义

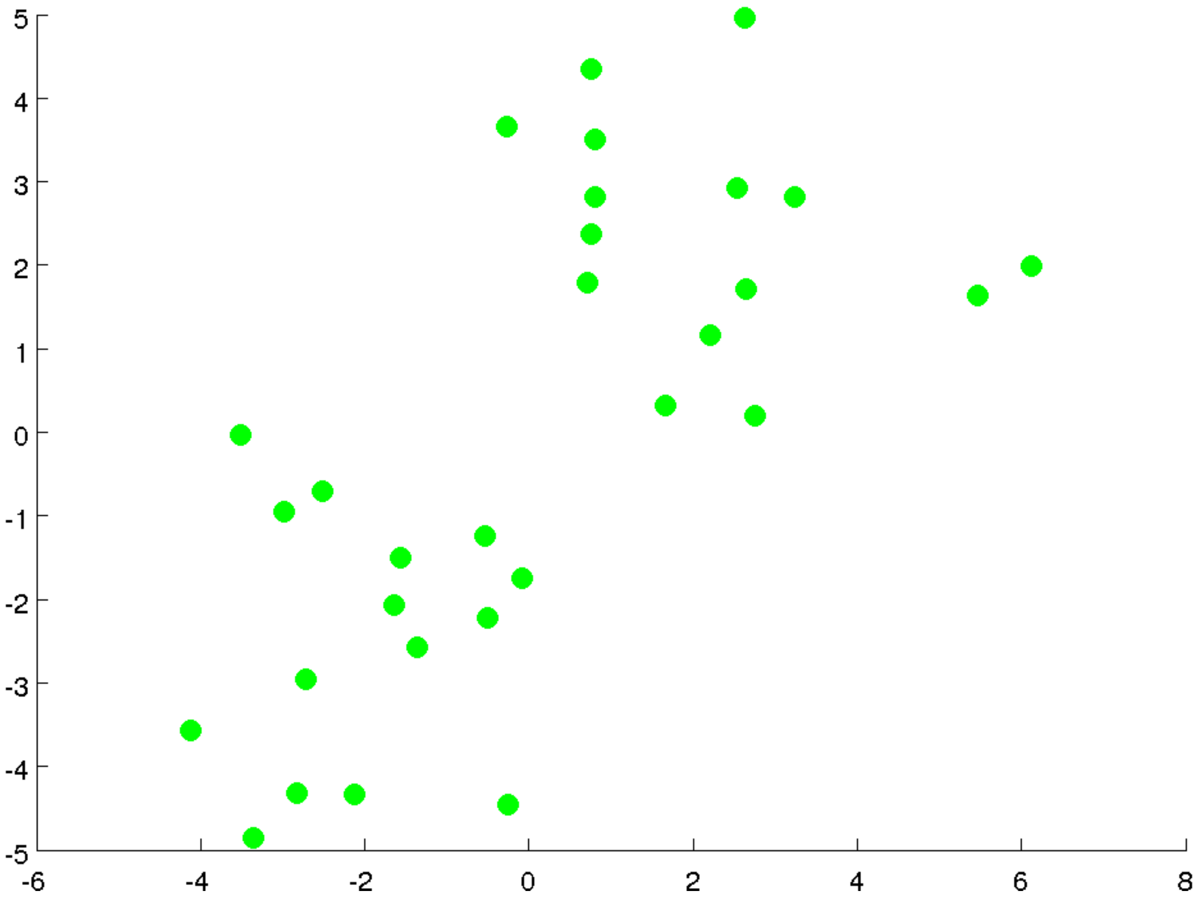
输入： $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \cdots; x_{in})$

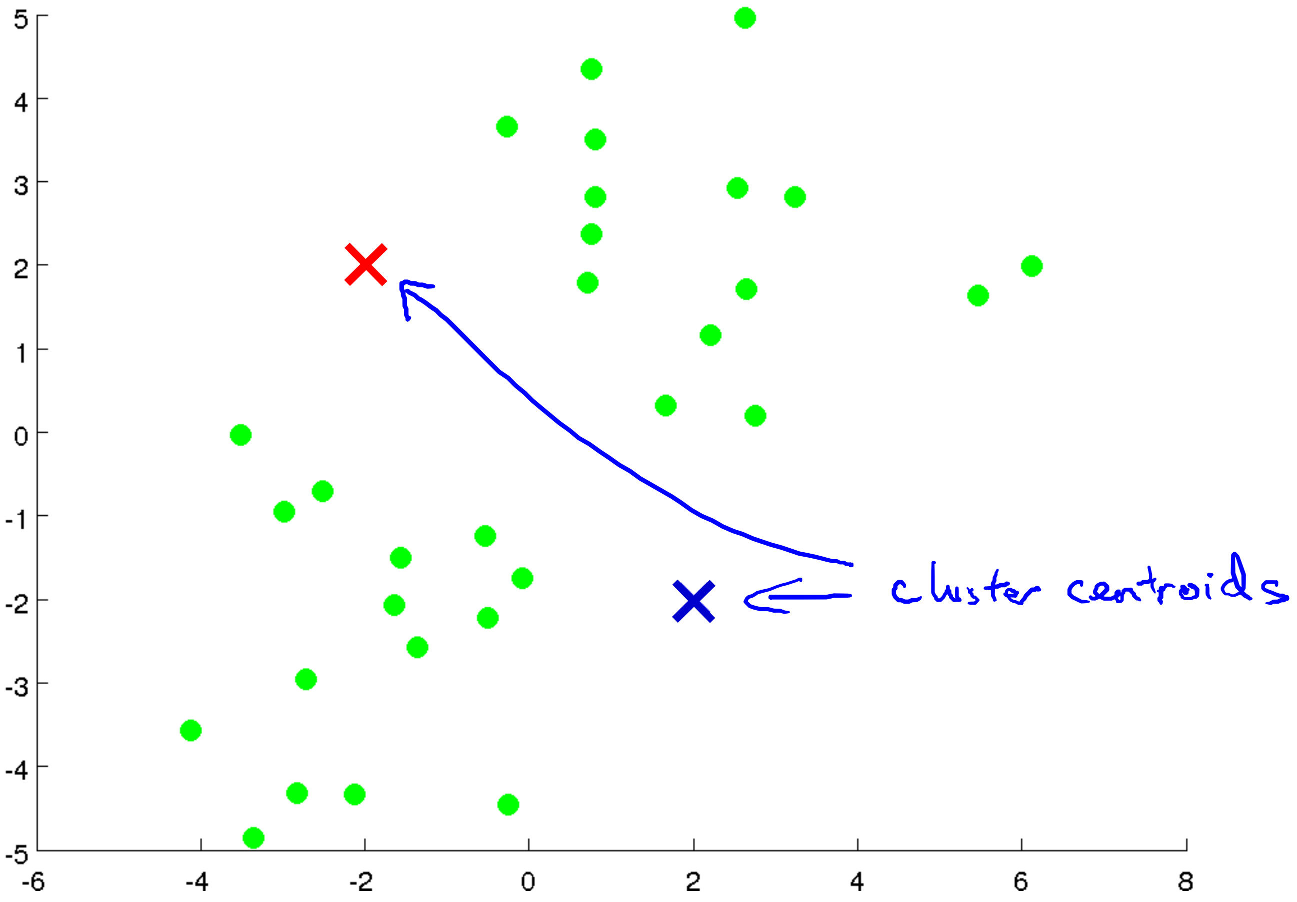
输出： $\boldsymbol{x} \in C_l$ 簇集合： $\{C_l \mid l = 1, 2, \cdots, k\}$

$C_{l'} \cap_{l' \neq l} C_l = \varnothing$ $D = \cup_{l=1}^{k} C_l$

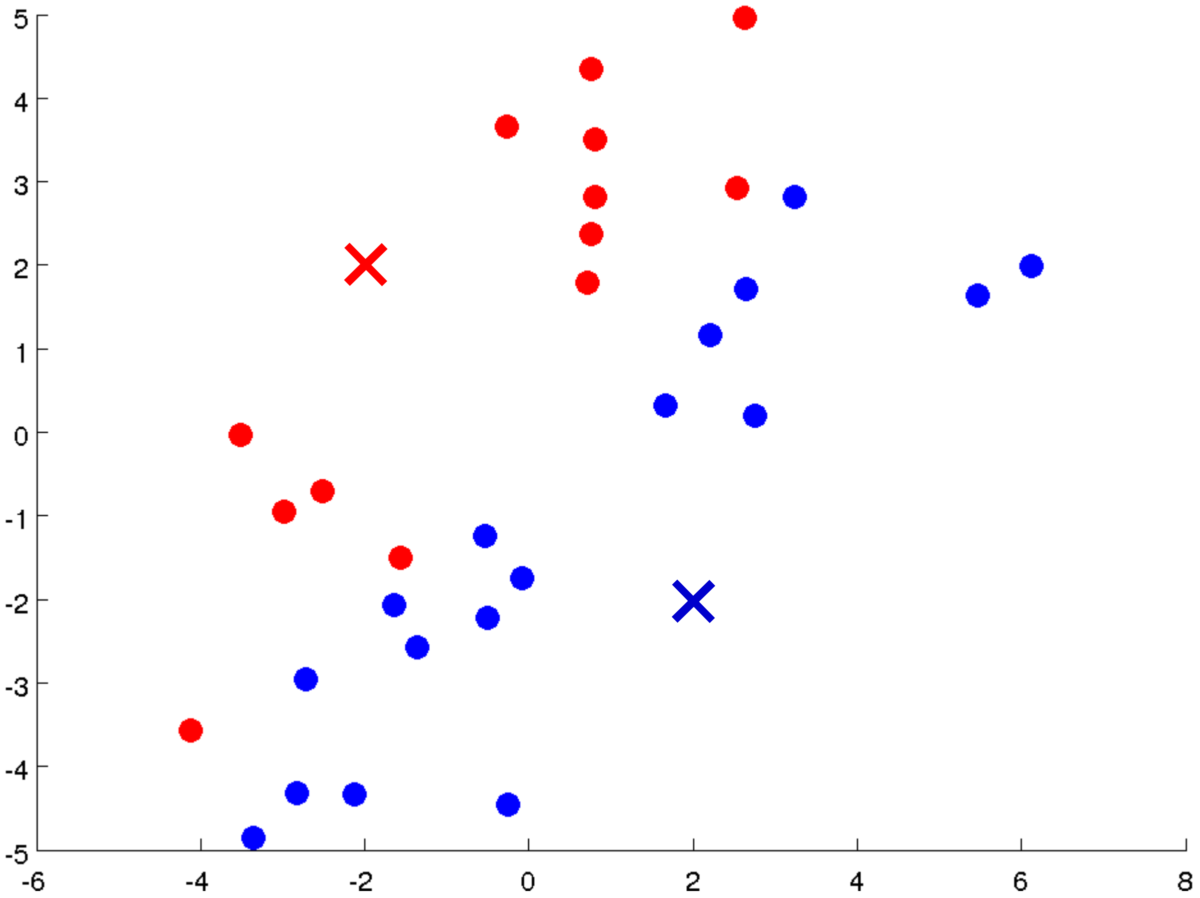
聚类中一种算法思路的示意（下页开始）

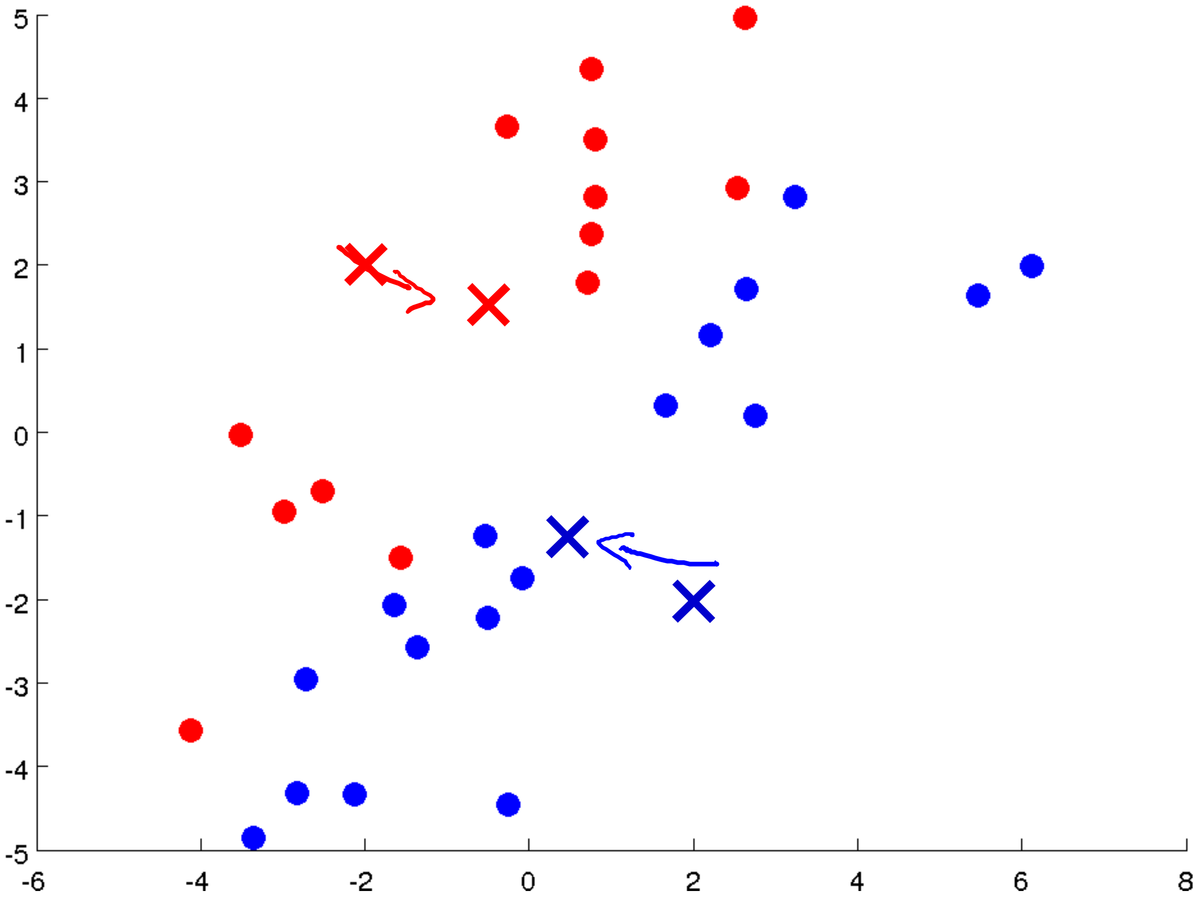
周志华 .《机器学习》 Chpt. 9. P197


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

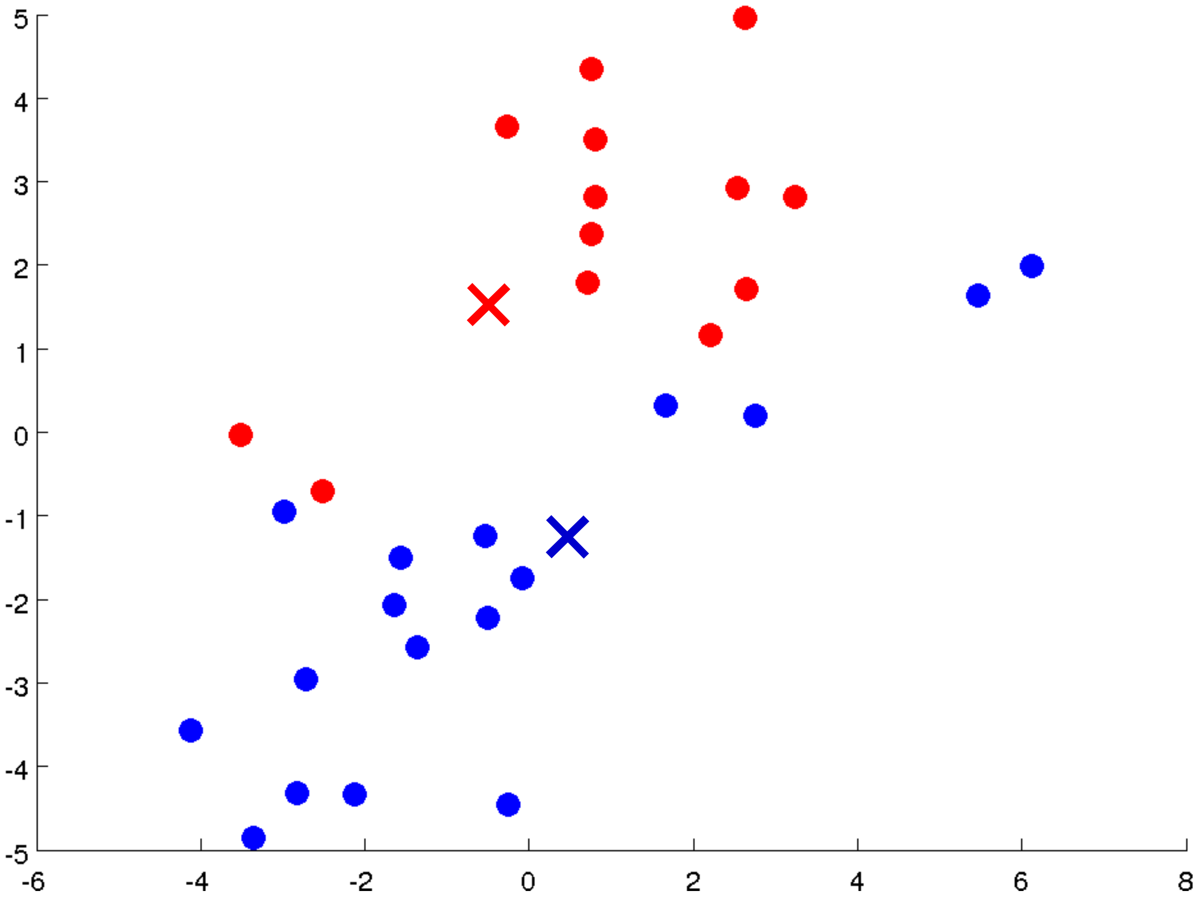


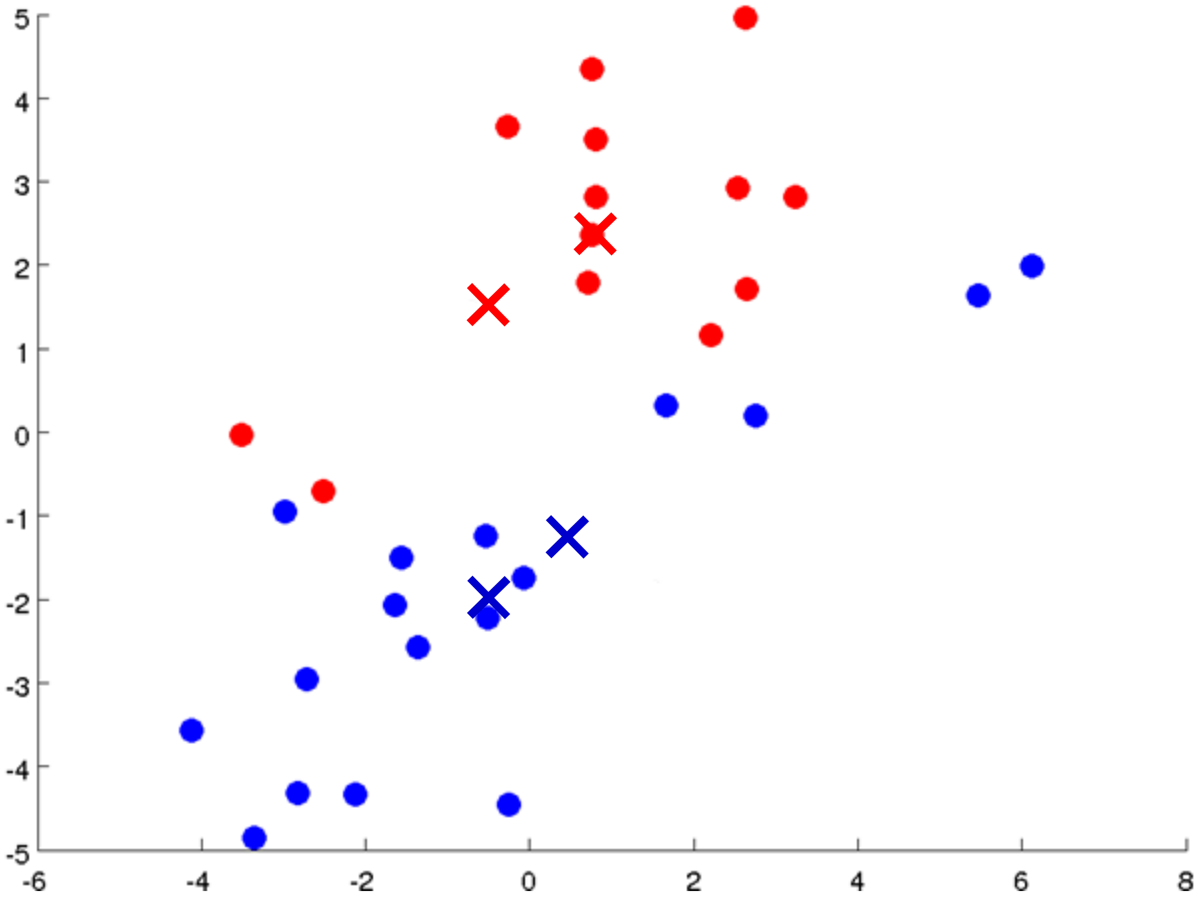
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8
cluster centroids

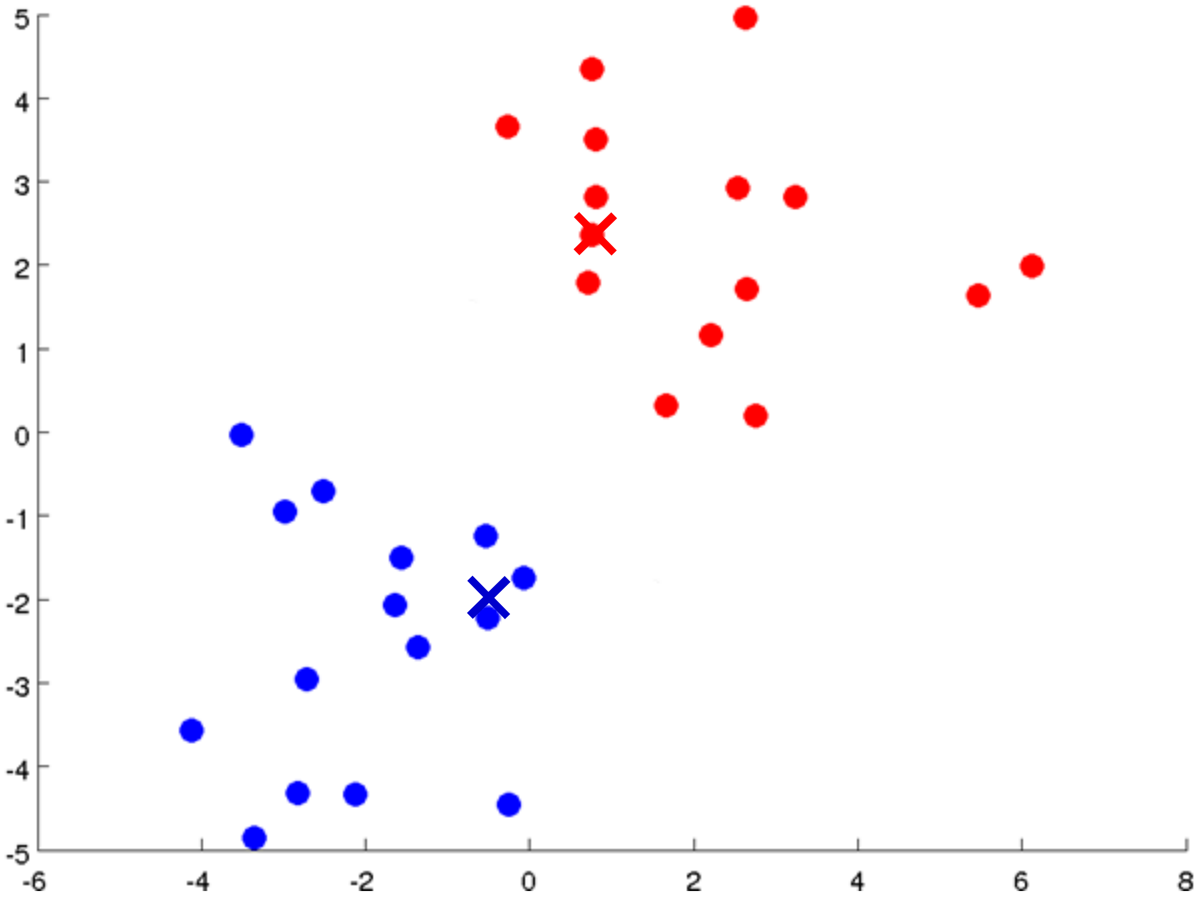

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

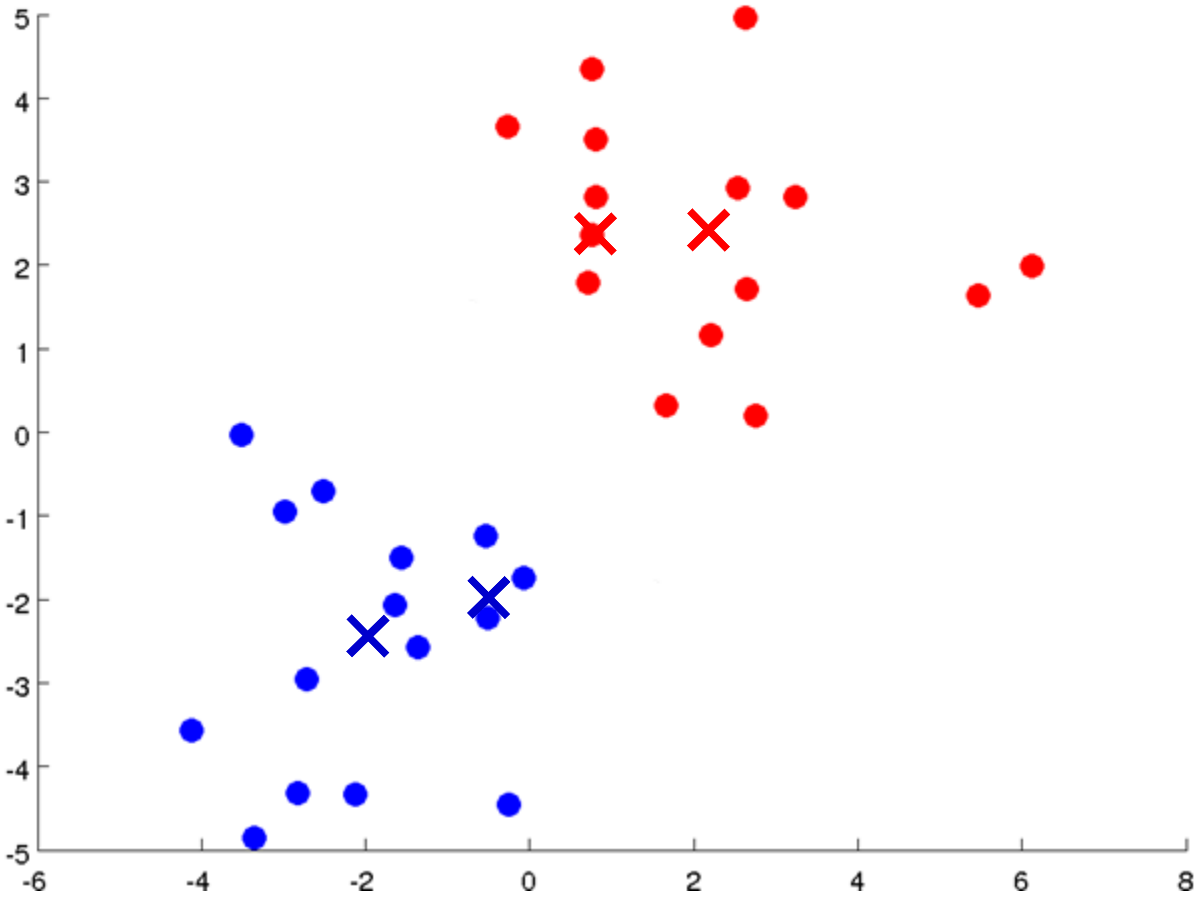


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

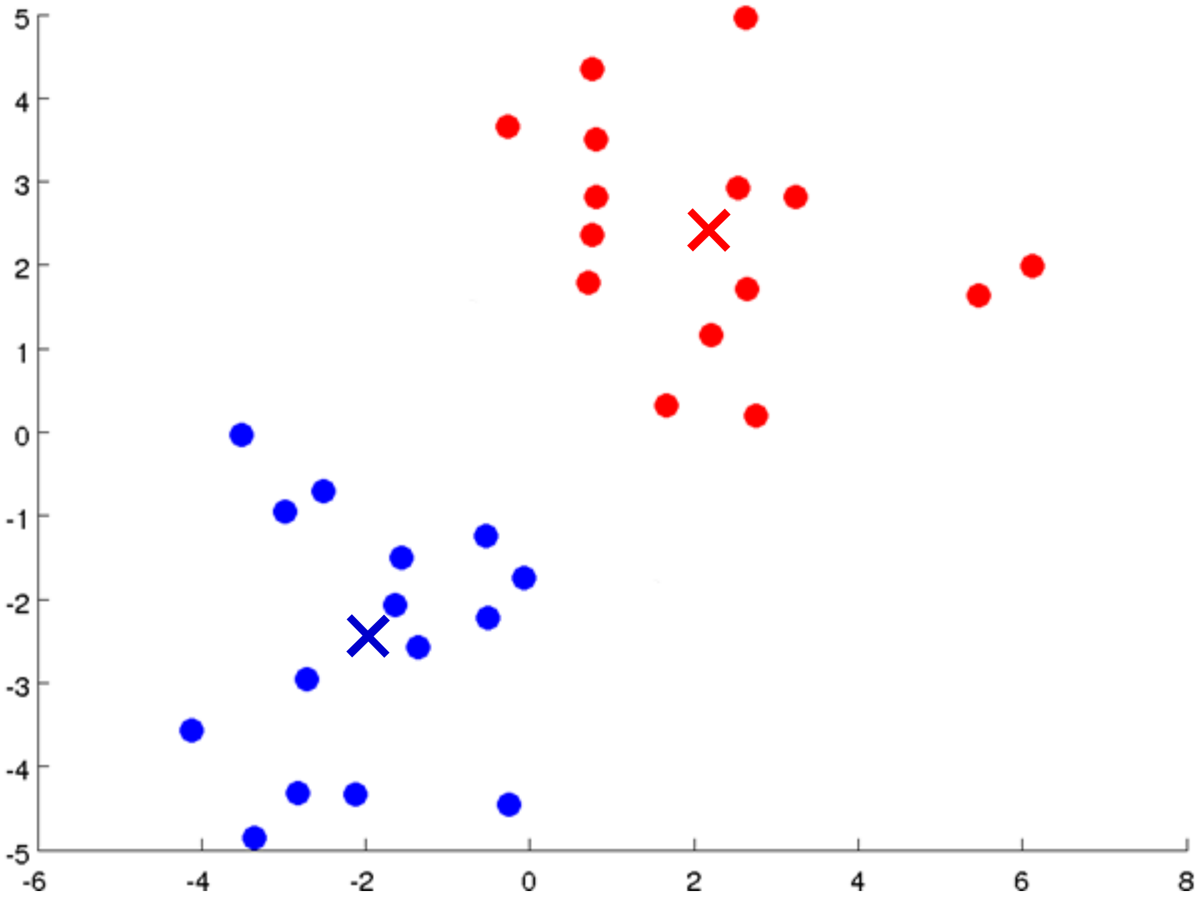

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

- 聚类性能指标（有效性指标，validity index），用来评价结果，进而作为优化目标。（物以类聚）

外部指标（External Index）

与某个“参考模型”进行比较

比如专家标注的簇标签

内部指标（Internal Index）

簇内相似度（intra-cluster sim）高

簇间相似度（inter-cluster sim）低

类内距、方差 …

2.1 聚类性能的外部指标（从“样本对”归属的角度)

对数据集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$, 假定通过聚类给出的簇划分为 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, 参考模型给出的簇划分为 $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_s^*\}$. 相应地, 令 $\boldsymbol{\lambda}$ 与 $\boldsymbol{\lambda}^*$ 分别表示与 $\mathcal{C}$ 和 $\mathcal{C}^*$ 对应的簇标记向量. 我们将样本两两配对考虑, 定义

$$a = |SS|, \quad SS = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j)\}, \tag{9.1}$$

$$b = |SD|, \quad SD = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \lambda_i = \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j)\}, \tag{9.2}$$

$$c = |DS|, \quad DS = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \lambda_i \neq \lambda_j, \lambda_i^* = \lambda_j^*, i < j)\}, \tag{9.3}$$

$$d = |DD|, \quad DD = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j)\}, \tag{9.4}$$

其中集合 SS 包含了在 $\mathcal{C}$ 中隶属于相同簇且在 $\mathcal{C}^*$ 中也隶属于相同簇的样本对, 集合 SD 包含了在 $\mathcal{C}$ 中隶属于相同簇但在 $\mathcal{C}^*$ 中隶属于不同簇的样本对, ……由于每个样本对 $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ $(i < j)$ 仅能出现在一个集合中, 因此有 $a + b + c + d = m(m-1)/2$ 成立.

2.1聚类性能的外部指标

- Jaccard系数 (Jaccard Index, Jaccard Similarity Coefficient)

$$JC = \frac{a}{a+b+c}$$

- FM指数（Fowlkers & Mallows Index, FMI）

$$FMI = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

- Rand指数（Rand Index, RI）

$$RI = \frac{2(a+d)}{m(m-1)}$$

		Groundtruth	
		同簇	异簇
Prediction	同簇	SS(a)	SD(b)
	异簇	DS(c)	DD(d)

2.2 聚类性能的内部指标

- 簇C内样本间平均距离

$$\operatorname{avg}(C)=\frac{2}{|C|(|C|-1)}\sum_{1\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)\ , \tag{9.8}$$

- 簇C内样本间最远距离

$$\operatorname{diam}(C)=\max_{1\leqslant i<j\leqslant|C|}\operatorname{dist}(\boldsymbol{x}_i,\boldsymbol{x}_j)\ , \tag{9.9}$$

2.2 聚类性能的内部指标

- 簇间距离

$$d_{\min}(C_i, C_j) = \min_{\boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j} \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) , \tag{9.10}$$

$$d_{\text{cen}}(C_i, C_j) = \operatorname{dist}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) , \tag{9.11}$$

- DB指数(Davies-Bouldin Index, DBI)

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\operatorname{avg}(C_i) + \operatorname{avg}(C_j)}{d_{\text{cen}}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)} \right) . \tag{9.12}$$

DBI越小聚类性能越好

2.2 聚类性能的内部指标

- 簇间距离

$$d_{\min}(C_i, C_j) = \min_{\boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j} \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \ , \tag{9.10}$$

$$d_{\text{cen}}(C_i, C_j) = \operatorname{dist}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \ , \tag{9.11}$$

- Dunn指数(Dunn Index, DI)

$$\mathrm{DI} = \min_{1 \leqslant i \leqslant k} \left\{ \min_{j \neq i} \left(\frac{d_{\min}(C_i, C_j)}{\max_{1 \leqslant l \leqslant k} \operatorname{diam}(C_l)} \right) \right\} . \tag{9.13}$$

DI越大聚类性能越好

3. 距离度量

- 非负性： $\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geqslant 0$
- 同一性： $\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ 当且仅当 $\boldsymbol{x}_i = \boldsymbol{x}_j$
- 对称性： $\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \operatorname{dist}(\boldsymbol{x}_j, \boldsymbol{x}_i)$
- 三角不等式： $\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_k) + \operatorname{dist}(\boldsymbol{x}_k, \boldsymbol{x}_j)$

3. 距离度量

- 明氏距离(Minkowski distance, L_p 范数)：

$$\mathrm{dist}_{\mathrm{mk}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p \right)^{\frac{1}{p}}$$

- 欧氏距离(Euclidean distance)：

$$\mathrm{dist}_{\mathrm{ed}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 = \sqrt{\sum_{u=1}^{n} |x_{iu} - x_{ju}|^2}$$

- 曼哈顿距离(Manhattan/City block distance)：

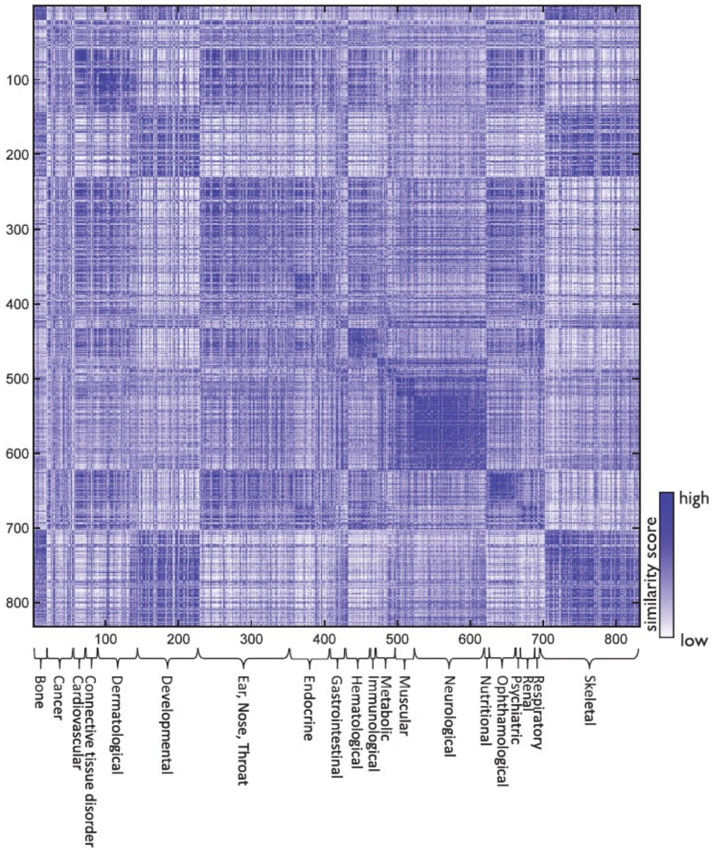
$$\mathrm{dist}_{\mathrm{man}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_1 = \sum_{u=1}^{n} |x_{iu} - x_{ju}|$$

3. 距离和空间的关系

- 距离是定义在空间定义之上，是空间的附属
- 常见的空间问题
  - 尺度伸缩
  - 球面距离
  - 流形（manifold）

距离、相似度

示例：可以使用Jaccard Index 作为语义相似度度量



K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

for k = 1 to K

μ_k := average (mean) of points assigned to k cluster

}

4. K-Means算法的优化目标

给定样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$，“$k$ 均值”(k-means)算法针对聚类所得簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$ 最小化平方误差

$$E=\sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i}||\boldsymbol{x}-\boldsymbol{\mu}_i||_2^2\ ,\tag{9.24}$$

其中 $\boldsymbol{\mu}_i=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$ 是簇 C_i 的均值向量. 直观来看, 式(9.24) 在一定程度上刻画了簇内样本围绕簇均值向量的紧密程度, E 值越小则簇内样本相似度越高.

簇中心的随机初始化

$K=2$

Should have $K < m$

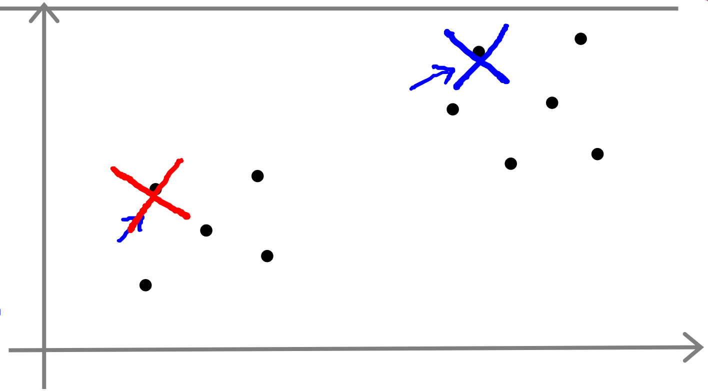
Randomly pick K training examples.

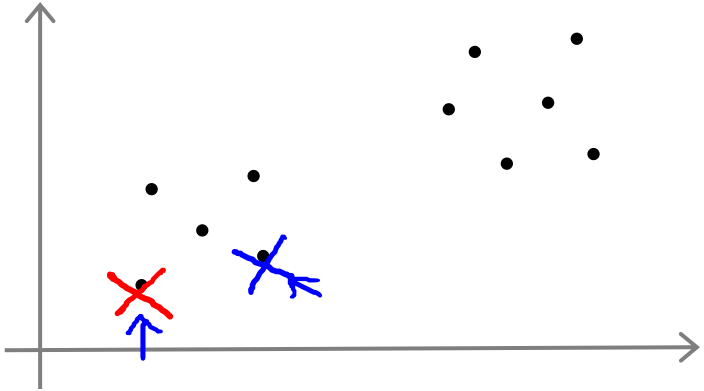
Set $\mu_1, \ldots, \mu_K$ equal to these K examples.

$\mu_1 = x^{(i)}$

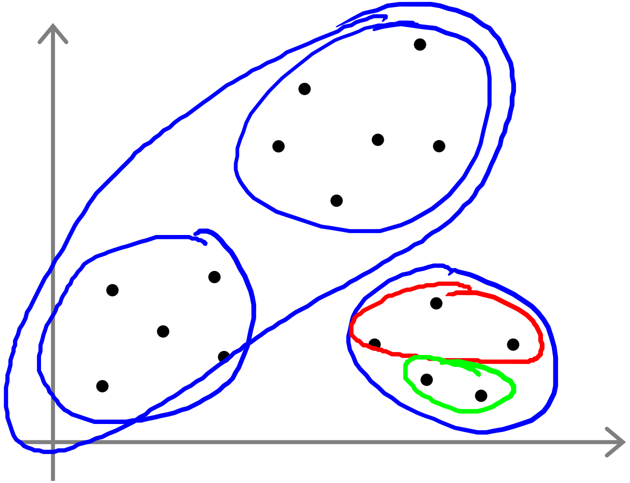
$\mu_2 = x^{(j)}$

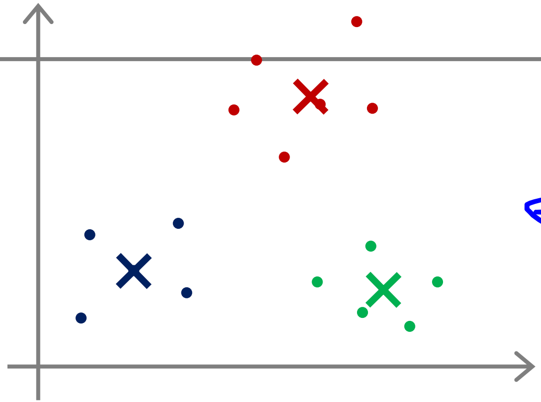
$\vdots$



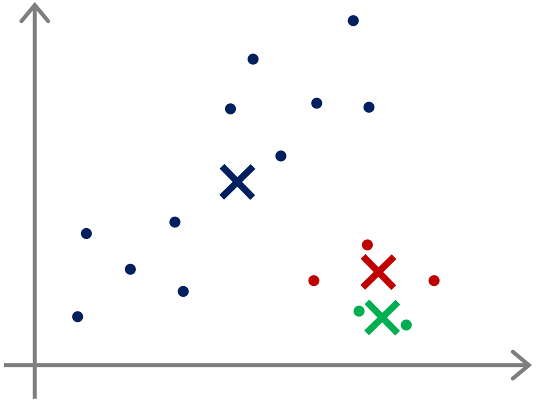


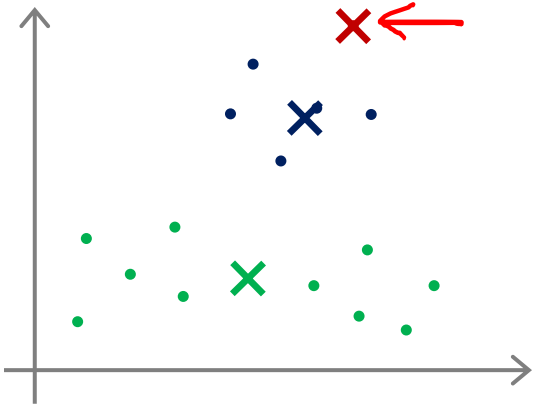
K-Means算法的局部收敛结果





$J(c^{(1)},\dots,c^{(m)},\mu_1,\dots,\mu_K)$





通过多次随机初始化解决陷入局部极值的问题

For i = 1 to 100 {

Randomly initialize K-means.

Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.

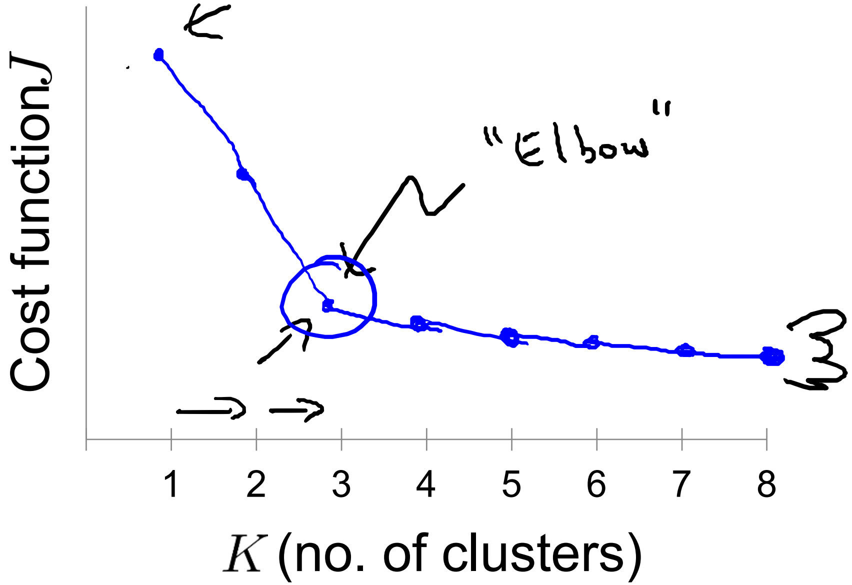
Compute cost function (distortion)

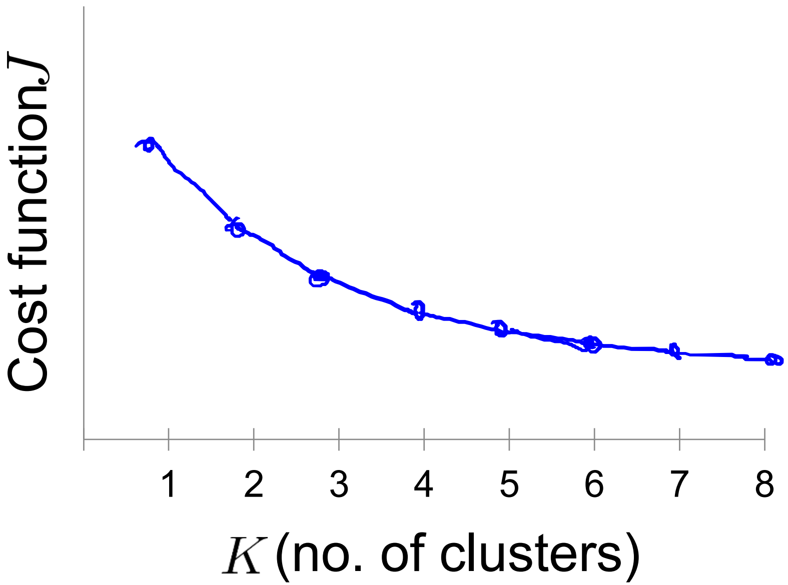
$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

}

Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

选择合适聚类簇数K（Elbow方法）



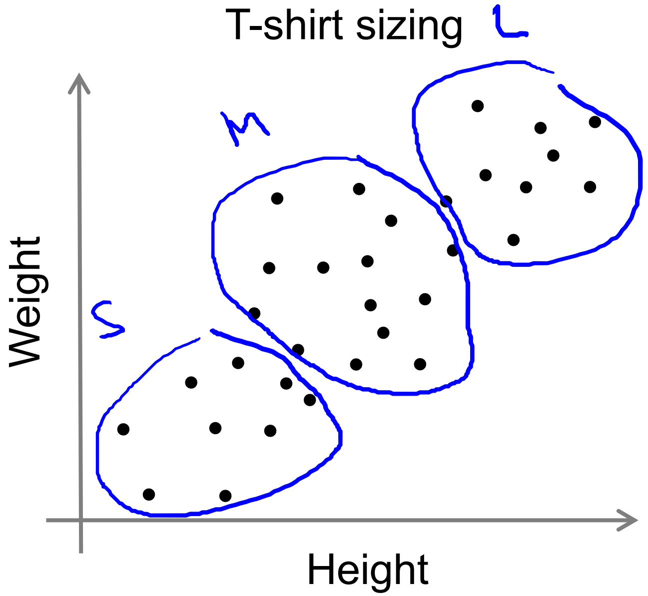


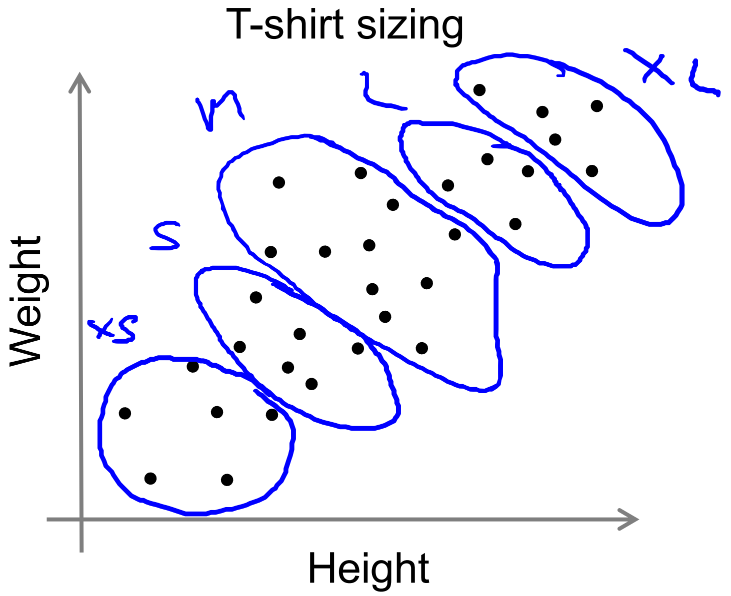
选择合适聚类簇数K

Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

K=3 S, M, L K=5 XS, S, M, L, XL

E.g.





4.2 Gaussian Mixture Model

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
高斯混合成分个数 k.

过程:

1: 初始化高斯混合分布的模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$
2: **repeat**
3: **for** $j = 1, 2, \ldots, m$ **do**
4: 根据式(9.30)计算 $\boldsymbol{x}_j$ 由各混合成分生成的后验概率, 即
$\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j)$ $(1 \leqslant i \leqslant k)$
5: **end for**

for $i = 1, 2, \ldots, k$ do

计算新均值向量: $\boldsymbol{\mu}_i' = \frac{\sum_{j=1}^m \gamma_{ji}\boldsymbol{x}_j}{\sum_{j=1}^m \gamma_{ji}}$;

计算新协方差矩阵: $\mathbf{\Sigma}_i' = \frac{\sum_{j=1}^m \gamma_{ji}(\boldsymbol{x}_j - \boldsymbol{\mu}_i')(\boldsymbol{x}_j - \boldsymbol{\mu}_i')^{\mathrm{T}}}{\sum_{j=1}^m \gamma_{ji}}$;

计算新混合系数: $\alpha_i' = \frac{\sum_{j=1}^m \gamma_{ji}}{m}$;

end for

将模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$ 更新为 $\{(\alpha_i', \boldsymbol{\mu}_i', \mathbf{\Sigma}_i') \mid 1 \leqslant i \leqslant k\}$

until 满足停止条件

$C_i = \varnothing \ (1 \leqslant i \leqslant k)$

for $j = 1, 2, \ldots, m$ do

根据式(9.31)确定 $\boldsymbol{x}_j$ 的簇标记 λ_j;

将 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{\boldsymbol{x}_j\}$

end for

输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

- DBSCAN(Density-Based Spatial Clustering of Applications with Noise) 连通簇簇定义(P212)

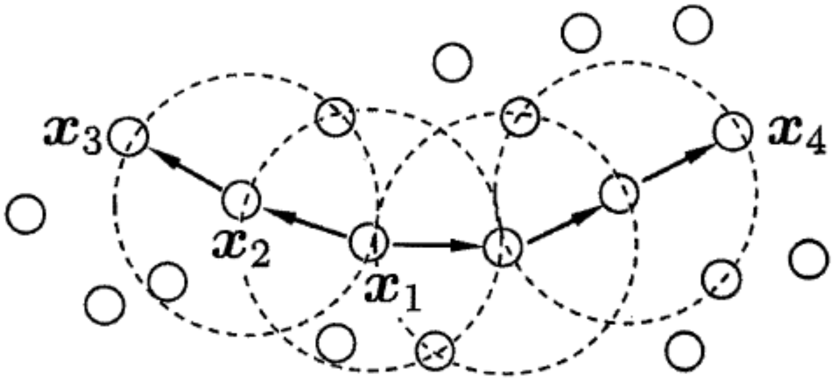


图 9.8 DBSCAN 定义的基本概念($MinPts = 3$): 虚线显示出 ϵ-邻域, x_1 是核心对象, x_2 由 x_1 密度直达, x_3 由 x_1 密度可达, x_3 与 x_4 密度相连.

5. DBSCAN

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
邻域参数 $(\epsilon, MinPts)$.

过程:

```
初始化核心对象集合: Ω = ∅
for j = 1, 2, ..., m do
    确定样本 x_j 的 ε-邻域 N_ε(x_j);
    if |N_ε(x_j)| ⩾ MinPts then
        将样本 x_j 加入核心对象集合: Ω = Ω ⋃ {x_j}
    end if
end for
```

初始化聚类簇数: $k = 0$

初始化未访问样本集合: $\Gamma = D$

while $\Omega \neq \varnothing$ **do**

  记录当前未访问样本集合: $\Gamma_{\text{old}} = \Gamma$;

  随机选取一个核心对象 $\boldsymbol{o} \in \Omega$, 初始化队列 $Q = <\boldsymbol{o}>$;

  $\Gamma = \Gamma \setminus \{\boldsymbol{o}\}$;

  while $Q \neq \varnothing$ **do**

    取出队列 Q 中的首个样本 $\boldsymbol{q}$;

    if $|N_\epsilon(\boldsymbol{q})| \geqslant MinPts$ **then**

      令 $\Delta = N_\epsilon(\boldsymbol{q}) \bigcap \Gamma$;

      将 Δ 中的样本加入队列 Q;

      $\Gamma = \Gamma \setminus \Delta$;

    end if

  end while

  $k = k + 1$, 生成聚类簇 $C_k = \Gamma_{\text{old}} \setminus \Gamma$;

  $\Omega = \Omega \setminus C_k$

end while

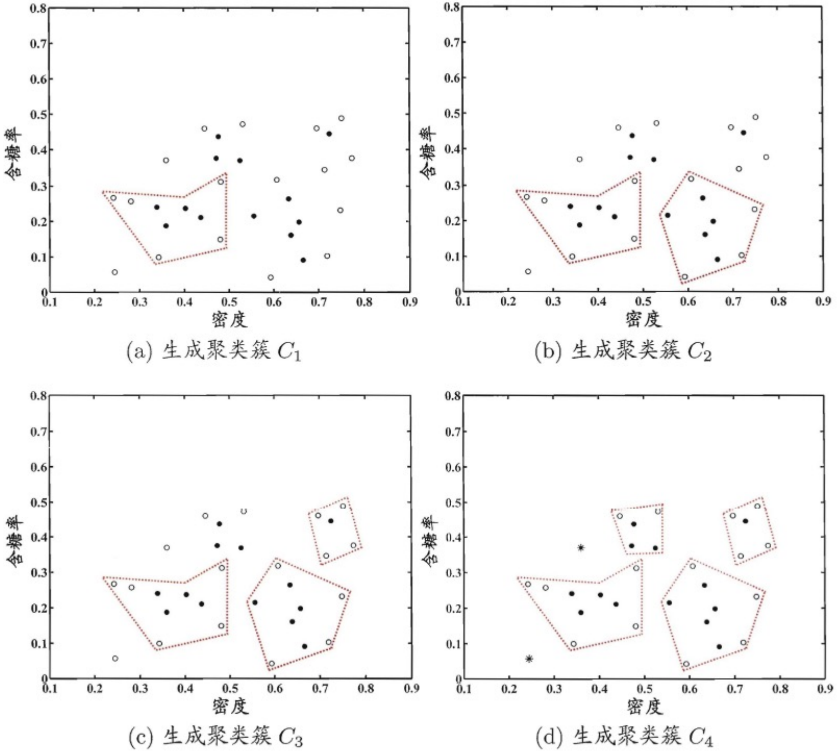


图 9.10 DBSCAN 算法($\epsilon = 0.11$, $MinPts = 5$)生成聚类簇的先后情况. 核心对象、非核心对象、噪声样本分别用“●” “○” “$*$”表示, 红色虚线显示出簇划分.

6. Hierarchical Clustering

- 在不同层次对数据集进行划分，从而形成树形的聚类结构
  - “自底向上”的聚合策略 (AGNES)
  - “自顶向下”的分拆策略
- AGNES (Agglomerative Nesting, 凝聚聚类)

6. AGNES

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
聚类簇距离度量函数 d;
聚类簇数 k.

过程:

for $j = 1, 2, \ldots, m$ **do**
  $C_j = \{\boldsymbol{x}_j\}$
end for
for $i = 1, 2, \ldots, m$ **do**
  for $j = 1, 2, \ldots, m$ **do**
    $M(i, j) = d(C_i, C_j)$;
    $M(j, i) = M(i, j)$
  end for
end for
设置当前聚类簇个数: $q = m$
while $q > k$ **do**
  找出距离最近的两个聚类簇 C_{i^*} 和 C_{j^*};
  合并 C_{i^*} 和 C_{j^*}: $C_{i^*} = C_{i^*} \bigcup C_{j^*}$;
  for $j = j^* + 1, j^* + 2, \ldots, q$ **do**
    将聚类簇 C_j 重编号为 C_{j-1}
  end for
  删除距离矩阵 M 的第 j^* 行与第 j^* 列;
  for $j = 1, 2, \ldots, q - 1$ **do**
    $M(i^*, j) = d(C_{i^*}, C_j)$;
    $M(j, i^*) = M(i^*, j)$
  end for
  $q = q - 1$
end while

输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

6. AGNES

- 类-类距离

$$\text{最小距离: } d_{\min}(C_i, C_j) = \min_{\boldsymbol{x} \in C_i, \boldsymbol{z} \in C_j} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) \ , \tag{9.41}$$

$$\text{最大距离: } d_{\max}(C_i, C_j) = \max_{\boldsymbol{x} \in C_i, \boldsymbol{z} \in C_j} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) \ , \tag{9.42}$$

$$\text{平均距离: } d_{\mathrm{avg}}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\boldsymbol{x} \in C_i} \sum_{\boldsymbol{z} \in C_j} \operatorname{dist}(\boldsymbol{x}, \boldsymbol{z}) \ . \tag{9.43}$$

- 对象-类距离

- 对象-对象距离

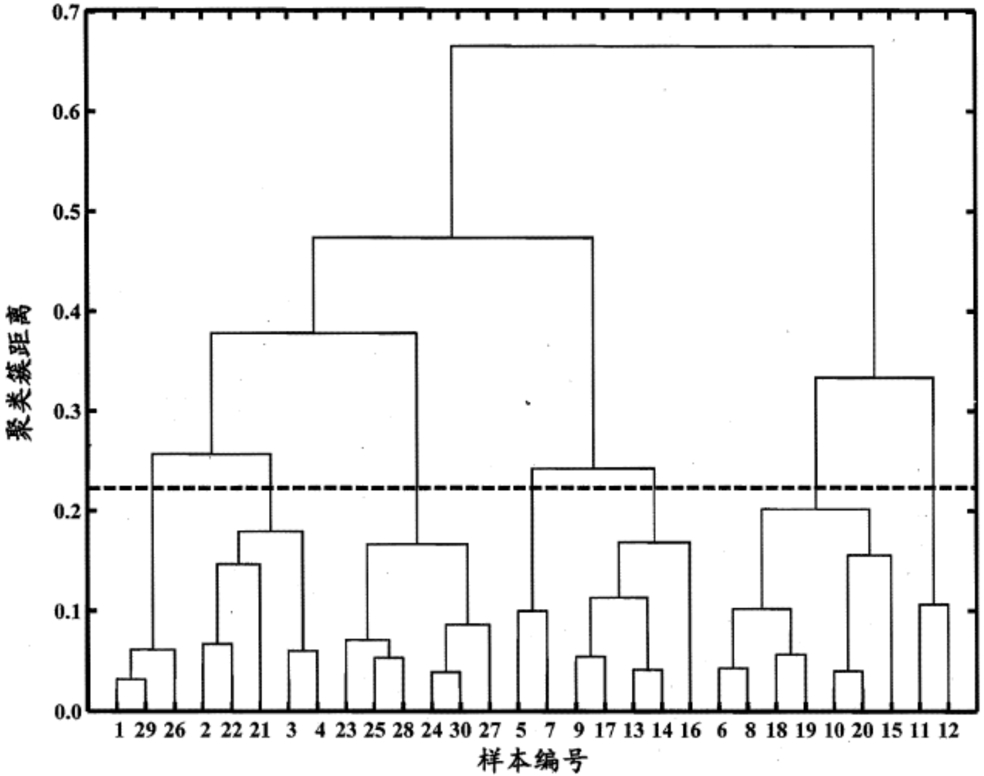
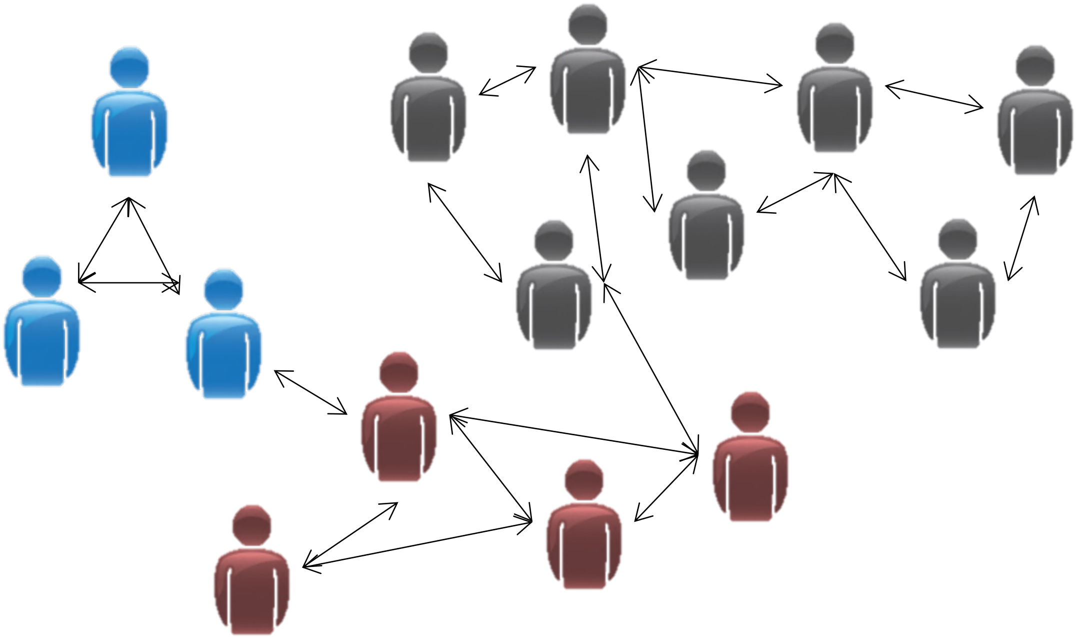


图 9.12 西瓜数据集 4.0 上 AGNES 算法生成的树状图(采用 $d_{\max}$). 横轴对应于样本编号, 纵轴对应于聚类簇距离.

其他：网络数据聚类算-Mincut



7. 双向层次聚类结果的可视化

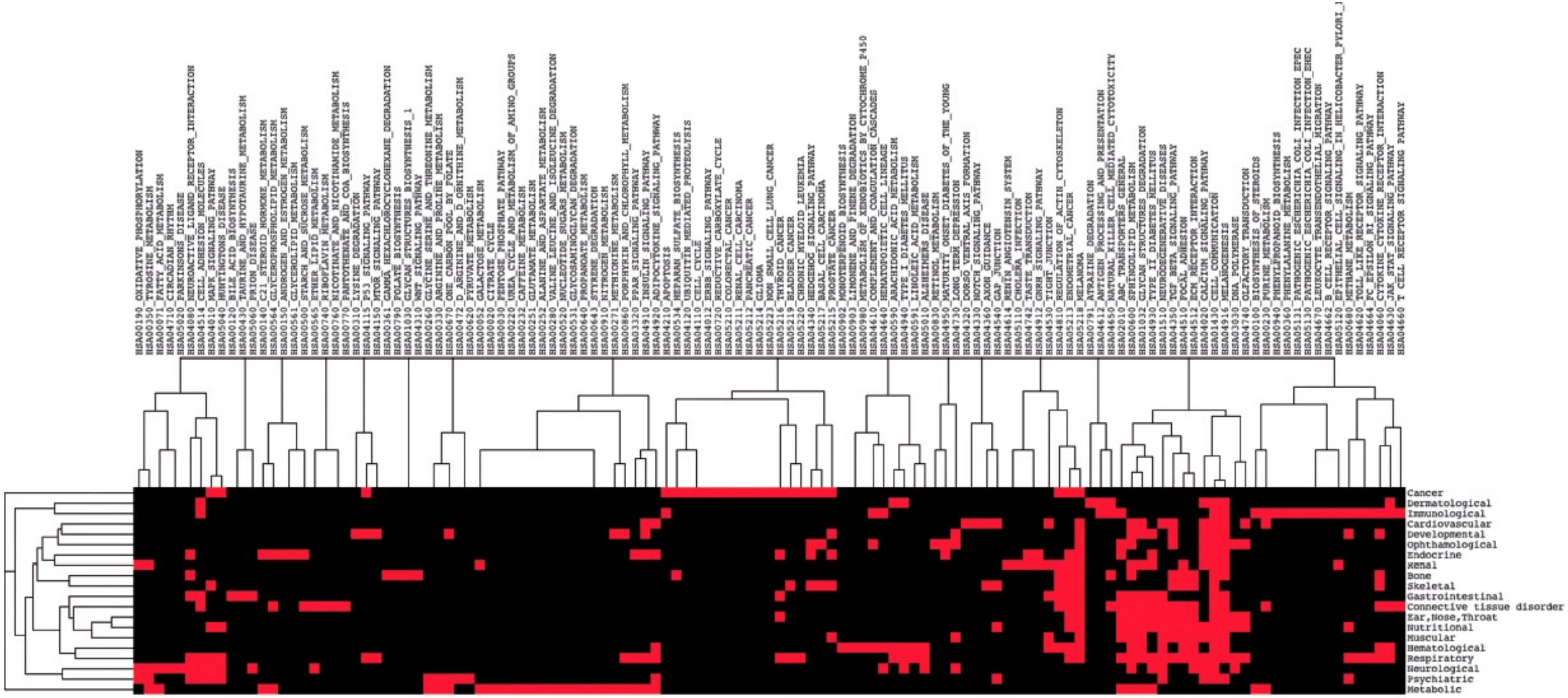
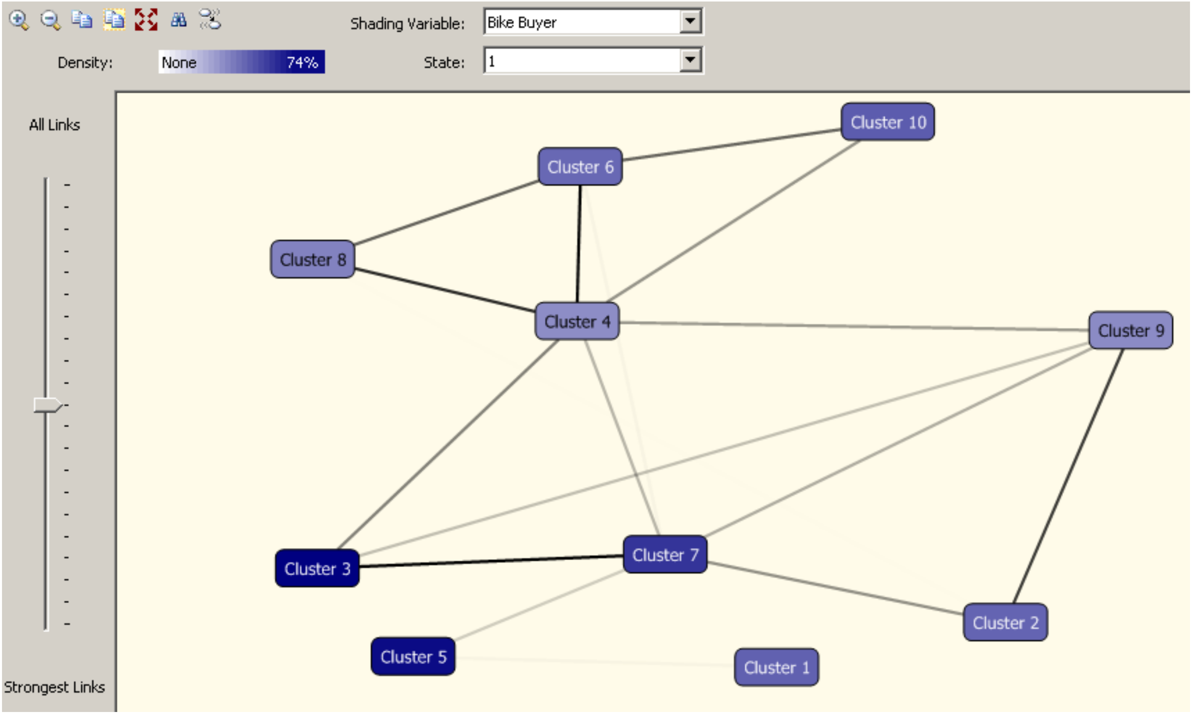


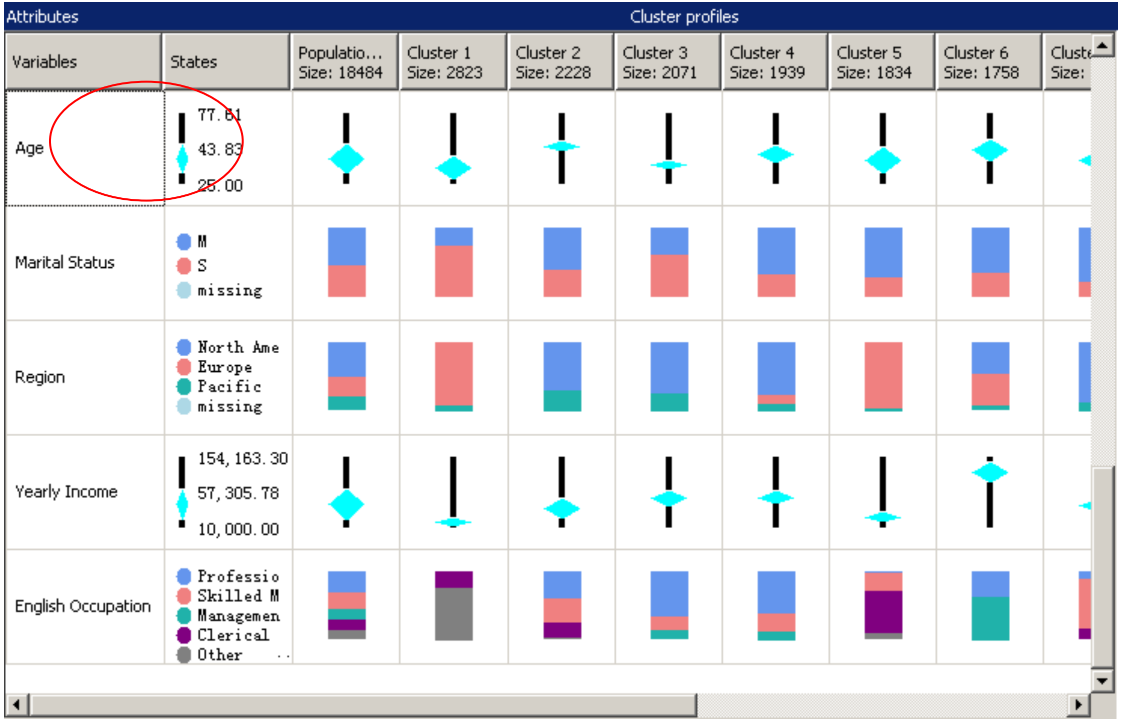
Figure 6. Predicted associations between disease classes and pathways. Each red entry represents a predicted association between 20 disease classes and 200 KEGG pathways.

7. 聚类结果可视化



- The lines between the clusters represent “closeness”
- The color of the cluster represents the frequency of the variable and state

其他：聚类结果可视化



- The distribution of an attribute's states for each cluster
- A colored bar: discrete attribute
- A diamond chart: continuous attribute

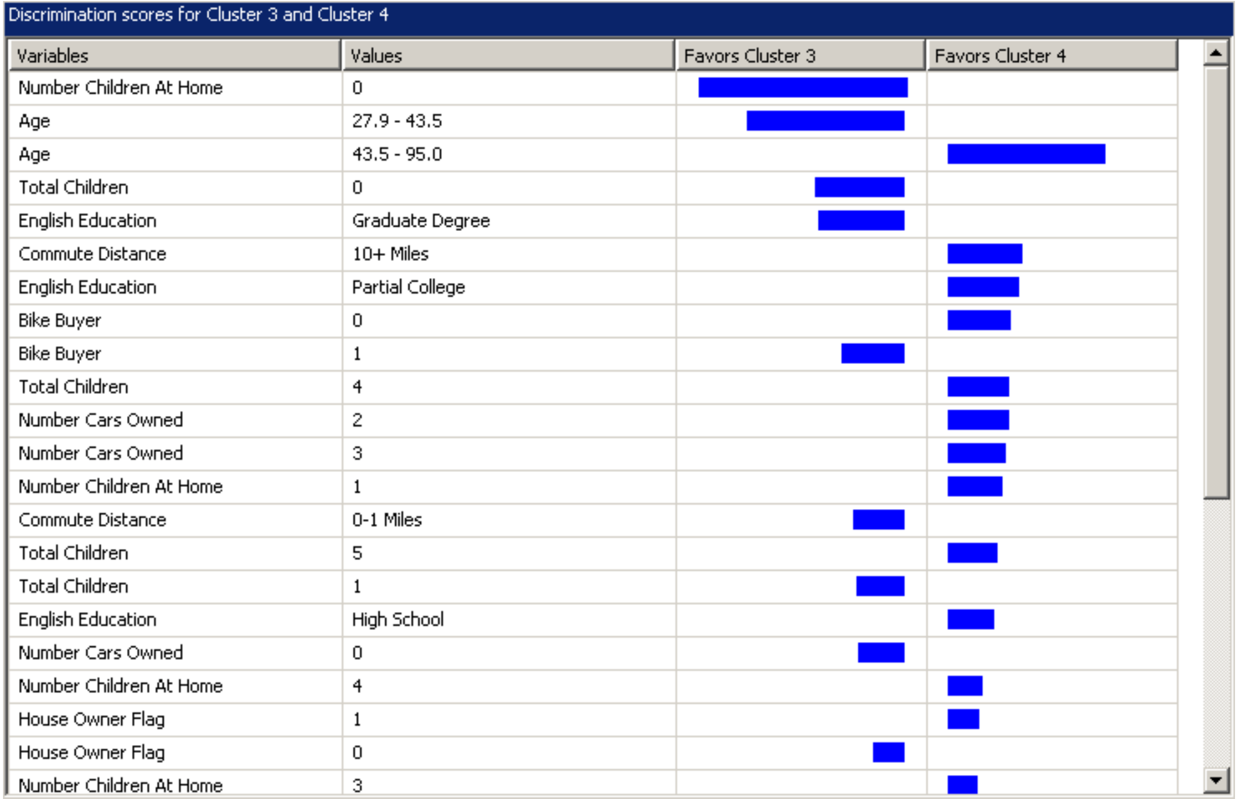
7. 聚簇特征

Characteristics for Cluster 3

Variables	Values	Probability
Bike Buyer	1	
Number Children At Home	0	
Region	North America	
Age	36.2 - 43.8	
English Occupation	Professional	
Marital Status	S	
House Owner Flag	1	
Gender	F	
First Name	missing	
English Education	Bachelors	
Commute Distance	0-1 Miles	
Yearly Income	57305.8 - 79082.2	
Number Cars Owned	0	
Gender	M	
Number Cars Owned	1	
English Education	Graduate Degree	
House Owner Flag	0	
Last Name	missing	
Marital Status	M	
Commute Distance	2-5 Miles	
Total Children	0	
Region	Pacific	

- The characteristics that make up a cluster in more detail

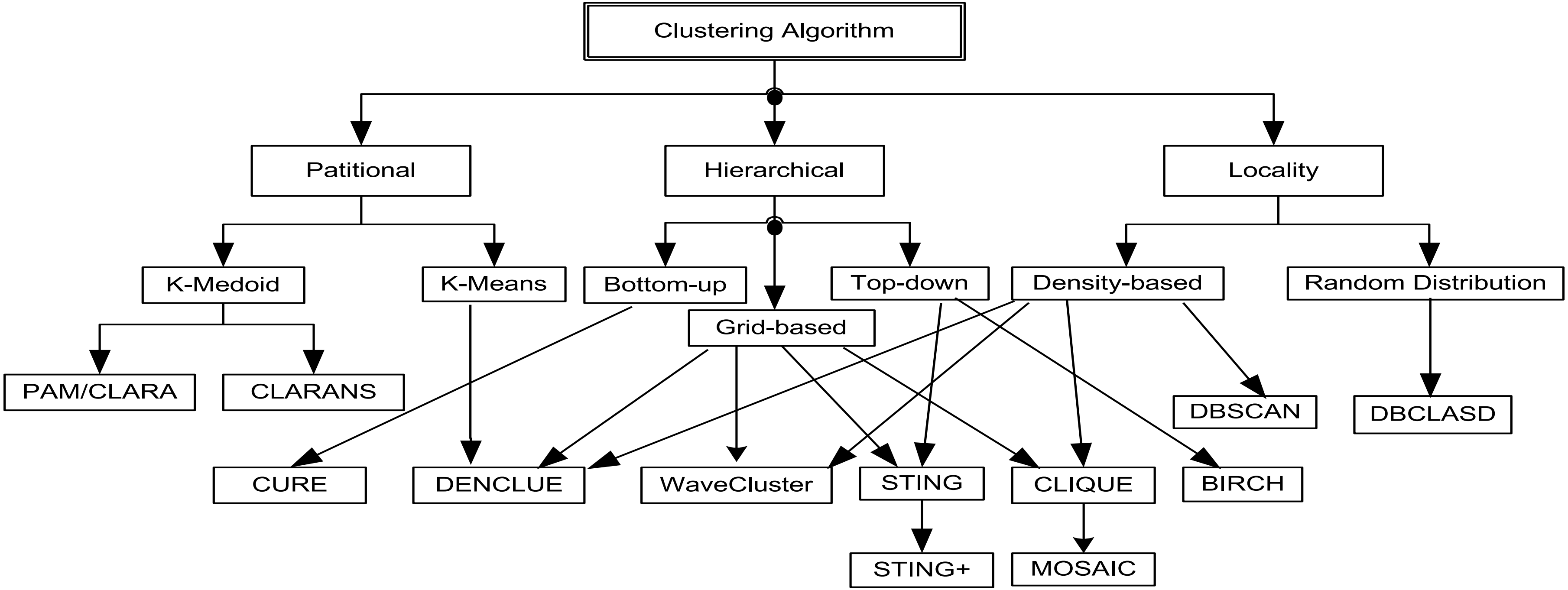
7. 区分簇的典型特征



Variables	Values	Favors Cluster 3	Favors Cluster 4
Number Children At Home	0		
Age	27.9 - 43.5		
Age	43.5 - 95.0		
Total Children	0		
English Education	Graduate Degree		
Commute Distance	10+ Miles		
English Education	Partial College		
Bike Buyer	0		
Bike Buyer	1		
Total Children	4		
Number Cars Owned	2		
Number Cars Owned	3		
Number Children At Home	1		
Commute Distance	0-1 Miles		
Total Children	5		
Total Children	1		
English Education	High School		
Number Cars Owned	0		
Number Children At Home	4		
House Owner Flag	1		
House Owner Flag	0		
Number Children At Home	3		

- The characteristics that distinguish one cluster from another

常见的聚类算法之间的关系



其他的聚类(Clustering)注意事项

- 可伸缩性
- 可以处理噪声
- 对输入次序不敏感
- 能够处理高维数据
- 可解释和可用性
- 处理不同类型属性的能力
- 可以发现任意形状的簇
- 用于决定输入参数的领域知识最小化

总结

1. 聚类简介
2. 聚类性能评价指标
3. 空间和距离（相似度）度量
4. 基于划分的聚类 (K-Means, GMM)
5. 基于密度的聚类 (DBSCAN)
6. 层次聚类方法 (AGNES)
7. 聚类可视化