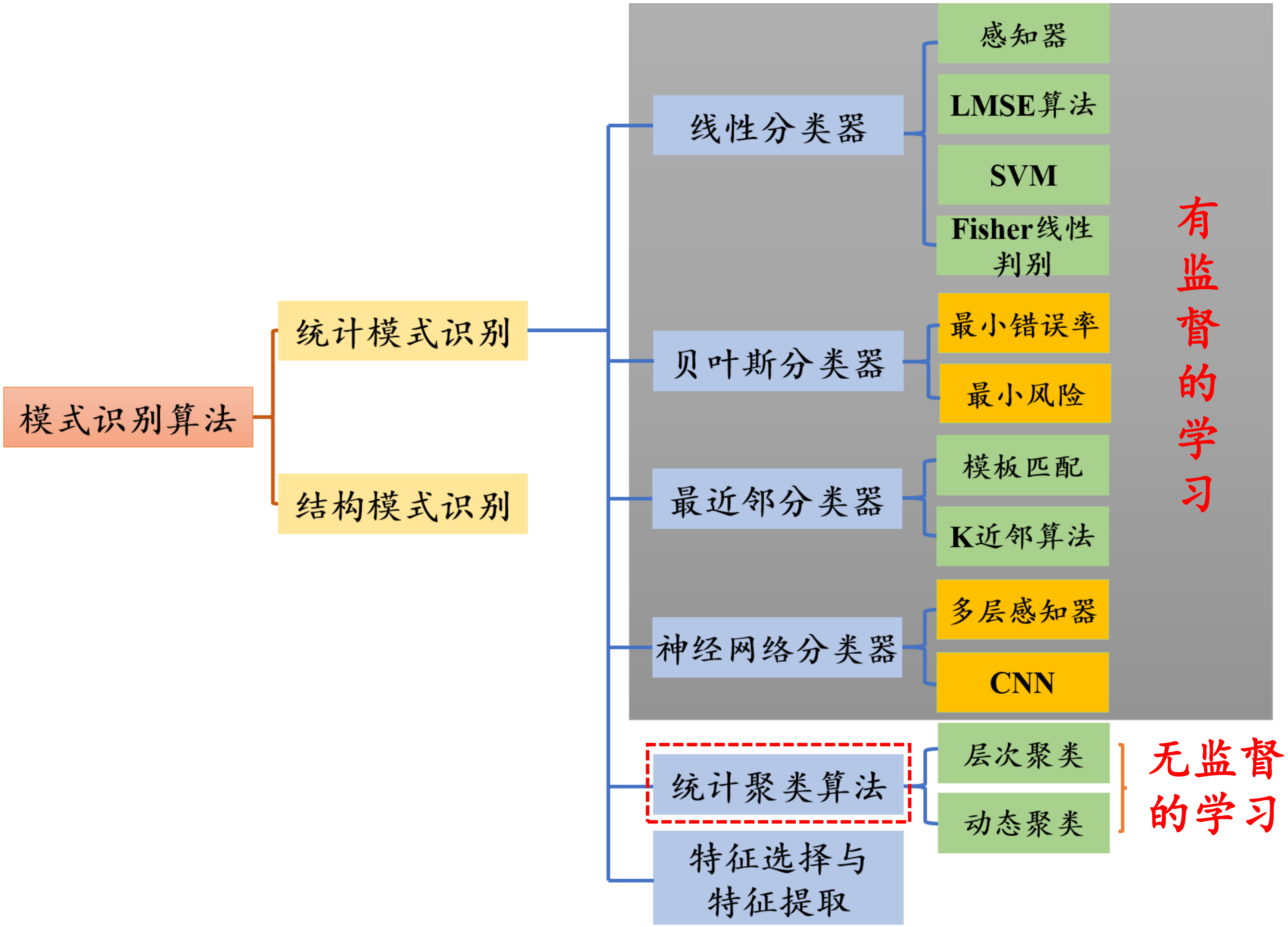
统计聚类算法

张俊超

中南大学
航空航天学院

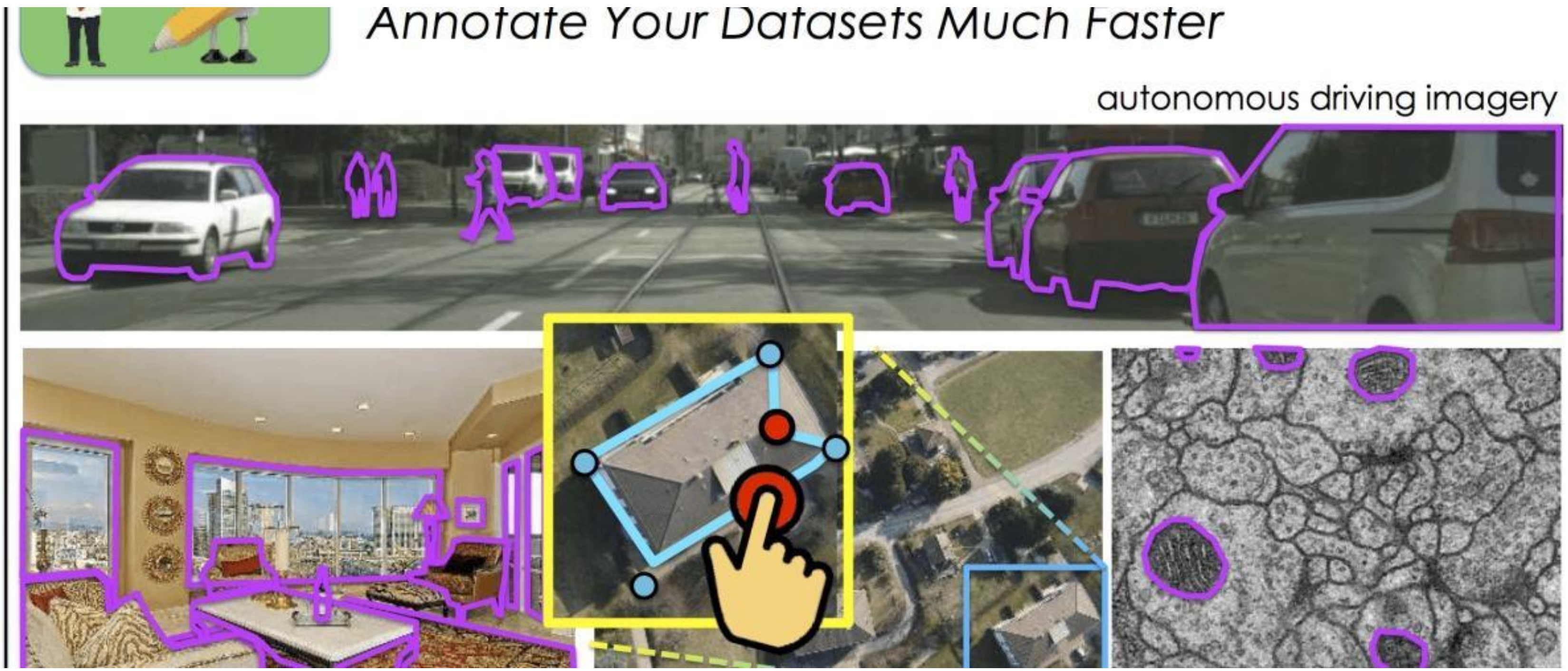
模式识别-统计聚类算法



模式识别算法
统计模式识别
结构模式识别
线性分类器
感知器
LMSE算法
SVM
Fisher线性判别
贝叶斯分类器
最小错误率
最小风险
最近邻分类器
模板匹配
K近邻算法
神经网络分类器
多层感知器
CNN
有监督的学习
统计聚类算法
层次聚类
动态聚类
无监督的学习
特征选择与特征提取

模式识别-统计聚类算法

AI发展背后的基础力量



模式识别-统计聚类算法

谷歌推出新方案，图像标注速度提高三倍

2018-11-26 10:15:19　来源：　科技日报

传统手动标记（中列）和流体标注（右）比较

核心思想：
数据聚类

模式识别-统计聚类算法

什么是聚类？和分类有什么区别？

模式识别-统计聚类算法

聚类的数学定义：聚类是指在模式空间 S 中，给定 N 个样本，**按照样本间的相似程度**，将 S 划分为 k 个决策区域 $S_i(i=1,2,\ldots,k)$ 的过程。该过程使得各样本均能归入其中一个类，且不会同时属于两个类。即：

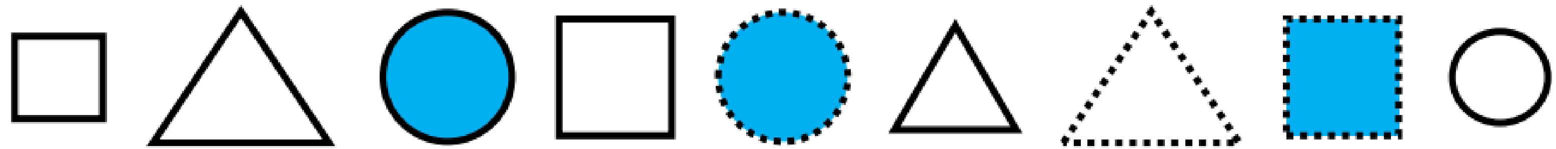
$$S_1 \cup S_2 \cup \cdots \cup S_k = S,\ 且 S_i \cap S_j = \varnothing,\quad i \neq j$$

根据定义，可以得出：

- 聚类是对整个样本集的划分，而不是对单个样本的识别；
- 聚类的依据是：样本间的相似程度；
- 聚类的结果是：无遗漏、无重复的。(不考虑样本可以属于多个类别的"软聚类")

聚类有什么特点呢？

模式识别-统计聚类算法



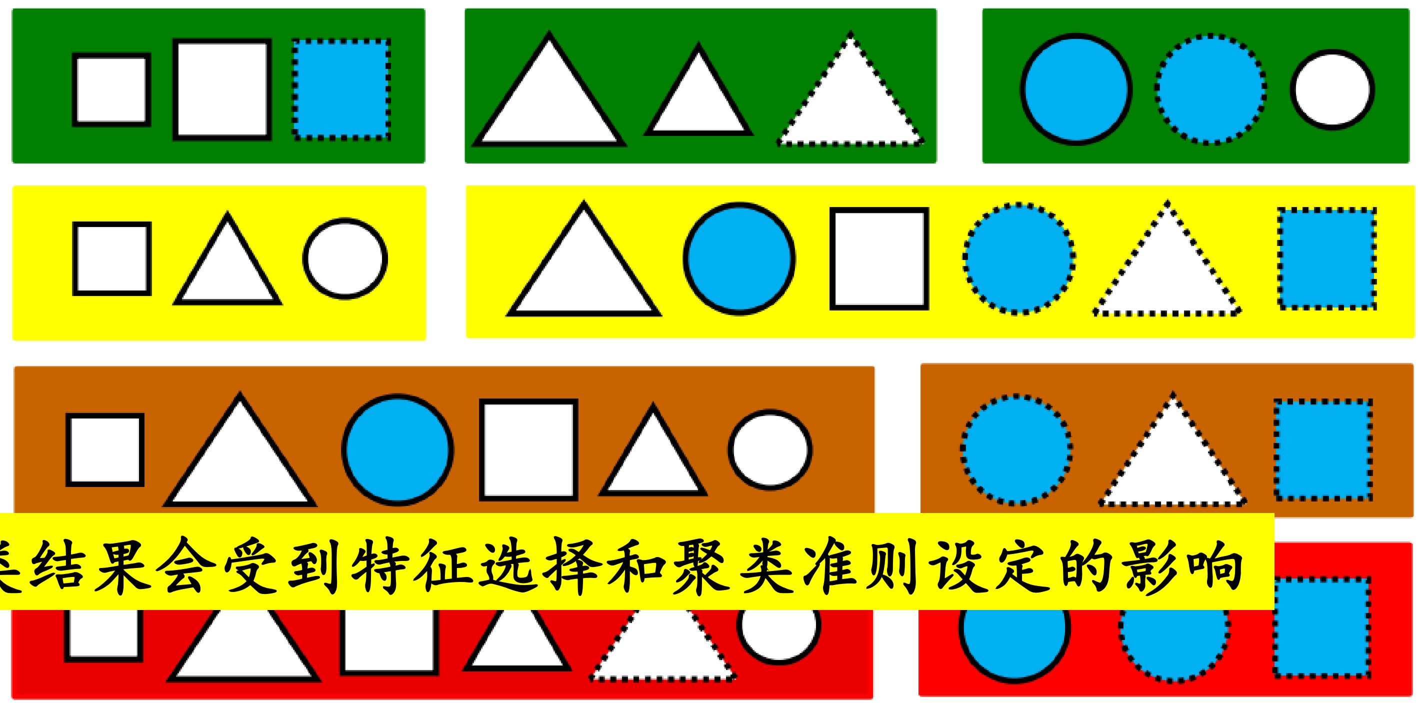
所选特征

形状

尺寸

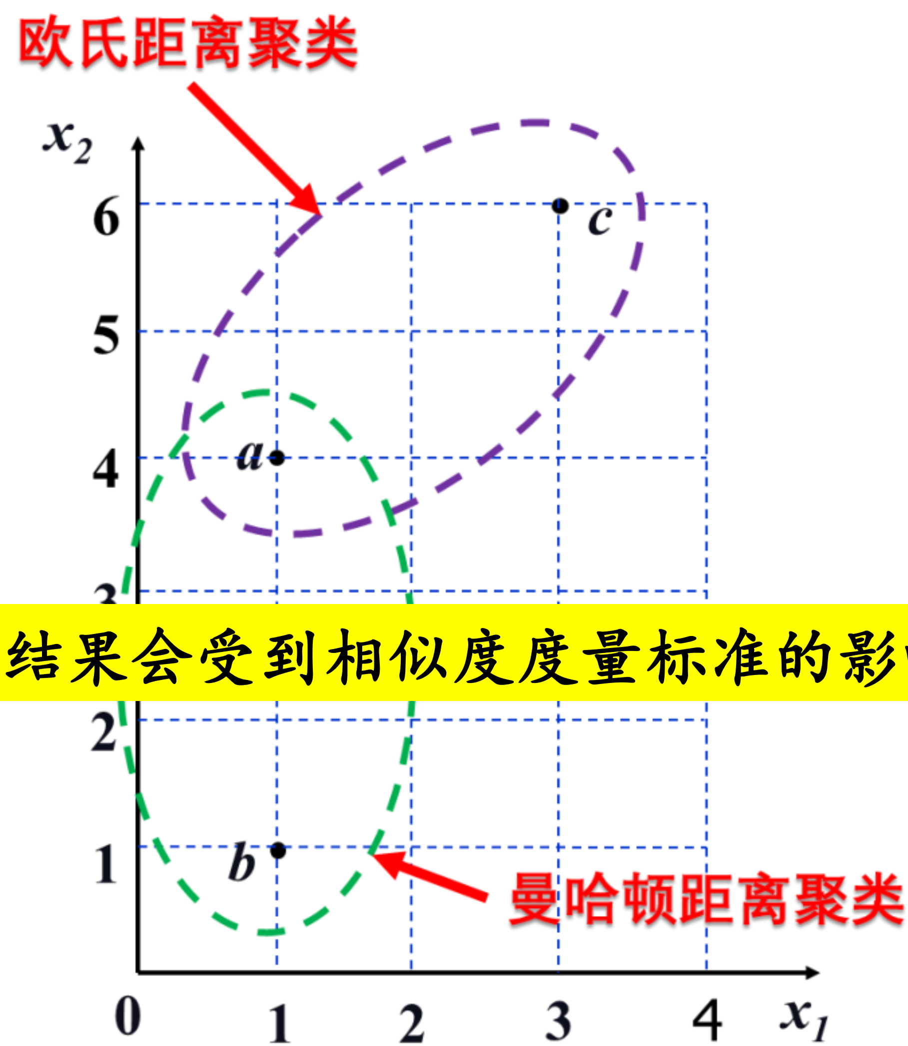
线种

颜色



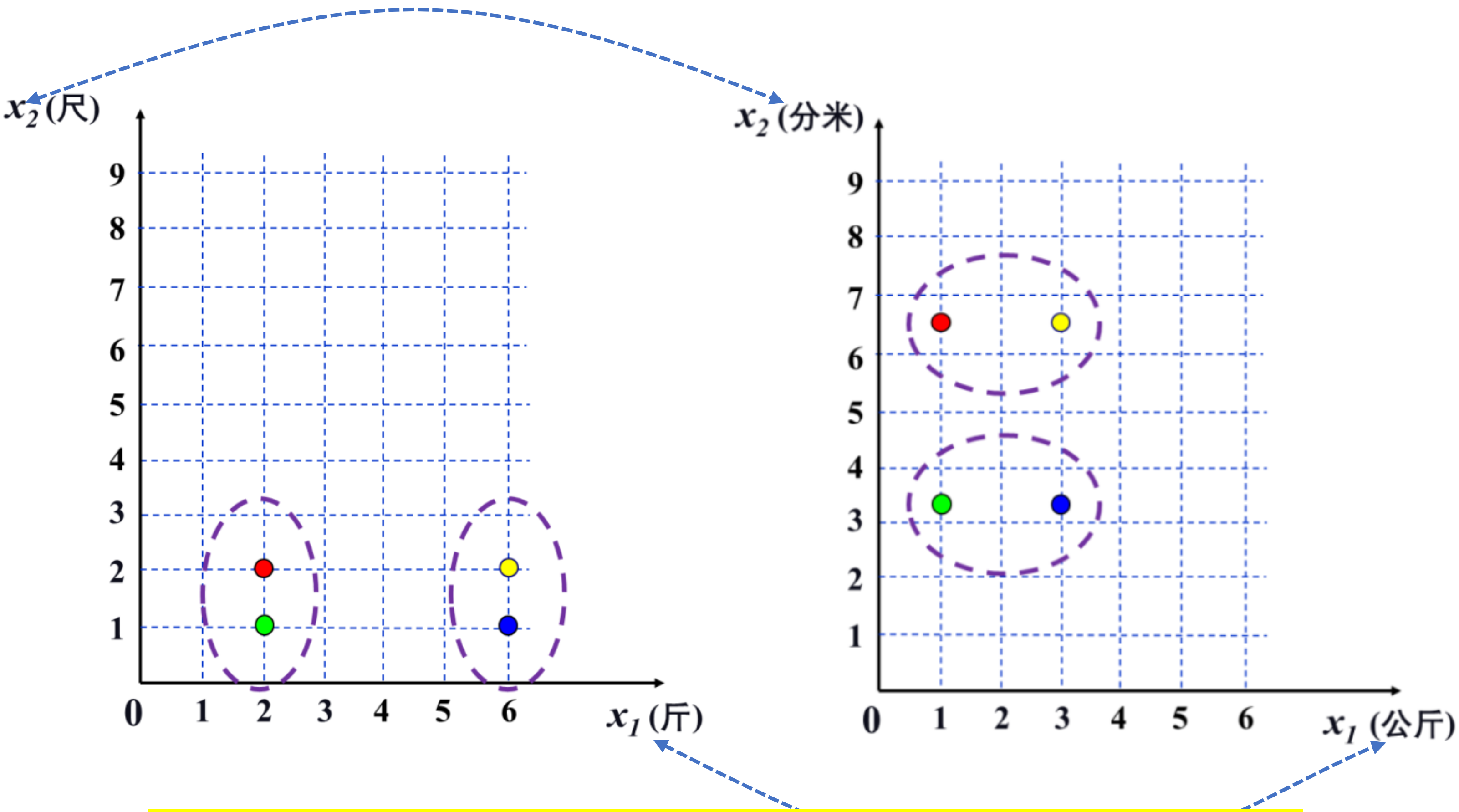
聚类结果会受到特征选择和聚类准则设定的影响

模式识别-统计聚类算法



聚类结果会受到相似度度量标准的影响

模式识别-统计聚类算法



聚类结果会受到各个特征的量纲标尺的影响

聚类的特点：

- 聚类结果会受到特征选择和聚类准则设定的影响
- 聚类结果会受到相似度度量标准的影响
- 聚类结果会受到各个特征的量纲标尺的影响

如何消除量纲标尺带来的影响？

归一化：

$$\bar{x} = \frac{x - x_{\mathrm{mix}}}{x_{\max} - x_{\min}}$$

而在某些聚类任务中，某些特征确实应该具有比其他特征更大的权重，所以进行归一化处理，反而会造成聚类结果变差。**归一化不是必做的操作。**

模式识别-统计聚类算法

聚类的一般流程：

模式识别-统计聚类算法

- 2. 确定相似度度量

样本间的相似度度量

- 曼哈顿距离 $d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$
- 欧几里德距离 $d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$
- 明考夫斯基距离 $d_{ij}(q) = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^q\right)^{\frac{1}{q}}$
- 切比雪夫距离 $d_{ij}(\infty) = \max_{1 \le k \le n} |x_{ik} - x_{jk}|$
- 非距离相似度

模式识别-统计聚类算法

类别间的相似度度量

- 最短距离：两类中相距最近的两样本间的距离

$$D_{h,k} = \min\{D(x_i, y_j)\}, x_i \in \text{类}h, y_j \in \text{类}k$$

- 最长距离：两类中相距最远的两样本间的距离

$$D_{h,k} = \max\{D(x_i, y_j)\}, x_i \in \text{类}h, y_j \in \text{类}k$$

- 重心距离：两类的均值点（重心）间的距离

$$D_{h,k} = D_{m_i,m_j}, m_i \text{为类}h\text{的重心，} m_j \text{为类}k\text{的重心}$$

- 类平均距离：两类中各个元素两两之间的距离相加后取平均值

$$D_{h,k} = \frac{1}{n_h n_k} \sum_{\substack{u \in h \\ m \in k}} d_{um}$$

- **3. 设定聚类准则**

  判定哪些样本应该聚为同一类。

① 误差平方和准则函数

$$J_e = \sum_{i=1}^{c} \sum_{k=1}^{n_i} \left\| \mathbf{x}_k^{(i)} - \mathbf{m}_i \right\|^2$$ → **最小化**

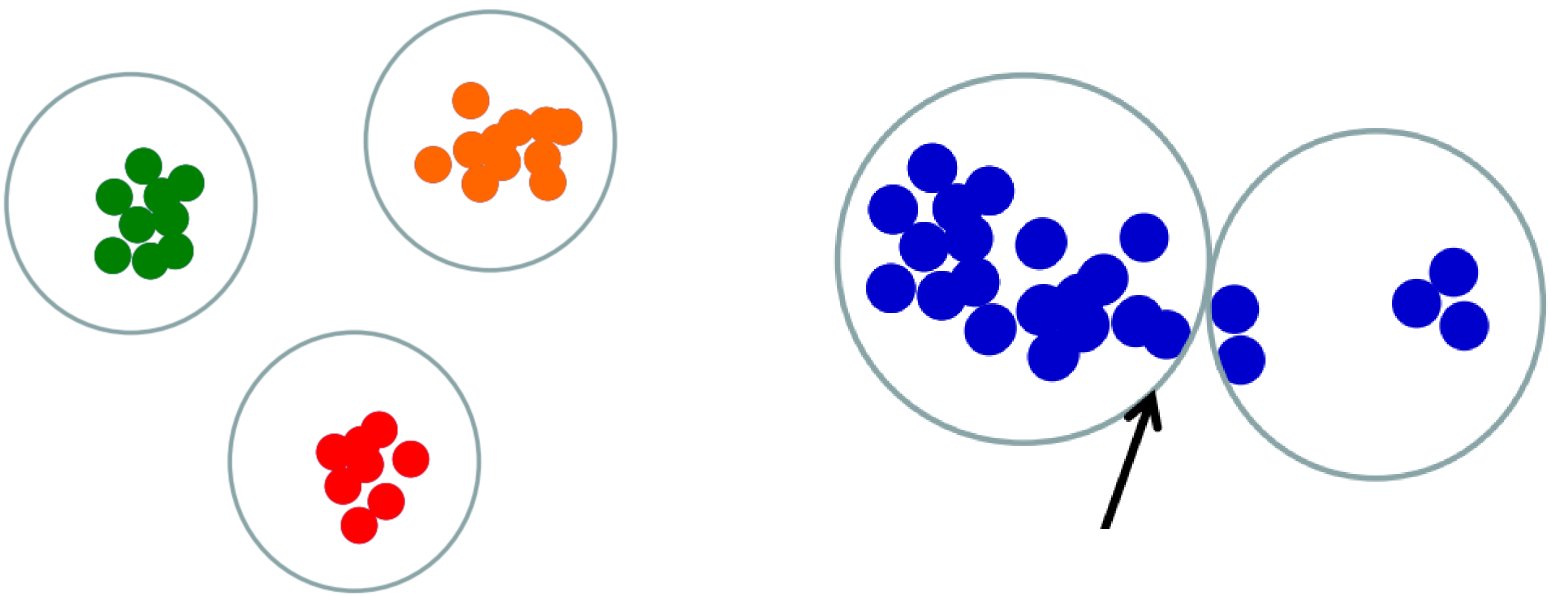
n_i : 类别ω_i的样本总数

$\mathbf{x}_k^{(i)}$: 类别ω_i的第k个样本

$\mathbf{m}_i$: 类别ω_i的样本均值, $\mathbf{m}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{x}_k^{(i)}$

【误差平方和准则函数】适用范围：

1. 同类样本分布相对密集；
2. 各类别所包含样本数相差不大、类间距离较大。



② 离散度准则函数

类内离散度矩阵：$\mathbf{S}_i = \frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)} - \mathbf{m}_i\right)\left(\mathbf{x}_k^{(i)} - \mathbf{m}_i\right)^T$

1. 总的类内离散度矩阵：$\mathbf{S}_w = \sum_{i=1}^{c} P_i \mathbf{S}_i$

2. 类间离散度矩阵：$\mathbf{S}_b = \sum_{i=1}^{c} P_i \left(\mathbf{m}_i - \mathbf{m}\right)\left(\mathbf{m}_i - \mathbf{m}\right)^T$

3. 总的离散度矩阵：$\mathbf{S}_t = \frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)} - \mathbf{m}\right)\left(\mathbf{x}_k^{(i)} - \mathbf{m}\right)^T$，其中$n = \sum_{i=1}^{c} n_i$

三者之间有什么关系呢？

模式识别-统计聚类算法

$$\mathbf{S}_t = \frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}\right)^T$$

$$=\sum_{i=1}^{c}\frac{n_i}{n}\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}\right)^T$$

$$=\sum_{i=1}^{c}P_i\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}\right)^T\right]$$

$$=\sum_{i=1}^{c}P_i\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i-\left(\mathbf{m}-\mathbf{m}_i\right)\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i-\left(\mathbf{m}-\mathbf{m}_i\right)\right)^T\right]$$

$$=\sum_{i=1}^{c}P_i\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)^T+\left(\mathbf{m}-\mathbf{m}_i\right)\left(\mathbf{m}-\mathbf{m}_i\right)^T\right]$$

$$=\sum_{i=1}^{c}P_i\left[\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)^T\right]+\sum_{i=1}^{c}P_i\left(\mathbf{m}-\mathbf{m}_i\right)\left(\mathbf{m}-\mathbf{m}_i\right)^T$$

$$=\sum_{i=1}^{c}P_i\mathbf{S}_i+\mathbf{S}_b=\mathbf{S}_w+\mathbf{S}_b$$

$$\mathbf{S}_t=\mathbf{S}_w+\mathbf{S}_b$$

三个离散度矩阵之间的关系

仅和样本全体的分布有关，
而和聚类结果无关。

此消彼长
相互依存
相互制约

$$S_t = S_w + S_b$$

与聚类结果直接关联
两者的总和保持不变
不便直接进行评估

离散度准则函数：基于迹的准则函数

$$\mathbf{S}_i=\frac{1}{n_i}\sum_{k=1}^{n_i}\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)^T=\frac{1}{n_i}\sum_{k=1}^{n_i}\mathbf{A}_i$$

$$令\left(\mathbf{x}_k^{(i)}-\mathbf{m}_i\right)^T=\begin{bmatrix}a_{i1} & a_{i2} & \cdots & a_{id}\end{bmatrix}$$

$$\mathbf{A}_i=\begin{bmatrix}a_{i1}\\a_{i2}\\\vdots\\a_{id}\end{bmatrix}\begin{bmatrix}a_{i1} & a_{i2} & \cdots & a_{id}\end{bmatrix}=\begin{bmatrix}a_{i1}^2 & a_{i1}a_{i2} & \cdots & a_{i1}a_{id}\\a_{i2}a_{i1} & a_{i2}^2 & \cdots & a_{i2}a_{id}\\\vdots & \vdots & \ddots & \vdots\\a_{id}a_{i1} & a_{id}a_{i2} & \cdots & a_{id}^2\end{bmatrix}$$

$$tr\left(\mathbf{A}_i\right)=\sum_{i=1}^{d}a_{id}^2=\left\|\mathbf{x}_k^{(i)}-\mathbf{m}_i\right\|^2$$

$$tr\left(\mathbf{S}_i\right)=\frac{1}{n_i}\sum_{k=1}^{n_i}tr\left(\mathbf{A}_i\right)=\frac{1}{n_i}\sum_{k=1}^{n_i}\left\|\mathbf{x}_k^{(i)}-\mathbf{m}_i\right\|^2$$

模式识别-统计聚类算法

$tr(\mathbf{S}_i)$：其值越小,类内样本聚集程度越高

$tr(\mathbf{S}_w)$：其值反映了同类样本的聚集程度$\rightarrow$加权的误差平方和准则函数

$tr(\mathbf{S}_b)$：其值反映了不同类样本的分离程度

准则函数：

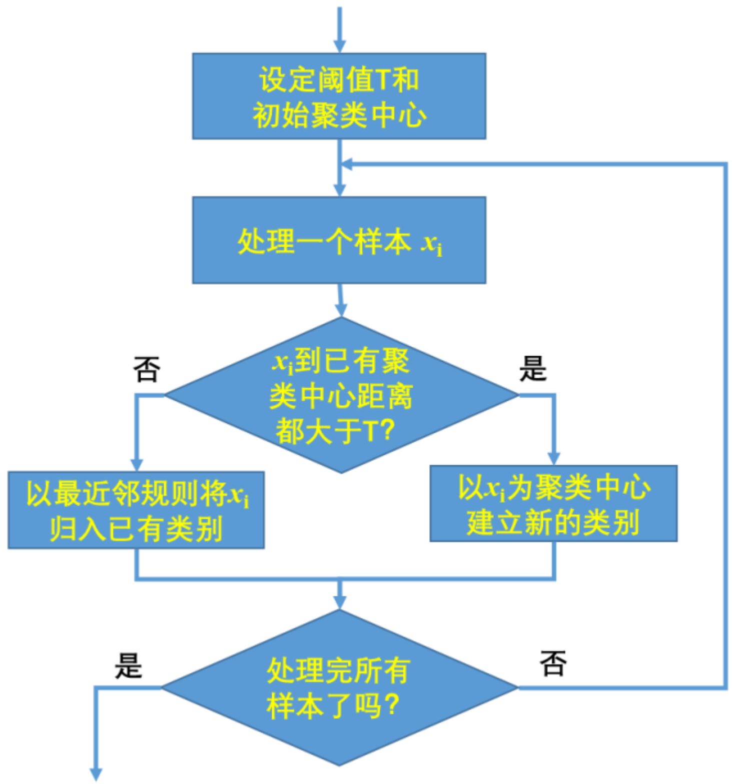
$$J = tr(\mathbf{S}_w)\text{：最小化}$$

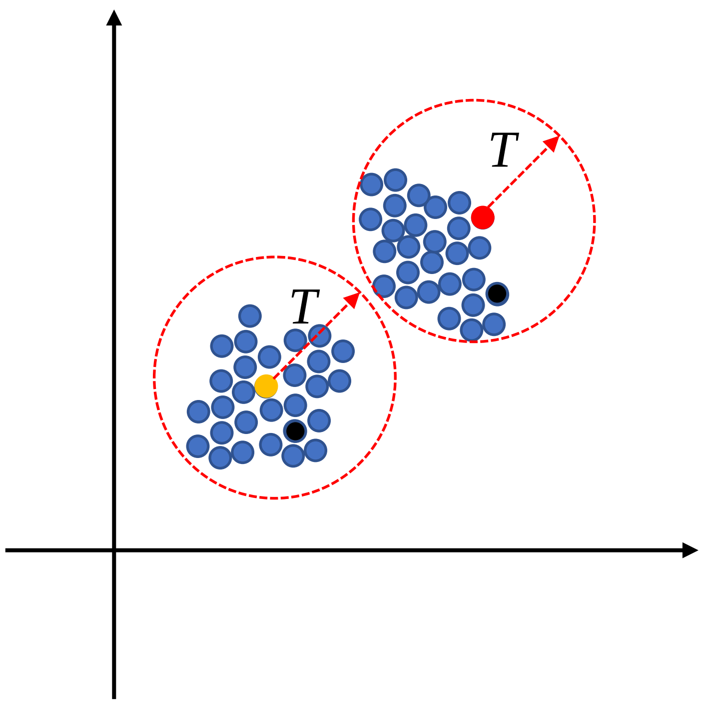
$$J = tr(\mathbf{S}_b)\text{：最大化}$$

$$J = tr(\mathbf{S}_w^{-1}\mathbf{S}_b)\text{：最大化}$$

- **4. 选择聚类算法**

① 试探法聚类(基于最近邻规则的试探聚类算法)





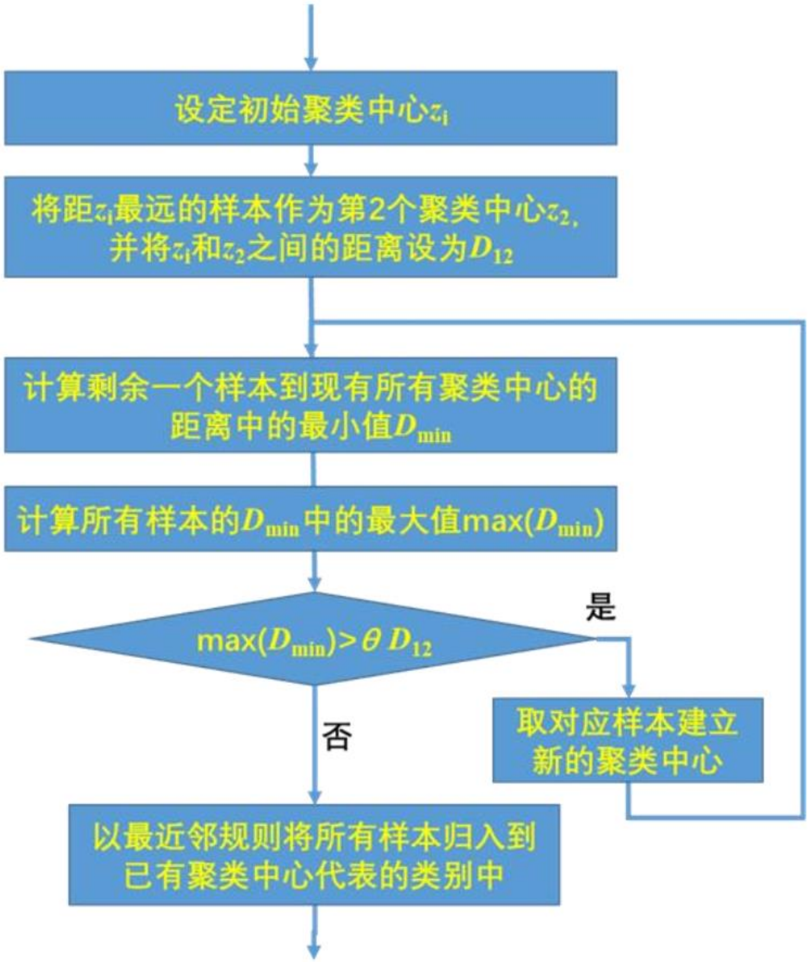
模式识别-统计聚类算法

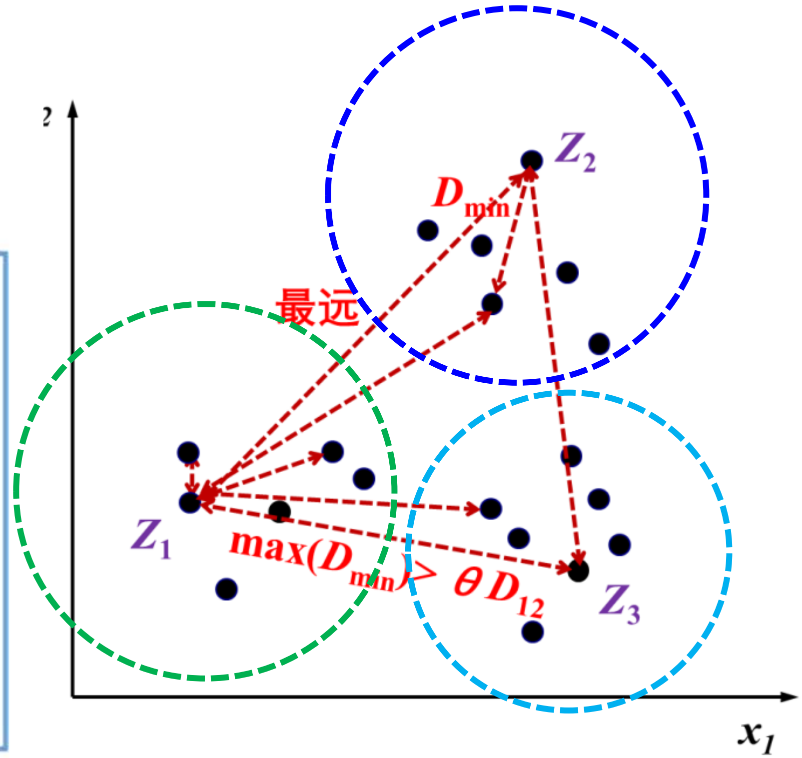
基于最近邻规则的试探聚类算法【特点】

- 聚类结果中所有类内的样本距聚类中心的距离都在以 T 为半径的范围内。
- 分类结果受第一个聚类中心的选择、待分类模式样本的排列顺序和阈值T的大小的影响。
- 一个样本一旦划归到某一类中之后，就无法再剔除或调整。

模式识别-统计聚类算法

① 试探法聚类(基于最大最小距离的试探聚类算法)





模式识别-统计聚类算法

```
%%
% 生成数据
randn('seed',2020);
mu1 = [2 3];
sigma1 = [0.2 0;
          0 0.2];
data1 = mvnrnd(mu1,sigma1,300);
randn('seed',2021);
mu2 = [4 5];
sigma2 = [0.2 0;
          0 0.2];
data2 = mvnrnd(mu2,sigma2,300);

Data = [data1;data2];
N = size(Data,1);
List = randperm(N);
Data = Data(List,:);
```

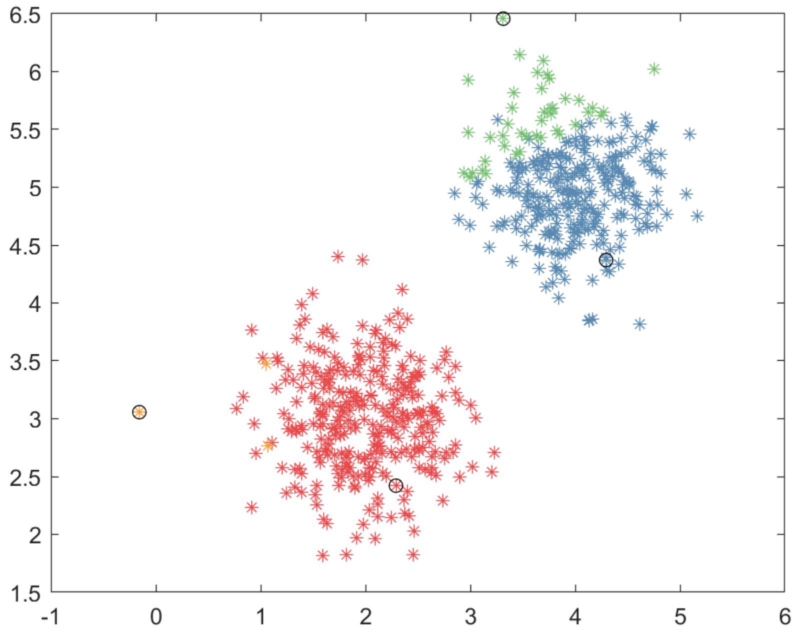
模式识别-统计聚类算法

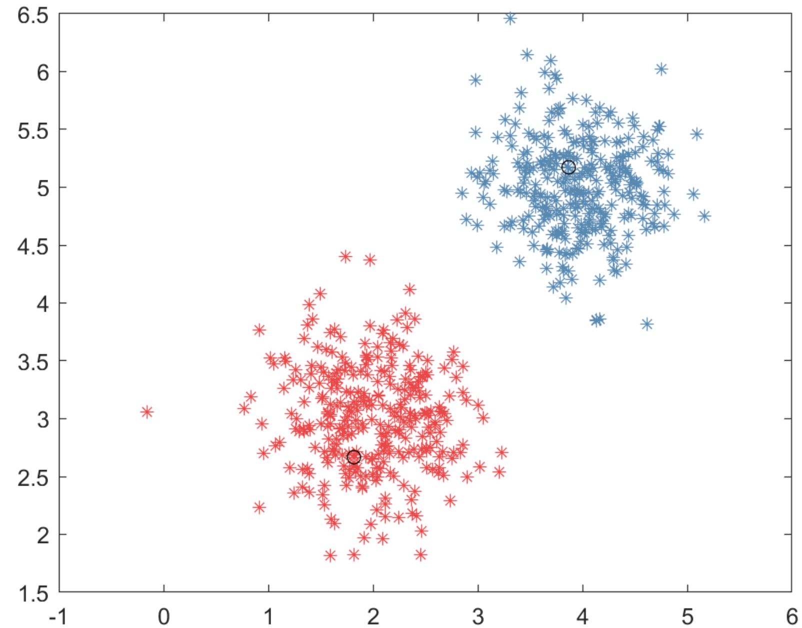
```
%% 试探聚类算法
% 基于最近邻规则的试探聚类算法
T = 5;
Labels = zeros(N,1);
start_L = randperm(N);
start = start_L(1);
Z = [];
Z_sub = []; % 保存中心点对应的位置
Z = [Z;Data(start,:)]; %保存中心点
Z_sub = [Z_sub;start];
label = 1;
Labels(start) =label;
```

模式识别-统计聚类算法

```
for i = 1:N
    dist = Z - repmat(Data(i,:),[size(Z,1),1]);
    dist = sum(dist.^2,2);
    [minv,idx] = min(dist);
    if minv>T
        Z = [Z;Data(i,:)];
        Z_sub = [Z_sub;i];
        label = label + 1;
        Labels(i) =label;
    else
        tmp = Z_sub(idx);
        Labels(i) = Labels(tmp);
    end
end
```

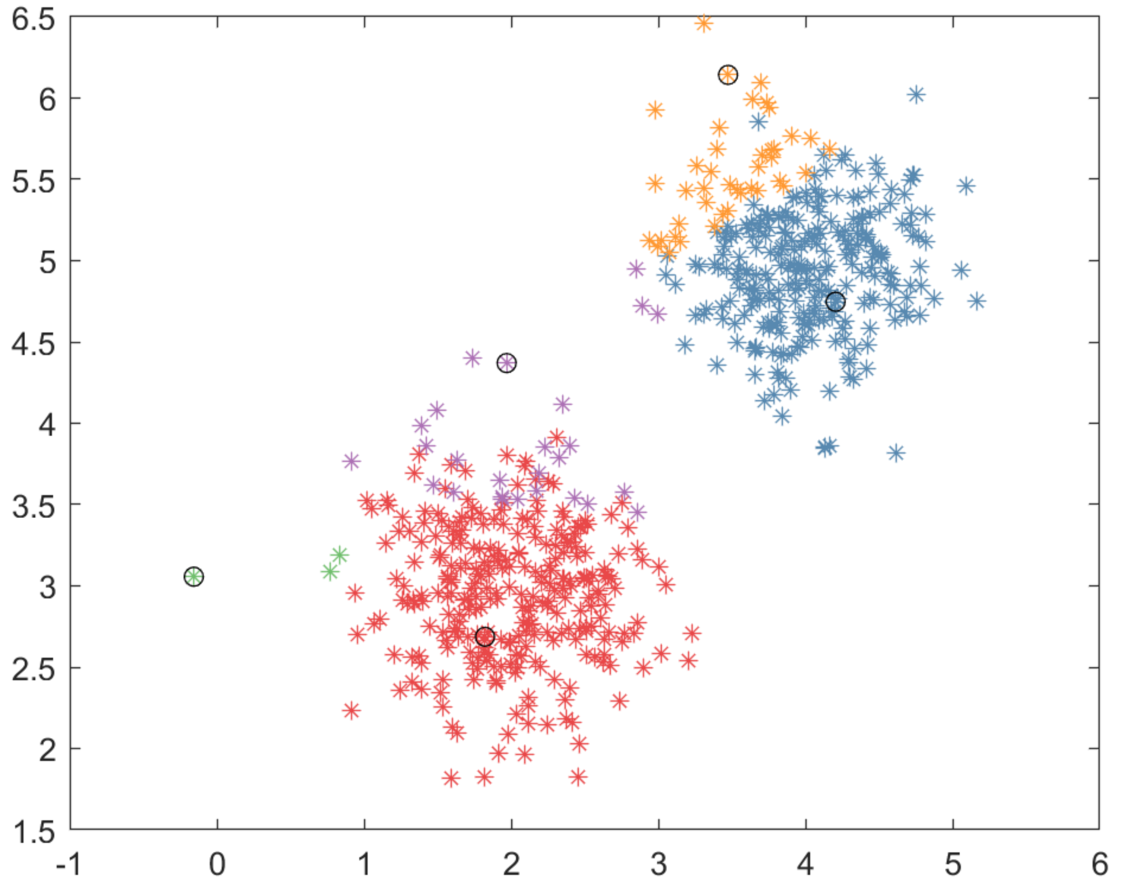
模式识别 -统计聚类算法





$T = 5$,不同初始样本点

模式识别 -统计聚类算法



$T=2$

【作业】编程实现基于最大最小距离的聚类算法。

说明：

1.数据可以自己生成(也可以用提供的)

2.编程语言：Matlab/Python

3. 提交实验报告和源代码(**命名规则：聚类_学号_姓名**)

4. 作业迟交n天，本次作业分数乘以0.98^n。


若干素材取自网络，特此致谢！


谢谢聆听！