模式识别的理论与方法

Pattern Recognition

裴继红

无监督学习和聚类

混合密度估计和k均值聚类

Chapter 10 (Part 1):10.1-10.4

混合密度估计和k均值聚类

1. 无监督与有监督
2. 混合密度
3. 最大似然估计
4. 混合正态密度估计
5. k-均值聚类

1、无监督与有监督

- 统计模式识别的任务：
  - 利用获得的样本集合中样本构造分类器，并进一步对（未知）样本进行分类。

➢ 有监督数据集合

- 训练样本集合中的每个样本已分类（已标记），即类别标记已知，这些样本用于构造分类器。

➢ 无监督数据集合

- 训练样本集合中的每个样本未分类（未标记），需要对这些样本分类，或构造分类器。

无监督学习假设

- **假设样本集的概率结构已知，只有参数未知：**

  ① 样本集合 D 中的样本类别未知（未被标记）；

  ② 所有样本来自 c 个类，类数c已知；

  ③ 每个类别的先验概率 $P(\omega_j)$ 已知，$j=1,2,\ldots,c$;

  ④ c 个类的类条件概率密度函数的形式 $P(x|\omega_j,\theta_j)$ 是已知的，$j = 1,2,\ldots,c$;

  ⑤ 参数向量 $\theta = (\theta_1, \theta_2,\ldots, \theta_c)$ 是未知的。

➢ 无监督学习的目的：由无监督样本集 D 估计参数向量 θ

混合密度估计和k均值聚类

1. 无监督与有监督
2. 混合密度
3. 最大似然估计
4. 混合正态密度估计
5. k-均值聚类

2、混合密度

- 在无监督学习假设下，样本x的生成密度：

$$p(\mathbf{x}|\boldsymbol{\theta})=\sum_{j=1}^{c}p(\mathbf{x}|\omega_j,\boldsymbol{\theta}_j)P(\omega_j)$$

其中：

$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_c)^T$ 为参数向量，未知； $p(\mathbf{x}|\boldsymbol{\theta})$ 为混合密度，

$p(\mathbf{x}|\omega_j,\boldsymbol{\theta}_j)$ 为分量密度， $P(\omega_j)$ 为分量先验概率（混合参数）

若可以估计出参数向量 $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_c)$ ，则可以估计出后验概率：

$$p(\omega_j|\mathbf{x})=\frac{p(\mathbf{x}|\omega_j,\boldsymbol{\theta}_j)P(\omega_j)}{P(\mathbf{x})}$$

混合密度估计和k均值聚类

1. 无监督与有监督
2. 混合密度
3. 最大似然估计
4. 混合正态密度估计
5. k-均值聚类

3、最大似然估计

最大似然估计器:

$$\boxed{P(D|\theta)=\prod_{k=1}^{n} p(x_k|\theta)}$$

上式称为：与数据集合 D 相关的参数向量 θ的最大似然函数。对于可能的参数值 θ，希望寻找出向量 θ'

使得 $P(D|\theta')$ 为似然函数的最大值.

最大似然函数

假设 $D=\{x_1,x_2,\cdots,x_n\}$ 的数据样本为独立同分布的，则

$$\boxed{P\left(D\middle|\theta\right)=\prod_{k=1}^{n}p\left(x_k\middle|\theta\right)}$$

- **注意：**在有监督学习中使用最大似然，其估计所使用的数据集合中的数据是属于同一类的；而在这里，由于所有样本都没有类别标签，所以该最大似然函数是作用在所有数据上的。
- 在本章的无监督学习中，使用的是混合密度：

混合密度 $\Longrightarrow$ $$\boxed{p\left(x_k\middle|\theta\right)=\sum_{j=1}^{c}p\left(x_k\middle|\omega_j,\theta_j\right)P\left(\omega_j\right)}$$

对数似然函数

在实际中，似然函数进行对数运算后，计算比较简单。此时，称为**对数似然函数，**如下:

$$\boxed{l(\theta) \equiv \ln p(D|\theta)}$$

这样 $$l(\theta) \equiv \ln\left(\prod_{k=1}^{n} p(x_k|\theta)\right) = \sum_{k=1}^{n} \ln p(x_k|\theta)$$

由于自然对数函数是**单调增函数**，因此**对数似然函数**和**原似然函数**的**极值点的位置相同**

最优解的必要条件

似然函数的定义导数为:

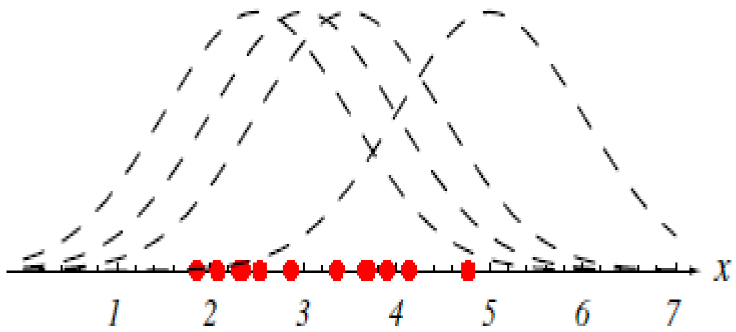
$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(x_k \mid \theta) = \sum_{k=1}^{n} \frac{\nabla_\theta p(x_k \mid \theta)}{p(x_k \mid \theta)}$$

则最优解的必要条件是:

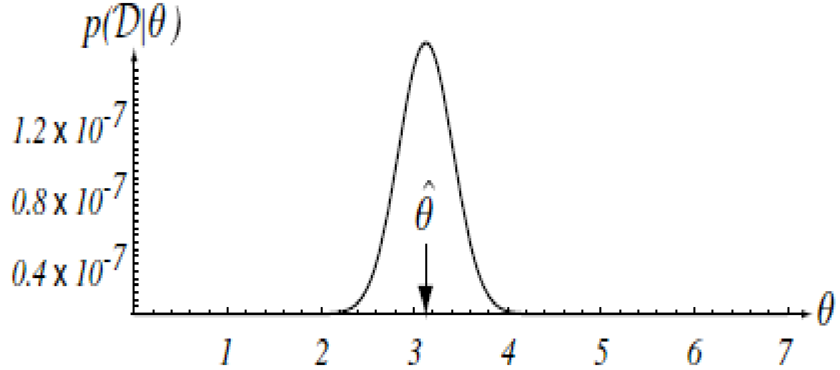
$$\nabla_\theta l = \sum_{k=1}^{n} \frac{\nabla_\theta p(x_k \mid \theta)}{p(x_k \mid \theta)} = 0$$

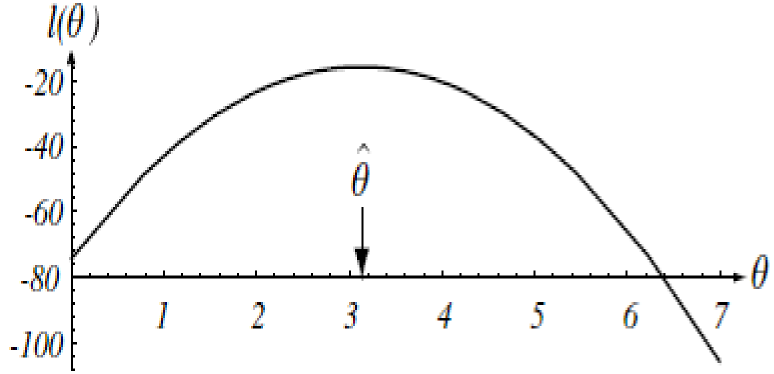
若 θ 由 q 个参数组成，则上式代表 q个方程组成的方程组

- 右上图示出了几个一维的训练样本点，假定它们是从一个方差已知，但均值未知的高斯分布抽样得到的。虚线表示了从源分布中的其中四种可能的分布。



- 右中间的图示出了以均值为变量的似然函数 $p(D|\theta)$ 。若有大量的训练样本点，该似然函数将非常窄。
- 使似然最大的值标记为 $\hat{\theta}$; 该参数也使得右下图中的对数似然 $l(\theta)$ 最大化。





注意：似然函数 $p(D|\theta)$和条件概率密度函数 $p(x|\theta)$很相似，但 $p(D|\theta)$是θ的函数，而 $p(x|\theta)$是以θ为参数的 x 的函数。

似然函数$p(D|\theta)$不是概率密度函数，其曲线下的面积没有实际意义

最大似然估计：求极值点

- 令 $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$， θ_i, θ_j 相互独立($i \neq j$)，D中的每个样本独立且服从混合密度

$$p\left(x_k \middle| \theta\right) = \sum_{j=1}^{c} p\left(x_k \middle| \omega_j, \theta_j\right) P\left(\omega_j\right)$$

令 ∇_θ 是梯度算子，则似然函数对 θ_i 梯度

$$\nabla_{\theta_i} l(\theta) = \sum_{k=1}^{n} \nabla_{\theta_i} \ln p\left(x_k \middle| \theta\right) = \sum_{k=1}^{n} \frac{1}{p(x_k \mid \theta)} \nabla_{\theta_i} \left[\sum_{j=1}^{c} p\left(x_k \middle| \omega_j, \theta_j\right) P\left(\omega_j\right) \right]$$

$$= \sum_{k=1}^{n} \frac{P\left(\omega_i\right)}{p(x_k \mid \theta)} \nabla_{\theta_i} p\left(x_k \middle| \omega_i, \theta_i\right) = \sum_{k=1}^{n} \frac{p\left(x_k \middle| \omega_i, \theta_i\right) P\left(\omega_i\right)}{p(x_k \mid \theta)} \cdot \frac{\nabla_{\theta_i} p\left(x_k \middle| \omega_i, \theta_i\right)}{p\left(x_k \middle| \omega_i, \theta_i\right)}$$

最大似然估计：求极值点（续）

考虑到 $P\left(\omega_i \middle| x_k,\theta\right)=\dfrac{p\left(x_k \middle| \omega_i,\theta_i\right)P\left(\omega_i\right)}{p(x_k\mid\theta)}$ $\qquad \nabla_{\theta_i}\ln p\left(x_k \middle| \omega_i,\theta_i\right)=\dfrac{\nabla_{\theta_i}p\left(x_k \middle| \omega_i,\theta_i\right)}{p\left(x_k \middle| \omega_i,\theta_i\right)}$

则最大似然参数满足

$$\nabla_{\theta_i}l\left(\hat{\theta}\right)=\sum_{k=1}^{n}\hat{P}\left(\omega_i \middle| x_k,\hat{\theta}\right)\nabla_{\theta_i}\ln p\left(x_k \middle| \omega_i,\hat{\theta}_i\right)=0$$

其中

$$\hat{P}\left(\omega_i \middle| x_k,\hat{\theta}\right)=\frac{p\left(x_k \middle| \omega_i,\hat{\theta}_i\right)\hat{P}\left(\omega_i\right)}{p(x_k\mid\hat{\theta})}=\frac{p\left(x_k \middle| \omega_i,\hat{\theta}_i\right)\hat{P}\left(\omega_i\right)}{\sum_{j=1}^{c}p\left(x_k \middle| \omega_j,\hat{\theta}_j\right)\hat{P}\left(\omega_j\right)}$$

且

$$\hat{P}\left(\omega_i\right)=\frac{1}{n}\sum_{k=1}^{n}\hat{P}\left(\omega_i \middle| x_k,\hat{\theta}\right)$$

混合密度估计和k均值聚类

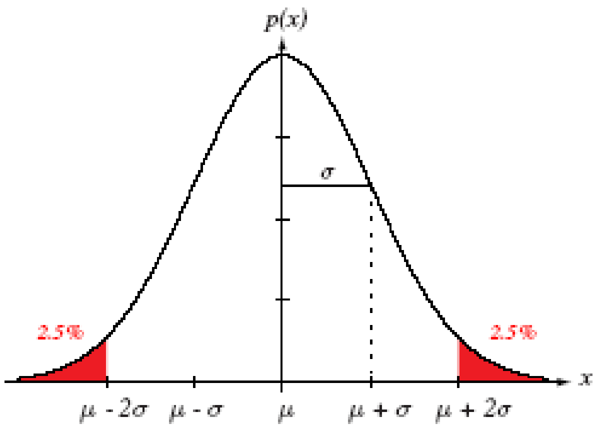
1. 无监督与有监督
2. 混合密度
3. 最大似然估计
4. 混合正态密度估计
5. k-均值聚类

4、混合正态密度估计

- 下面讨论概率密度为混合正态密度的情况
  - 假设每个分量密度都是多元正态分布

$$p\left(\mathbf{x}\middle|\omega_j,\boldsymbol{\theta}_j\right) \rightarrow N\left(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)$$

$$p\left(\mathbf{x}\middle|\omega_j,\boldsymbol{\theta}_j\right)=N\left(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)=\frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}}e^{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)^T\boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)}$$



单变量正态分布情况，大约 95% 的样本分布在 $|x-\mu|\leq 2\sigma$ 区间中

混合正态密度——均值矢量未知

- 在协方差矩阵已知，仅均值矢量未知的情况下：

$$\ln p\left(\mathbf{x}_k \middle| \omega_i, \boldsymbol{\theta}_i\right) = -\left[(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}\right] - \frac{1}{2}\left(\mathbf{x}_k - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_k - \boldsymbol{\mu}_i\right)$$

$$\nabla_{\boldsymbol{\mu}_i} \ln p\left(\mathbf{x}_k \middle| \omega_i, \boldsymbol{\mu}_i\right) = \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_k - \boldsymbol{\mu}_i\right)$$

$$\nabla_{\theta_i} l\left(\hat{\theta}\right) = \sum_{k=1}^{n} \hat{P}\left(\omega_i \middle| x_k, \hat{\theta}\right) \nabla_{\theta_i} \ln p\left(x_k \middle| \omega_i, \hat{\theta}_i\right) = 0$$

$$\sum_{k=1}^{n} \hat{P}\left(\omega_i \middle| x_k, \hat{\theta}\right) \boldsymbol{\Sigma}_i^{-1}\left(x_k - \mu_i\right) = 0$$

混合正态密度——均值矢量未知

- 在仅均值矢量未知的情况下，由最大似然估计可以得到：

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{k=1}^{n} P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}\right) \mathbf{x}_k}{\sum_{k=1}^{n} P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}\right)}$$

$P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}\right)$ 可解释为样本 $\mathrm{x_k}$ 属于ω_j类的可能性。

可以由迭代公式计算：

$$\hat{\boldsymbol{\mu}}_j (i+1) = \frac{\sum_{k=1}^{n} P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}(i)\right) \mathbf{x}_k}{\sum_{k=1}^{n} P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}(i)\right)}$$

若令

$$P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}(i)\right) = \begin{cases} 1, & P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}(i)\right) = \max\limits_i \left\{ P\left(\omega_i \middle| \mathbf{x}_k, \hat{\boldsymbol{\mu}}(i)\right) \right\} \\ 0, & otherwise \end{cases}$$

则均值矢量在第i+1次的迭代结果，可看成是在第i次迭代后属于ω_j类的所有样本 $\mathrm{x_k}$的均值。

混合正态密度——均值、协方差都未知

- 在均值矢量、协方差矩阵都未知的情况下，由最大似然估计可以得到：

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{k=1}^{n} \hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right) \mathbf{x}_k}{\sum_{k=1}^{n} \hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)}$$

$$\hat{P}\left(\omega_j\right) = \frac{1}{n} \sum_{k=1}^{n} \hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{k=1}^{n} \hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right) \left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j\right) \left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_j\right)^T}{\sum_{k=1}^{n} \hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)}$$

$$\boldsymbol{\theta} = \left\{\boldsymbol{\mu}，\ \boldsymbol{\Sigma}\right\}$$

其中：

$$\hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right) = \frac{\hat{P}\left(\mathbf{x}_k \middle| \omega_j, \hat{\boldsymbol{\theta}}_j\right) \hat{P}\left(\omega_j\right)}{\sum_{i=1}^{c} \hat{P}\left(\mathbf{x}_k \middle| \omega_i, \hat{\boldsymbol{\theta}}_i\right) \hat{P}\left(\omega_i\right)}$$

混合正态密度——均值、协方差都未知

令
$$P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)=\begin{cases}1, & P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)=\max_i\left\{P\left(\omega_i \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)\right\}\\ 0, & otherwise\end{cases}$$

$$\boldsymbol{\theta}=\{\boldsymbol{\mu},\ \boldsymbol{\Sigma}\}$$

则
$$\hat{\boldsymbol{\mu}}_j=\frac{1}{n_{D_j}}\sum_{\mathbf{x}_k\in D_j}\mathbf{x}_k \qquad \hat{P}\left(\omega_j\right)=\frac{n_{D_j}}{n}$$

$$\hat{\boldsymbol{\Sigma}}_j=\frac{1}{n_{D_j}}\sum_{\mathbf{x}_k\in D_j}\left(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_j\right)\left(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_j\right)^T$$

其中：
$$\hat{P}\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)\propto\frac{1}{d_M\left(\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j\right)} \qquad d_M\left(\mathbf{x}_k,\boldsymbol{\theta}_j\right)=\left(\mathbf{x}-\boldsymbol{\mu}_j\right)^T\boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)$$

$$P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)=\max_i\left\{P\left(\omega_i \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right)\right\}\Leftrightarrow d_M\left(\mathbf{x}_k,\hat{\boldsymbol{\theta}}_j\right)=\min_i\left\{d_M\left(\mathbf{x}_k,\hat{\boldsymbol{\theta}}_i\right)\right\}$$

混合密度估计和k均值聚类

1. 无监督与有监督
2. 混合密度
3. 最大似然估计
4. 混合正态密度估计
5. k-均值聚类

5、k-均值聚类（c-均值聚类）

- 在均值矢量、协方差矩阵都未知的混合正态密度情况下，若

令协方差矩阵为单位阵： $\boldsymbol{\Sigma} = I$

则马氏距离退化为欧氏距离

$$d_M\left(\mathbf{x}_k, \hat{\boldsymbol{\theta}}_j\right) = \left(\mathbf{x} - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_j\right) = \left\|\mathbf{x} - \boldsymbol{\mu}_j\right\|^2 = d_E\left(\mathbf{x}_k, \boldsymbol{\mu}_j\right)$$

进一步，令 $\hat{P}\left(\omega_j\right) = \frac{1}{c}$

则： $$P\left(\omega_j \middle| \mathbf{x}_k, \hat{\boldsymbol{\theta}}\right) = \begin{cases} 1, & d_E\left(\mathbf{x}_k, \hat{\boldsymbol{\mu}}_j\right) = \min\limits_i \left\{d_E\left(\mathbf{x}_k, \hat{\boldsymbol{\mu}}_i\right)\right\} \\ 0, & otherwise \end{cases}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_{D_j}} \sum\nolimits_{\mathbf{x}_k \in D_j} \mathbf{x}_k$$

通过这两个公式迭代估计均值参数的方法称为：**c-均值聚类**

c-均值聚类的目标函数

c-均值聚类也可以等价描述为优化下面的目标函数的过程，

$$J_1(U,V)=\sum_{i=1}^{c}\left(\sum_{x\in S_i}(d_{ik})^2\right)$$

式中，

$$d_{ik}=d(x_k,v_i)=\|x_k-v_i\|=\left[\sum_{j=1}^{p}(x_{kj}-v_{ij})^2\right]^{1/2}$$

C-均值聚类的迭代公式

$$u_{ik} = \begin{cases} 1; & d_{ik} = \min_{1 \le j \le c} \{d_{jk}\} \\ 0; & otherwise \end{cases} \qquad 1 \le i \le c \qquad 1 \le k \le n$$

$$v_i = \sum_{k=1}^{n} u_{ik} x_k \Big/ \sum_{k=1}^{n} u_{ik} \qquad 1 \le i \le c$$

C-均值聚类的具体迭代算法

1） 确定聚类类别数 c，$2 \le c < n$，n 是数据个数。

2） 初始化分类矩阵$U^{(0)}$。设置聚类停止误差$\varepsilon > 0$，令$t = 1$

3） 根据$U^{(t-1)}$计算 C 均值（聚类中心）矢量$v_i^{(t)}$，$1 \le i \le c$。

4） 用 C 均值矢量$v_i^{(t)}$，$1 \le i \le c$计算新的划分矩阵$U^{(t)}$。

5） 以一个合适的矩阵范数比较$U^{(t)}$和$U^{(t-1)}$，若$\left\| U^{(t)} - U^{(t-1)} \right\| < \varepsilon$，

停止；否则置 t=t+1，返回步骤 3）。

模糊c-均值聚类（FCM）

模糊c-均值聚类目标函数，

$$J_1(U,V)=\sum_{i=1}^{c}\left(\sum_{x\in S_i}(u_{ik})^m(d_{ik})^2\right)$$

式中，

$$d_{ik}=d(x_k,v_i)=\|x_k-v_i\|=\left[\sum_{j=1}^{p}(x_{kj}-v_{ij})^2\right]^{1/2}$$

模糊C-均值聚类的迭代公式

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right]^{-1} \qquad 1 \le i \le c \quad 1 \le k \le n$$

$$v_i = \sum_{k=1}^{n} (u_{ik})^m x_k \Big/ \sum_{k=1}^{n} (u_{ik})^m \qquad 1 \le i \le c$$

模糊C-均值聚类的具体迭代算法

设 $X=\{x_1,x_2,\cdots,x_n\}\in R^p$ 为数据集合，n 为数据项数。整数 c 为类别数，$2\le c<n$。

$V=\{v_1,v_2,\cdots,v_c\}\in R^p$ 为 c 个聚类中心的集合。

1）确定加权指数：$1\le m<\infty$；确定聚类停止误差 $\varepsilon>0$。

2）初始化模糊划分矩阵 $U^{(0)}$，令 t=1。

3）根据 $U^{(t-1)}$ 计算 C 均值矢量 $v_i^{(t)}$，$1\le i\le c$。

4）用 C 均值矢量 $v_i^{(t)}$，$1\le i\le c$ 计算新的划分矩阵 $U^{(t)}$：

5）以一个合适的矩阵范数比较 $U^{(t)}$ 和 $U^{(t-1)}$，若 $\left\|U^{(t)}-U^{(t-1)}\right\|<\varepsilon$，停止；否则置 t=t+1

返回步骤 3）。