PART5 聚类

基本内容：

1. 什么是聚类？什么是分类？二者的区别与联系。

2. **样本点之间**欧式距离、马氏距离、切比雪夫距离、曼哈顿距离？

   样本点与簇、**簇与簇**之间的距离计算；

   最小距离、最远距离、平均距离法度量簇间距离。

3. 掌握 K-Means Clustering 算法，影响该算法性能的因素有哪些？

4. 以聚合式系统聚类为例，掌握系统聚类。
5. 理解基于高斯混合模型进行聚类的基本流程，了解什么是高斯混合模型？
6. 以 DBSCAN 算法为例，理解密度聚类实现的基本流程，掌握有关概念，哪些因素会影响聚类的效果。

7. 理解聚类算法有关性能的评价指标。

练习：

1. 聚类. 给定数据集$\boldsymbol{D}=\{\boldsymbol{x_i}, i=1,\dots,m\}$，其中$\boldsymbol{x_i} \in R^d$.若采用 K-均值聚类算法将该数据集$\boldsymbol{D}$划分为 K 簇$\{C_1,\cdots,C_K\}$，请完成如下工作：

(1) 写出 K-均值聚类算法的准则函数及其意义，并给出必要参数说明；

(2) 对 K-均值聚类算法的实现过程进行描述.

(3) 哪些因素会影响 K-均值聚类的性能？

(4) 如何弱化这些因素的不良影响？

答：

(1) 准则函数---总的簇内误差平方和最小

$$E\left(\mu_1,...,\mu_k,C_1,...,C_k\right)=\sum_{i=1}^{k}\sum_{x\in C_i}\left\|x-\mu_i\right\|^2$$

C_i --第i簇, $i=1,...,k$

μ_i --第i簇的中心， $\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$

(2)

输入： 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
聚类簇数 k.
过程：
从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_k\}$
repeat
  令 $C_i=\varnothing\ (1\leqslant i\leqslant k)$
  for $j=1,2,\ldots,m$ **do**
    计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i\ (1\leqslant i\leqslant k)$ 的距离: $d_{ji}=\|\boldsymbol{x}_j-\boldsymbol{\mu}_i\|_2$;
    根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j=\arg\min_{i\in\{1,2,\ldots,k\}}d_{ji}$;
    将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j}=C_{\lambda_j}\bigcup\{\boldsymbol{x}_j\}$;
  end for
  for $i=1,2,\ldots,k$ **do**
    计算新均值向量: $\boldsymbol{\mu}_i'=\frac{1}{|C_i|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}$;
    if $\boldsymbol{\mu}_i'\neq\boldsymbol{\mu}_i$ **then**
      将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}_i'$
    else
      保持当前均值向量不变
    end if
  end for
until 当前均值向量均未更新
输出： 簇划分 $\mathcal{C}=\{C_1,C_2,\ldots,C_k\}$

(2) 样本集是否规范化预处理、距离度量方式的不同、K 值的大小、以及初始化方式均会影响 K 均值聚类性能。

2. 若要采用合并式层次聚类(也称：聚合式系统聚类)将样本集 $\boldsymbol{D}=\{\boldsymbol{x}_i,\mathrm{i}=1,\cdots,m\}$ 划分为K簇，其中 $\boldsymbol{x}_i\in R^d$. 请完成如下工作：

(1)请对该聚类算法的实现流程予以描述.

(2)上述聚类过程中，需要进行不同聚类簇之间的距离计算，请分别采用最近距离、最远距离，估算任意两簇 $\boldsymbol{C}_j,\boldsymbol{C}_l$ 之间的距离.

解：

AGNES算法描述

STEP1.初始化聚类簇

STEP2.初始化距离矩阵

STEP3.初始化聚类簇的数目

STEP4.逐级合并最相似的两簇，直到满足规定的聚类簇数目：
(1) 合并最相似的两簇；
(2) 更新相关簇的序号；
(3) 更新相关距离矩阵；
(4) 更新簇的数目

STEP5.输出结果。

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,...,m do
    C_j = {x_j}
end for
for i = 1,2,...,m do
    for j = 1,2,...,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* ⋃ C_j*;
    for j = j* + 1, j* + 2, ..., 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,...,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 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

(1)

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,...,m do
    C_j = {x_j}
end for
for i = 1,2,...,m do
    for j = 1,2,...,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* ⋃ C_j*;
    for j = j* + 1, j* + 2, ..., 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,...,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 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

(2)

最近距离 $\boldsymbol{d}_{\min}\left(\boldsymbol{C}_j,\boldsymbol{C}_l\right)=\min_{\substack{x\in C_j\\ z\in C_l}}\boldsymbol{dist}\left(x,z\right)$

最远距离 $\boldsymbol{d}_{\max}\left(\boldsymbol{C}_j,\boldsymbol{C}_l\right)=\max_{\substack{x\in C_j\\ z\in C_l}}\boldsymbol{dist}\left(x,z\right)$

3. DBSCAN是一种基于密度的聚类算法，若要基于该算法,对观测点集 $\boldsymbol{D}=\left\{\boldsymbol{x}_i,\mathrm{i}=1,\cdots,N\right\}$ 进行聚类，需要提供两个全局参数$\left(\varepsilon,MinPts\right)$，回答如下问题：

(1)两个参数的意义是什么？如何确定核心对象？边界对象？噪声对象？什么是密度直达？什么是密度可达？密度相连？

(2)理解DBSCAN算法的实现流程。

解：

(1)两个**全局邻域参数**$\left(\varepsilon,\boldsymbol{MinPts}\right)$

ε**--**邻域最大半径

***MinPts*--**给定样本的$\varepsilon-$**邻域**内最小样本数.

其中：

$\varepsilon-$**邻域** 对于$\forall\boldsymbol{x}_j\in\boldsymbol{D}$， $\boldsymbol{x}_j$的$\varepsilon-$邻域为$\boldsymbol{N}_\varepsilon\left(\boldsymbol{x}_j\right)=\left\{\boldsymbol{x}_i\in\boldsymbol{D}\mid\boldsymbol{dist}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\le\varepsilon\right\}$

(2)**核心对象**$\left(\boldsymbol{core\ \ object}\right)$

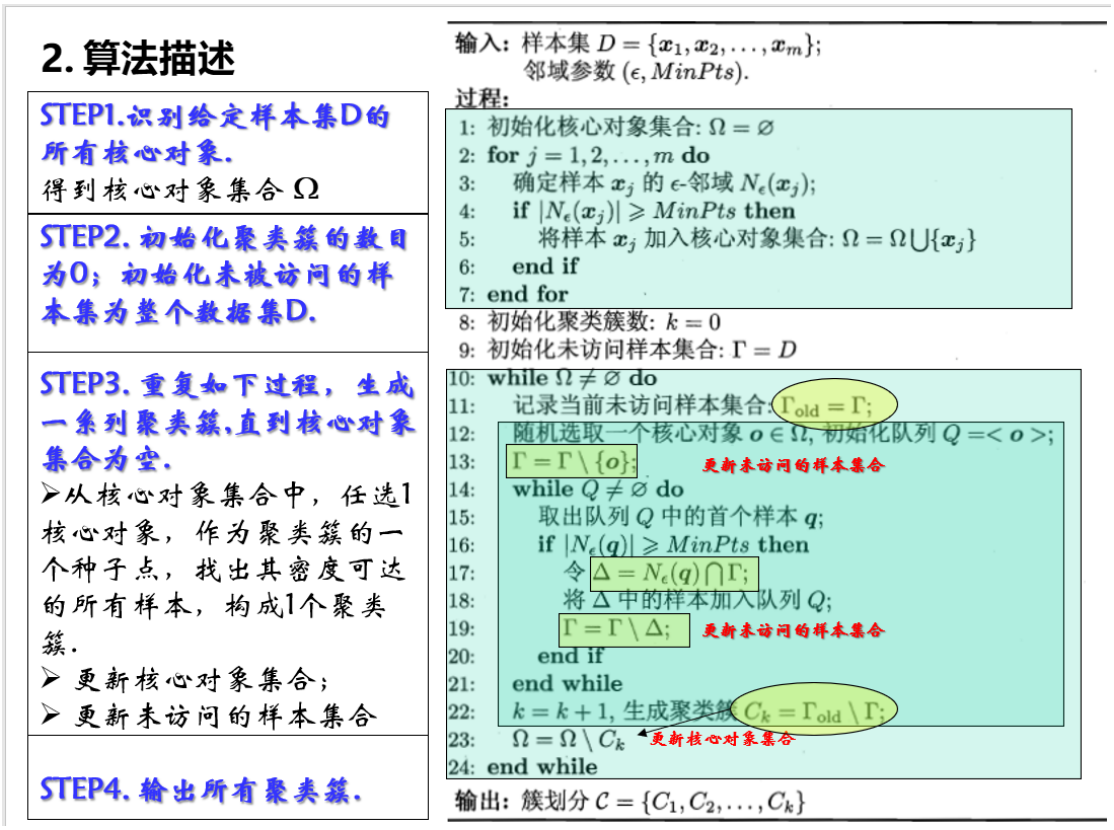
若$\boldsymbol{x}_j\in\boldsymbol{D}$并且$\left|\boldsymbol{N}_\varepsilon\left(\boldsymbol{x}_j\right)\right|\ge\boldsymbol{MinPts}$，则称$\boldsymbol{x}_j$为一个**核心对象**.

(3)**密度直达**$\left(\boldsymbol{directly\ \ density-reacheable}\right)$

若$\boldsymbol{x}_j\in\boldsymbol{N}_\varepsilon\left(\boldsymbol{x}_i\right)$,并且$\boldsymbol{x}_i$为一个核心对象，则称$\boldsymbol{x}_j$为由$\boldsymbol{x}_i$**密度直达**.

(4)**密度可达**$\left(\boldsymbol{density-reacheable}\right)$

对于$\boldsymbol{x}_i,\boldsymbol{x}_j$，若存在样本序列$\boldsymbol{p}_1,\boldsymbol{p}_2,...,\boldsymbol{p}_n$,其中$\boldsymbol{p}_1=\boldsymbol{x}_i,\boldsymbol{p}_n=\boldsymbol{x}_j$,且$\boldsymbol{p}_{i+1}$由$\boldsymbol{p}_i$密度直达，则称$\boldsymbol{x}_j$由$\boldsymbol{x}_i$**密度可达**.



4. 给定数据集$\boldsymbol{D}=\{\boldsymbol{x_i},\boldsymbol{i}=\boldsymbol{1},\ldots,\boldsymbol{m}\}$，其中$\boldsymbol{x_i}\in\boldsymbol{R^d}$，若采用高斯混合模型将数据集$\boldsymbol{D}$划分为K簇$\{\boldsymbol{C_1},\cdots,\boldsymbol{C_K}\}$.按要求完成如下工作：

(1)写出高斯混合概率密度函数的具体表达形式，并指出该模型各参数的意义.

(2)若采用EM算法基于该数据集$\boldsymbol{D}$已经完成了高斯混合模型的参数估计，如何由高斯混合模型得到关于数据集D的最终划分结果？

解：

(1)

$$p_M(x)=\sum_{j=1}^{K}\alpha_j p(x\mid\mu_j,\Sigma_j)=\sum_{j=1}^{K}\frac{\alpha_j}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right]$$

其中

- K--高斯成分的数目
- α_j--第j个高斯成分的混合系数　　$\sum_{j=1}^{K}\alpha_j=1$
- $p(x\mid\mu_j,\Sigma_j)$--第j个高斯成分的概率密度函数

$$p(x\mid\mu_j,\Sigma_j)=\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right]$$

- μ_j,Σ_j--第j个高斯成分的期望向量、协方差矩阵
- 待估计的**参数集合**：$\left\{(\alpha_j,\mu_j,\Sigma_j)\mid j=1,...,K\right\}$

(2)设参数估计结果 $\left\{(\hat{\alpha}_j,\hat{\mu}_j,\hat{\Sigma}_j)\mid j=1,...,K\right\}$，则

对于数据集**D**的观测样本$x_i, i=1,....,m$

其相应聚类簇的划分结果为:

$$\lambda_i=\arg\max_{j\in\{1,2,...,K\}}\hat{\alpha}_j\,p(x_i\mid\hat{\mu}_j,\hat{\Sigma}_j)\qquad i=1,....,m$$

(3)其它：理解模型学习中的交叉迭代过程：

GMM模型参数估计的交叉迭代

混合高斯模型 $p_M(x)=\sum_{i=1}^{k}\alpha_i p(x\mid\mu_i,\Sigma_i)$

(1)**初始化** $\hat{\mu}_i(0),\hat{\Sigma}_i(0),\hat{\alpha}_i(0),i=1,...,k$

(2)**交叉迭代**直到满足终止条件：

$$E-STEP\quad \gamma_{ji}(n+1)=\frac{\alpha_i(n)\,p(x_j\mid\mu_i(n),\Sigma_i(n))}{\sum_{l=1}^{k}\alpha_l(n)\,p(x\mid\mu_l(n),\Sigma_l(n))}\qquad \begin{cases} i=1,...,k \\ j=1,...,m \end{cases}$$

$$\gamma_{ji}=P_M(z_j=i\mid x_j)$$

$$M-STEP\begin{cases}\hat{\mu}_i(n+1)=\dfrac{\sum_{j=1}^{m}\gamma_{ji}(n+1)x_j}{\sum_{j=1}^{m}\gamma_{ji}(n+1)} & i=1,...,k \\ \hat{\Sigma}_i(n+1)=\dfrac{\sum_{j=1}^{m}\gamma_{ji}(n+1)\left(x_j-\hat{\mu}_i(n+1)\right)\left(x_j-\hat{\mu}_i(n+1)\right)^T}{\sum_{j=1}^{m}\gamma_{ji}(n+1)} & i=1,...,k \\ \hat{\alpha}_i(n+1)=\dfrac{1}{m}\sum_{j=1}^{m}\gamma_{ji}(n+1) & i=1,...,k\end{cases}$$