南昌大学
NANCHANG UNIVERSITY

统计机器学习

主讲人：彭振华

数学与计算机学院

2026年

目录

CONTENTS

01. 机器学习基础
02. 线性模型
03. 决策树
04. 支持向量机
05. 神经网络基础
06. 贝叶斯分类器
07. 集成学习
08. 聚类
09. 降维与度量学习
10. 特征选择与稀疏学习
11. 概率图模型

聚类

- 聚类（clustering）是将样本集合中相似的样本（实例）分配到相同的类，不相似的样本分配到不同的类。
- 聚类时，样本通常是欧氏空间中的向量，类别不是事先给定， 而是从数据中自动发现，但类别个数通常是事先给定的。样本之间的相似度或距离 由应用决定。
- 如果一个样本只能属于一个类，则称为硬聚类（hard clustering)

$$z_i = g_\theta(x_i), i = 1, 2, \cdots, N$$

- 如果一个样本可以属于多个类，则称为软聚类（soft clustering)

$$P_\theta(z_i|x_i), i = 1, 2, \cdots, N.$$

- 希望“物以类聚”，即同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同。换言之，聚类结果的“簇内相似度”（intra-cluster similarity）高，且“簇间相似度”（inter-cluster similarity）低，这样的聚类效果较好.

聚类

- 聚类性能度量：
  - 外部指标 (external index)

    将聚类结果与某个“参考模型”(reference model)进行比较。
  - 内部指标 (internal index)

    直接考察聚类结果而不用任何参考模型。

对数据集 $D=\{x_1,x_2,\ldots,x_m\}$，假定通过聚类得到的簇划分为 $C=\{C_1,C_2,\ldots,C_k\}$ 参考模型给出的簇划分为 $C^*=\{C_1^*,C_2^*,\ldots,C_s^*\}$.相应地，令 λ 与 λ^* 分别表示与 C 和 C^* 对应的簇标记向量.

将样本两两配对考虑，定义

$$a=|SS|,SS=\{(x_i,x_j)|\lambda_i=\lambda_j,\lambda_i^*=\lambda_j^*,i<j\}$$
$$b=|SD|,SD=\{(x_i,x_j)|\lambda_i=\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\}$$
$$c=|DS|,DS=\{(x_i,x_j)|\lambda_i\neq\lambda_j,\lambda_i^*=\lambda_j^*,i<j\}$$
$$d=|DD|,DD=\{(x_i,x_j)|\lambda_i\neq\lambda_j,\lambda_i^*\neq\lambda_j^*,i<j\}$$

聚类

- Jaccard系数（Jaccard Coefficient, JC）

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

- FM指数（Fowlkes and Mallows Index, FMI）

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

- Rand指数（Rand Index, RI）

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

[0,1]区间内，越大越好.

- 考虑聚类结果的簇划分 $C = \{C_1, C_2, \ldots, C_k\}$ ，定义

簇 C 内样本间的平均距离

$$avg(C) = \frac{2}{|C|(|C|-1)} \sum_{1 \le i \le j \le |C|} dist(x_i, x_j)$$

簇 C 内样本间的最远距离

$$diam(C) = max_{1 \le i \le j \le |C|} dist(x_i, x_j)$$

簇 C_i 与簇 C_j 最近样本间的距离

$$d_{min}(C) = min_{x_i \in C_i, x_j \in C_j} dist(x_i, x_j)$$

簇 C_i 与簇 C_j 中心点间的距离

$$d_{cen}(C) = dist(\mu_i, \mu_j)$$

- DB指数（Davies-Bouldin Index, DBI）

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{cen}(C_i,C_j)}\right)$$

越小越好.

- Dunn指数（Dunn Index, DI）

$$DI = \min_{1\leq i\leq k}\left\{\min_{j\neq i}\left(\frac{d_{min}(C_i,C_j)}{\max_{1\leq l\leq k} diam(C_l)}\right)\right\}$$

越大越好.

聚类

- 假设有n个样本，每个样本由m个属性的特征向量组成，样本合集可以用矩阵X表示

$$X = [x_{ij}]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

- 聚类的核心概念是相似度（similarity）或距离（distance），有多种相似度或距离定义。因为相似度直接影响聚类的结果，所以其选择是聚类的根本问题。

聚类

- 闵可夫斯基距离越大相似度越小，距离越小相似度越大。
- 给定样本集合X，X是m维实数向量空间R^m中点的集合，其中

$$x_i, x_j \in X,\ x_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}},\ x_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}}$$

- 样本x_i与样本x_j 的闵可夫斯基距离（Minkowski distance）定义为

$$d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p \right)^{\frac{1}{p}} \qquad p \geqslant 1$$

聚类

- 当p=2时称为欧氏距离（Euclidean distance)

$$d_{ij}=\left(\sum_{k=1}^{m}|x_{ki}-x_{kj}|^2\right)^{\frac{1}{2}}$$

- 当p=1时称为曼哈顿距离（Manhattan distance)

$$d_{ij}=\sum_{k=1}^{m}|x_{ki}-x_{kj}|$$

- 当p= ∞ 时称为切比雪夫距离（Chebyshev distance)

$$d_{ij}=\max_{k}|x_{ki}-x_{kj}|$$

聚类

- 马哈拉诺比斯距离（Mahalanobis distance），简称马氏距离，也是另一种常用的相似度，考虑各个分量（特征）之间的相关性并与各个分量的尺度无关。
- 马哈拉诺比斯距离越大相似度越小，距离越小相似度越大。
- 给定一个样本集合X，$X = [x_{ij}]_{m\times n}$ ，其协方差矩阵记作S。样本x_i 与样本x_j之间的马哈拉诺比斯距离d_{ij}定义为

$$d_{ij} = \left[(x_i - x_j)^{\mathrm{T}} S^{-1} (x_i - x_j)\right]^{\frac{1}{2}}$$

$$x_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}, \quad x_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}}$$

- 样本之间的相似度也可以用相关系数（correlation coefficient）来表示。
- 相关系数的绝对值越接近于1，表示样本越相似
- 越接近于0，表示样本越不相似。
- 样本x_i与样本x_j之间的相关系数定义为

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\left[\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{m}(x_{kj}-\bar{x}_j)^2\right]^{\frac{1}{2}}}$$

$$\bar{x}_i=\frac{1}{m}\sum_{k=1}^{m}x_{ki},\quad \bar{x}_j=\frac{1}{m}\sum_{k=1}^{m}x_{kj}$$

- 样本之间的相似度也可以用夹角余弦（cosine）来表示。
- 夹角余弦越接近于1，表示样本越相似
- 越接近于0，表示样本越不相似。
- 样本x_i与样本x_j之间的夹角余弦定义为

$$s_{ij} = \frac{\sum_{k=1}^{m} x_{ki}x_{kj}}{\left[\sum_{k=1}^{m} x_{ki}^2 \sum_{k=1}^{m} x_{kj}^2\right]^{\frac{1}{2}}}$$

- Value Difference Metric, VDM（处理无序属性）：

令 $m_{u,a}$ 表示属性 u 上取值为 a 的样本数， $m_{u,a,i}$ 表示在第 i 个样本簇中在属性 u 上取值为 a 的样本数，k 为样本数，则属性 u 上两个离散值 a 与 b 之间的VDM距离为

$$VDM_p(a,b) = \sum_{i=1}^{k} \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p$$

- MinkovDMp（处理混合属性）：

$$MinkovDM_p(x_i, x_j) = \left(\sum_{u=1}^{n_c} |x_{iu} - x_{ju}|^p + \sum_{u=n_c+1}^{n} VDM_p(x_{iu}, x_{ju}) \right)^{\frac{1}{p}}$$

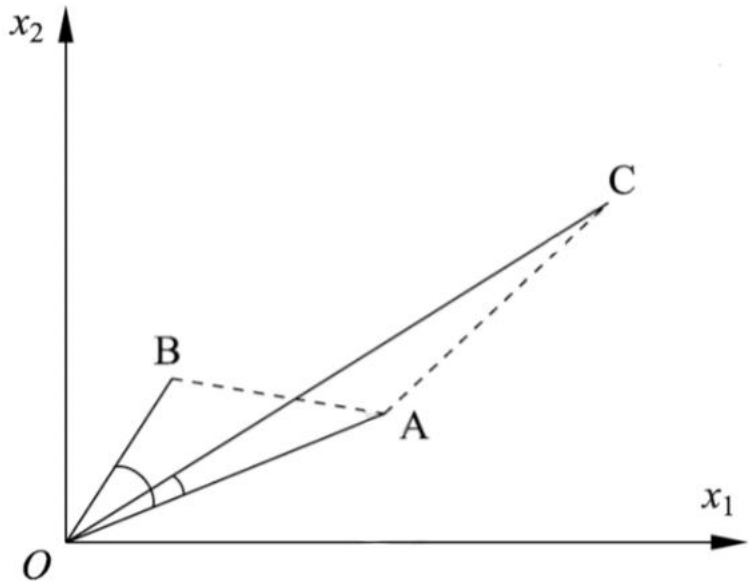
- 加权距离（样本中不同属性的重要性不同时）：

$$dist(x_i, x_j) = (\omega_1 \cdot |x_{i1} - x_{j1}|^p + \cdots + \omega_n \cdot |x_{in} - x_{jn}|^p)^{\frac{1}{p}} \qquad \omega_i \geq 0, \sum_{i=1}^{n} \omega_i = 1$$

聚类

- 用距离度量相似度时，距离越小样本越相似
- 用相关系数时，相关系数越大样本越相似
- 注意不同相似度度量得到的结果并不一定一致。

- 从右图可以看出，如果从距离的角度看，
  A和B比A和C更相似
- 但从相关系数的角度看，
  - A和C比A和B更相似。



聚类

- 通过聚类得到的类或簇，本质是样本的子集。
- 如果一个聚类方法假定一个样本只能属于一个类，或类的交集为空集，那么该方法称为硬聚类（hard clustering）方法。
- 如果一个样本可以属于多个类，或类的交集不为空集，那么该方法称为软聚类（soft clustering）方法。
- 用G表示类或簇（cluster)，用xi, xj表示类中的样本，用nG表示G中样本的个数，用dij表示样本xi与样本xj之间的距离。
- 类或簇有多种定义，下面给出几个常见的定义。

设 T 为给定的正数，若集合 G 中任意两个样本 x_i, x_j，有 $d_{ij} \leqslant T$ 则称 G 为一个类或簇。

设 T 为给定的正数，若对集合 G 的任意样本 x_i，一定存在 G 中的另一个样本 x_j，使得 $d_{ij} \leqslant T$ 则称 G 为一个类或簇。

设 T 为给定的正数，若对集合 G 中任意一个样本 x_i，G 中的另一个样本 x_j 满足 $\frac{1}{n_G - 1} \sum_{x_j \in G} d_{ij} \leqslant T$ 其中 n_G 为 G 中样本的个数，则称 G 为一个类或簇。

设 T 和 V 为给定的两个正数，如果集合 G 中任意两个样本 x_i, x_j 的距离 d_{ij} 满足

$$\frac{1}{n_G(n_G-1)}\sum_{x_i\in G}\sum_{x_j\in G} d_{ij} \leqslant T$$

$$d_{ij} \leqslant V$$

则称 G 为一个类或簇。

- 类的特征可以通过不同角度来刻画，常用的特征有下面三种：

（1）类的均值 $\bar{x}_G$，又称为类的中心

$$\bar{x}_G = \frac{1}{n_G}\sum_{i=1}^{n_G} x_i$$

式中 n_G 是类 G 的样本个数。

（2）类的直径（diameter）D_G

类的直径 D_G 是类中任意两个样本之间的最大距离，即

$$D_G = \max_{x_i, x_j \in G} d_{ij}$$

（3）类的样本散布矩阵（scatter matrix）A_G 与样本协方差矩阵（covariance matrix）S_G

类的样本散布矩阵 A_G 为 $A_G = \sum_{i=1}^{n_G} (x_i - \bar{x}_G)(x_i - \bar{x}_G)^{\mathrm{T}}$

样本协方差矩阵 S_G 为

$$S_G = \frac{1}{m-1} A_G$$

$$= \frac{1}{m-1} \sum_{i=1}^{n_G} (x_i - \bar{x}_G)(x_i - \bar{x}_G)^{\mathrm{T}}$$

其中 m 为样本的维数（样本属性的个数）。

聚类

- 下面考虑类G_p与类G_q之间的距离D(p,q)，也称为连接（linkage)。类与类之间的距离也有多种定义。
- 设类G_p包含 n_p个样本，G_q包含 n_q个样本，分别用 $\bar{x}_p$ 和 $\bar{x}_q$ 表示G_p和G_q的均值，即类的中心 。
- 最短距离或单连接（single linkage)
- 定义类G_p的样本与G_q的样本之间的最短距离为两类之间的距离

$$D_{pq} = \min\{d_{ij} | x_i \in G_p, x_j \in G_q\}$$

聚类

- 最长距离或完全连接（complete linkage）
- 定义类G_p的样本与G_q的样本之间的最长距离为两类之间的距离

$$D_{pq} = \max\{d_{ij} | x_i \in G_p, x_j \in G_q\}$$

- 中心距离
- 定义类G_p与G_q的中心 $\bar{x}_p$ 与 $\bar{x}_q$ 之间的距离为两类之间的距离 $D_{pq} = d_{\bar{x}_p \bar{x}_q}$
- 平均距离
- 定义类G_p与G_q任意两个样本之间距离的平均值为两类之间的距离

$$D_{pq} = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}$$

聚类

- 层次聚类假设类别之间存在层次结构，将样本聚到层次化的类中。
- 层次聚类又有聚合（agglomerative）或自下而上（bottom-up）聚类、分裂（divisive）或自上而下（top-down）聚类两种方法。
- 因为每个样本只属于一个类，所以层次聚类属于硬聚类
- 聚合聚类开始将每个样本各自分到一个类，之后将相距最近的两类合并，建立一个新的类，重复此操作直到满足停止条件，得到层次化的类别。
- 分裂聚类开始将所有样本分到一个类，之后将已有类中相距最远的样本分到两个新的类，重复此操作直到满足停止条件，得到层次化的类别。

输入：n 个样本组成的样本集合及样本之间的距离；

输出：对样本集合的一个层次化聚类。

（1）计算 n 个样本两两之间的欧氏距离 $\{d_{ij}\}$，记作矩阵 $D=[d_{ij}]_{n\times n}$。

（2）构造 n 个类，每个类只包含一个样本。

（3）合并类间距离最小的两个类，其中最短距离为类间距离，构建一个新类。

（4）计算新类与当前各类的距离。若类的个数为 1，终止计算，否则回到步 (3)。■

可以看出聚合层次聚类算法的复杂度是 $O(n^3m)$，其中 m 是样本的维数，n 是样本个数。

聚类

- 给定5个样本的集合，样本之间的欧氏距离由如下矩阵D表示

$$D = [d_{ij}]_{5\times 5} = \begin{bmatrix} 0 & 7 & 2 & 9 & 3 \\ 7 & 0 & 5 & 4 & 6 \\ 2 & 5 & 0 & 8 & 1 \\ 9 & 4 & 8 & 0 & 5 \\ 3 & 6 & 1 & 5 & 0 \end{bmatrix}$$

- 其中d_{ij}表示第i个样本与第j个样本之间的欧氏距离。
- 显然D为对称矩阵。应用聚合层次聚类法对这5个样本进行聚类。

聚类

- （1）
- 首先用5个样本构建5个类，$G_i = \{x_i\}, i = 1, 2, \cdots, 5,$
- 这样，样本之间的距离也就变成类之间的距离，所以5个类之间的距离矩阵亦为D
- （2）
- 由矩阵D可以看出，$D_{35} = D_{53} = 1$ 为最小，所以把G_3和G_5合并为一 个新类，记作 $G_6 = \{x_3, x_5\}$.

- （3）
- 计算G_6与G_1，G_2，G_4之间的最短距离，有

$$D_{61}=2,\quad D_{62}=5,\quad D_{64}=5$$

- 又注意到其余两类之间的距离是

$$D_{12}=7,\quad D_{14}=9,\quad D_{24}=4$$

- 显然，D_{61}=2最小，所以将G_1与G_6合并成一个新类，记作 $G_7=\{x_1,x_3,x_5\}$。

- （4）
- 计算G_7与G_2，G_4之间的最短距离，

$$D_{72} = 5, \quad D_{74} = 5$$

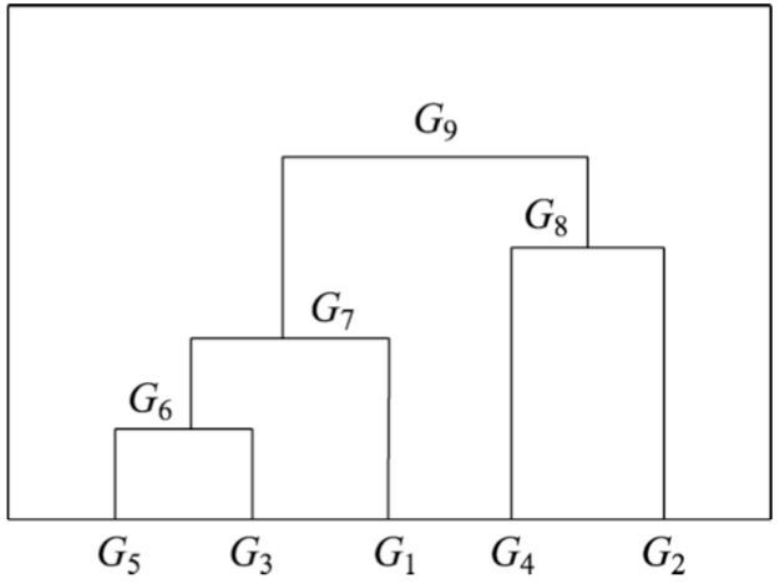
- 又注意到

$$D_{24} = 4$$

- 显然，其中D_{24}=4最小，所以将G_2与G_4合并成一个新类，记作 $G_8 = \{x_2, x_4\}$

聚类

- （5）
- 将G_7与G_8合并成一个新类，记作 $G_9 = \{x_1, x_2, x_3, x_4, x_5\}$
- 即将全部样本聚成1类，聚类终止



聚类

- k均值聚类是基于样本集合划分的聚类算法。
- k均值聚类将样本集合划分为k个子集，构成k个类，将n个样本分到k个类中，每个样本到其所属类的中心的距离最小。
- 每个样本只能属于一个类，所以k均值聚类是硬聚类。

给定数据集$D=\{x_1,x_2,\cdots,x_m\}$，*k*均值算法针对聚类所得簇划分$C=\{C_1,C_2,\cdots,C_k\}$最小化平方误差

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2$$

其中，μ_i是簇C_i的均值向量。

E值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。

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$C_i = \emptyset \ (1 \leq i \leq k)$
  for $j = 1, \ldots, m$ **do**
    计算样本$\boldsymbol{x}_j$与各均值向量$\boldsymbol{\mu}_i \ (1 \leq i \leq k)$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 $i = 1, \ldots, k$ **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
return 簇划分结果

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

- *k*均值算法实例

接下来以表9-1的西瓜数据集4.0为例，来演示*k*均值算法的学习过程。将编号为 i 的样本称为 x_i .

编号	密度	含糖率	编号	密度	含糖率	编号	密度	含糖率
1	0.697	0.460	11	0.245	0.057	21	0.748	0.232
2	0.774	0.376	12	0.343	0.099	22	0.714	0.346
3	0.634	0.264	13	0.639	0.161	23	0.483	0.312
4	0.608	0.318	14	0.657	0.198	24	0.478	0.437
5	0.556	0.215	15	0.360	0.370	25	0.525	0.369
6	0.403	0.237	16	0.593	0.042	26	0.751	0.489
7	0.481	0.149	17	0.719	0.103	27	0.532	0.472
8	0.437	0.211	18	0.359	0.188	28	0.473	0.376
9	0.666	0.091	19	0.339	0.241	29	0.725	0.445
10	0.243	0.267	20	0.282	0.257	30	0.446	0.459

- *k*均值算法实例

假定聚类簇数*k* =3，算法开始时，随机选择3个样本 x_6, x_{12}, x_{24} 作为初始均值向量，即 $\mu_1 = (0.403; 0.237), \mu_2 = (0.343; 0.099), \mu_3 = (0.478; 0.437)$ 。

考察样本$x_1 = (0.697; 0.460)$，它与当前均值向量 μ_1, μ_2, μ_3 的距离分别为0.369，0.506，0.220，因此 x_1 将被划入簇C_3 中。类似的，对数据集中的所有样本考察一遍后，可得当前簇划分为

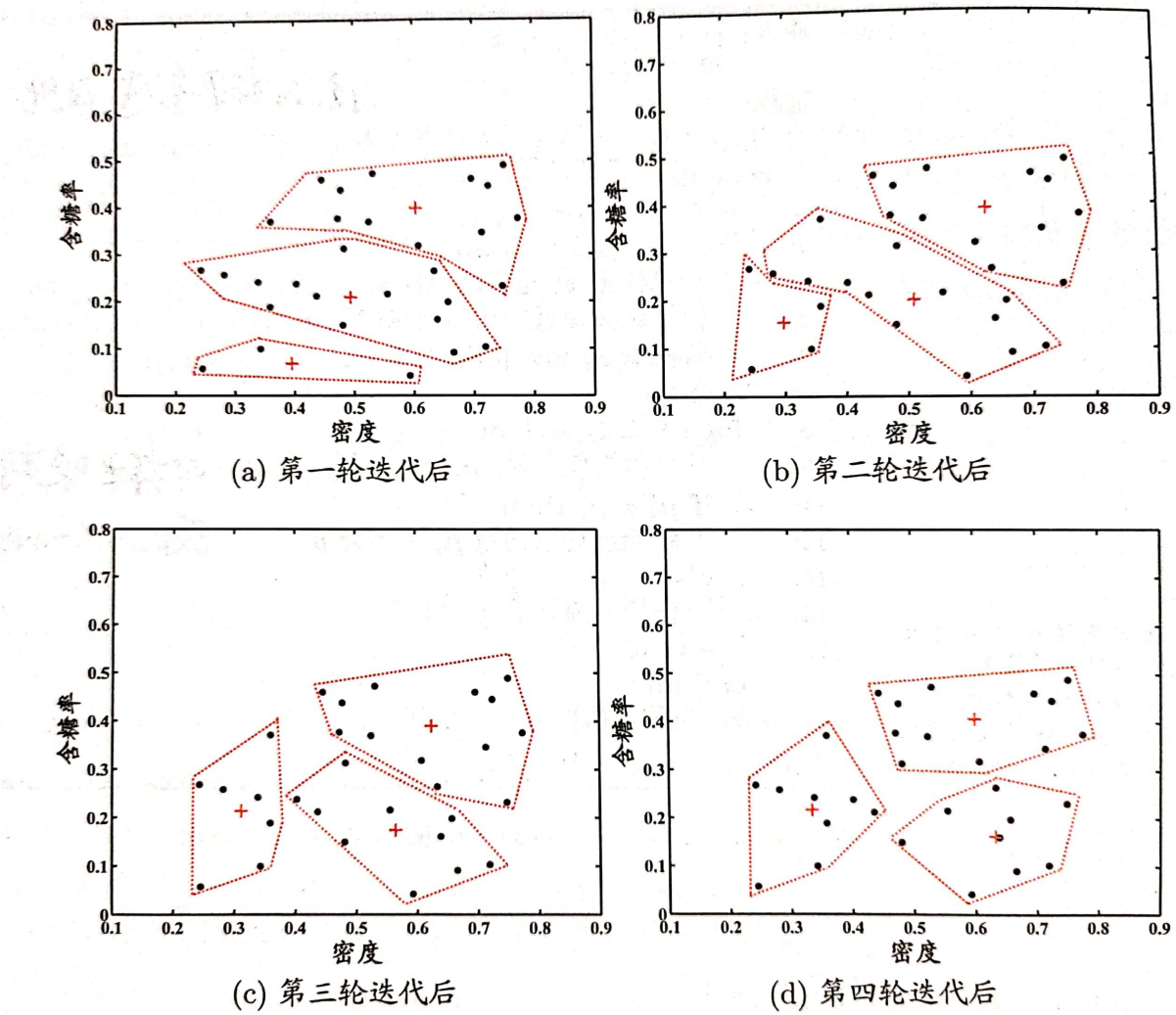
$$C_1 = \{x_3, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{13}, x_{14}, x_{17}, x_{18}, x_{19}, x_{20}, x_{23}\}$$
$$C_2 = \{x_{11}, x_{12}, x_{16}\}$$
$$C_3 = \{x_1, x_2, x_4, x_{15}, x_{21}, x_{22}, x_{24}, x_{25}, x_{26}, x_{27}, x_{28}, x_{29}, x_{30}\}$$

于是，可以从分别求得新的均值向量$\mu_1' = (0.493; 0.207), \mu_2' = (0.394; 0.066), \mu_3' = (0.602; 0.396)$

不断重复上述过程，如下图所示。

聚类

- 聚类结果：



(a) 第一轮迭代后　(b) 第二轮迭代后

(c) 第三轮迭代后　(d) 第四轮迭代后

聚类

- 给定含有5个样本的集合

$$X = \begin{bmatrix} 0 & 0 & 1 & 5 & 5 \\ 2 & 0 & 0 & 0 & 2 \end{bmatrix}$$

- 试用k均值聚类算法将样本聚到2个类中。

（1）选择两个样本点作为类的中心。假设选择 $m_1^{(0)} = x_1 = (0,2)^{\mathrm{T}}$，$m_2^{(0)} = x_2 = (0,0)^{\mathrm{T}}$。

（2）以 $m_1^{(0)}$，$m_2^{(0)}$ 为类 $G_1^{(0)}$，$G_2^{(0)}$ 的中心，计算 $x_3 = (1,0)^{\mathrm{T}}$，$x_4 = (5,0)^{\mathrm{T}}$，$x_5 = (5,2)^{\mathrm{T}}$ 与 $m_1^{(0)} = (0,2)^{\mathrm{T}}$，$m_2^{(0)} = (0,0)^{\mathrm{T}}$ 的欧氏距离平方。

对 $x_3 = (1,0)^{\mathrm{T}}$，$d(x_3, m_1^{(0)}) = 5$，$d(x_3, m_2^{(0)}) = 1$，将 x_3 分到类 $G_2^{(0)}$。

对 $x_4 = (5,0)^{\mathrm{T}}$，$d(x_4, m_1^{(0)}) = 29$，$d(x_4, m_2^{(0)}) = 25$，将 x_4 分到类 $G_2^{(0)}$。

对 $x_5 = (5,2)^{\mathrm{T}}$，$d(x_5, m_1^{(0)}) = 25$，$d(x_5, m_2^{(0)}) = 29$，将 x_5 分到类 $G_1^{(0)}$。

(3) 得到新的类 $G_1^{(1)}=\{x_1,x_5\}$, $G_2^{(1)}=\{x_2,x_3,x_4\}$, 计算类的中心 $m_1^{(1)}$, $m_2^{(1)}$:

$$m_1^{(1)}=(2.5,2.0)^{\mathrm{T}},\quad m_2^{(1)}=(2,0)^{\mathrm{T}}$$

(4) 重复步骤 (2) 和步骤 (3)。

将 x_1 分到类 $G_1^{(1)}$, 将 x_2 分到类 $G_2^{(1)}$, x_3 分到类 $G_2^{(1)}$, x_4 分到类 $G_2^{(1)}$, x_5 分到类 $G_1^{(1)}$。

得到新的类 $G_1^{(2)}=\{x_1,x_5\}$, $G_2^{(2)}=\{x_2,x_3,x_4\}$。

由于得到的新的类没有改变, 聚类停止。得到聚类结果:

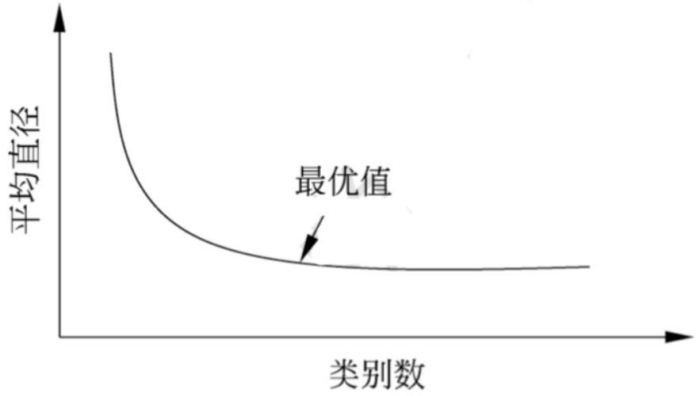
$$G_1^{*}=\{x_1,x_5\},\quad G_2^{*}=\{x_2,x_3,x_4\}$$

聚类

- 收敛性
- k均值聚类属于启发式方法，不能保证收敛到全局最优，初始中心的选择会直接影响聚类结果。
- 注意，类中心在聚类的过程中会发生移动，但是往往不会移动太大，因为在每一步，样本被分到与其最近的中心的类中。
- 初始类的选择
- 选择不同的初始中心，会得到不同的聚类结果。
- 初始中心的选择，比如可以用层次聚类对样本进行聚类，得到k个类时停止。然后从每个类中选取一个与中心距离最近的点。

聚类

- 类别数k的选择



- k均值聚类中的类别数k值需要预先指定，而在实际应用中最优的k值是不知道的。
- 尝试用不同的k值聚类，检验得到聚类结果的质量，推测最优的k值。
- 聚类结果的质量可以用类的平均直径来衡量。
- 一般地，类别数变小时，平均直径会增加
- 类别数变大超过某个值以后，平均直径会不变，而这个值正是最优的k值。实验时，可以采用二分查找，快速找到最优的k值。

学习向量量化（Learning Vector Quantization, LVQ）

与一般聚类算法不同的是，LVQ假设数据样本带有类别标记，学习过程中利用样本的这些监督信息来辅助聚类.

给定样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ ，LVQ的目标是学得一组 维原型向量 $\{p_1,p_2,\cdots,p_q\}$ ，每个原型向量代表一个聚类簇。

输入: 样本集$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$;
原型向量个数q, 各原型向量预设的类别标记$\{t_1, t_2, \ldots, t_q\}$;
学习率$\eta \in (0, 1)$.

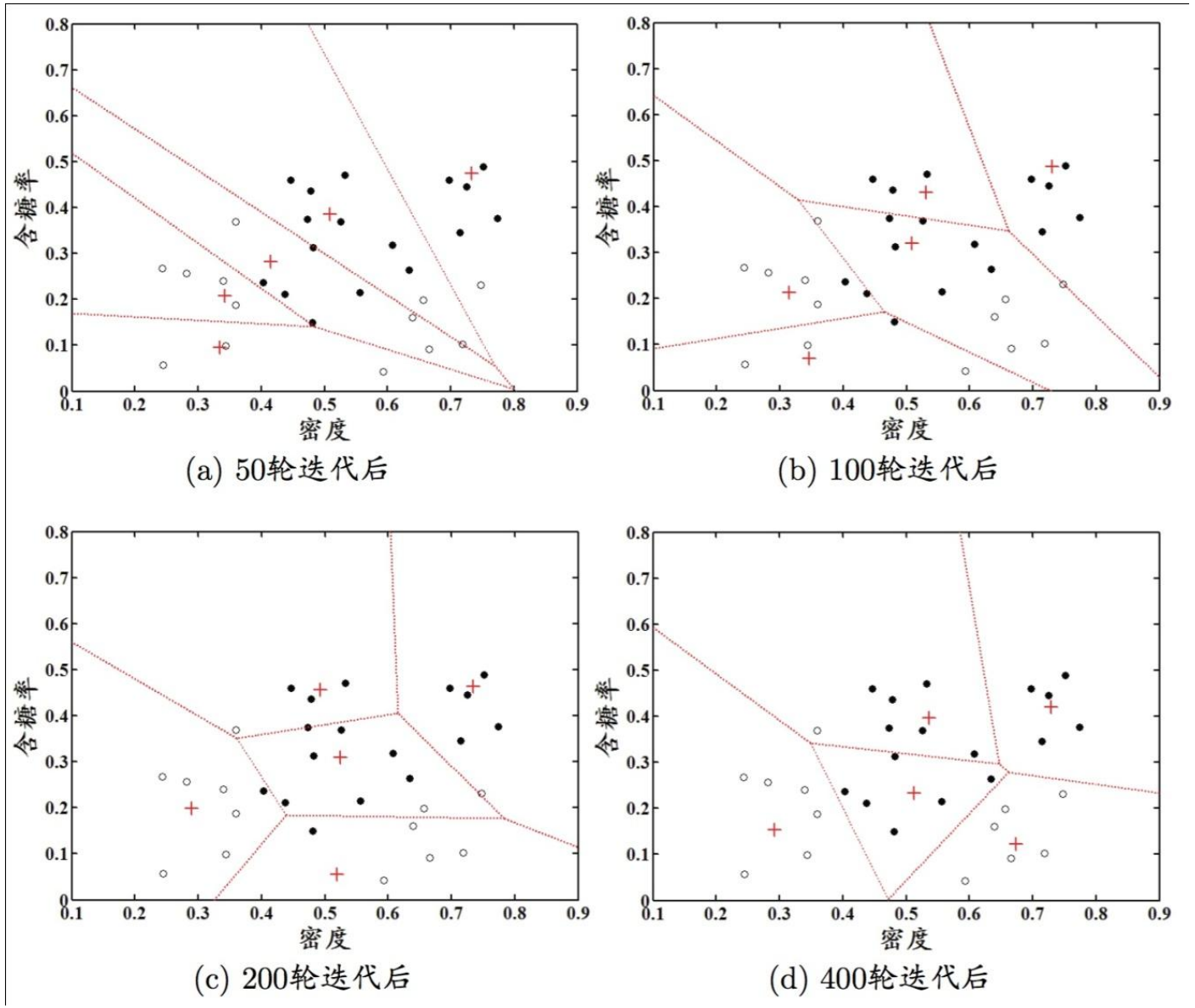
过程:

初始化一组原型向量$\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q\}$
repeat
  从样本集D随机选取样本$(\boldsymbol{x}_j, y_j)$;
  计算样本$\boldsymbol{x}_j$与$\boldsymbol{p}_i$ $(1 \leq i \leq q)$的距离: $d_{ji} = ||\boldsymbol{x}_j - \boldsymbol{p}_i||_2$;
  找出与$\boldsymbol{x}_j$距离最近的原型向量; $i^* = \arg\min_{i \in \{1,2,\ldots,q\}} d_{ji}$;
  if $y_j = t_{i^*}$ **then**
    $\boldsymbol{p}' = \boldsymbol{p}_{i^*} + \eta \cdot (\boldsymbol{x}_j - \boldsymbol{p}_{i^*})$
  else
    $\boldsymbol{p}' = \boldsymbol{p}_{i^*} - \eta \cdot (\boldsymbol{x}_j - \boldsymbol{p}_{i^*})$
  end if
  将原型向量$\boldsymbol{p}_{i^*}$更新为$\boldsymbol{p}'$
until 满足停止条件
return 当前原型向量

输出: 原型向量$\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q\}$

聚类

- 聚类效果：


含糖率
密度
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
(a) 50轮迭代后
(b) 100轮迭代后
(c) 200轮迭代后
(d) 400轮迭代后

与*k*均值、LVQ用原型向量来刻画聚类结构不同，高斯混合聚类 （Mixture-of-Gaussian）采用概率模型来表达聚类原型：

- 多元高斯分布的定义

对 n 维样本空间中的随机向量 x ，若 x 服从高斯分布，其概率密度函数为

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

其中 μ 是 n 维均值向量，Σ 是 $n \times n$ 的协方差矩阵。也可将概率密度函数记作 $p(x|\mu, \Sigma)$ 。

- 高斯混合分布的定义

$$p_M(x) = \sum_{i=1}^{k} \alpha_i p(x|\mu_i, \Sigma_i)$$

该分布由 k 个混合分布组成，每个成分对应一个高斯分布。其中，μ_i 与 Σ_i 是第 i个高斯混合成分的参数。而$\alpha_i > 0$ 为相应的“混合系数”，$\sum_{i=1}^{k} \alpha_i = 1$ 。

- 假设样本的生成过程由高斯混合分布给出：

首先，根据 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 定义的先验分布选择高斯混合成分，α_i 为选择第 i 个混合成分的概率；

然后，根据被选择的混合成分的概率密度函数进行采样，从而生成相应的样本。

- 模型求解：最大化（对数）似然

$$LL(D) = \ln\left(\prod_{j=1}^{m} p_M(x_j)\right)$$

$$= \sum_{j=1}^{m} \ln\left(\sum_{i=1}^{k} \alpha_i \cdot p\left(x_j|\mu_i, \Sigma_i\right)\right)$$

令：

$$\frac{\partial LL(D)}{\partial \mu_i} = 0 \quad \longrightarrow \quad \mu_i = \frac{\sum_{j=1}^{m} \gamma_{ji} x_j}{\sum_{j=1}^{m} \gamma_{ji}}$$

令：

$$\frac{\partial LL(D)}{\partial \Sigma_i} = 0 \longrightarrow \Sigma_i = \frac{\sum_{j=1}^{m} \gamma_{ji}(x_j - \mu_i)(x_j - \mu_i)^T}{\sum_{j=1}^{m} \gamma_{ji}}$$

拉格朗日乘子法 $\longrightarrow$ $\alpha_i = \frac{1}{m}\sum_{j=1}^{m} \gamma_{ji}$

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.

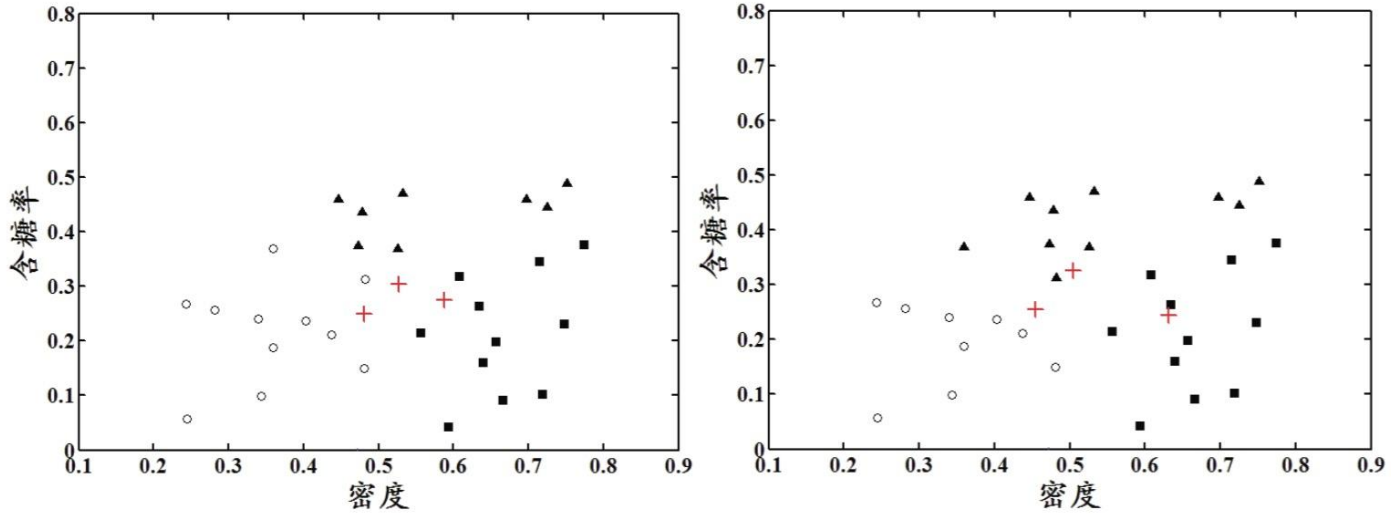
过程:

初始化高斯混合分布的模型参数$\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$
repeat
  for $j = 1, \ldots, m$ **do**
    根据(9.30)计算$\boldsymbol{x}_j$由各混合成分生成的后验概率, 即
    $\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j)$ $(1 \leq i \leq k)$
  end for
  for $i = 1, \ldots, k$ **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: $\boldsymbol{\Sigma}_i' = \frac{\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i')(\boldsymbol{x}_j - \boldsymbol{\mu}_i')^{\top}}{\sum_{j=1}^{m} \gamma_{ji}}$;
    计算新混合系数: $\alpha_i' = \frac{\sum_{j=1}^{m} \gamma_{ji}}{m}$;
  end for
  将模型参数$\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leq i \leq k\}$更新为$\{(\alpha_i', \boldsymbol{\mu}_i', \boldsymbol{\Sigma}_i') \mid 1 \leq i \leq k\}$
until 满足停止条件
$C_i = \emptyset$ $(1 \leq i \leq k)$
for $j = 1, \ldots, m$ **do**
  根据(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
return 簇划分结果

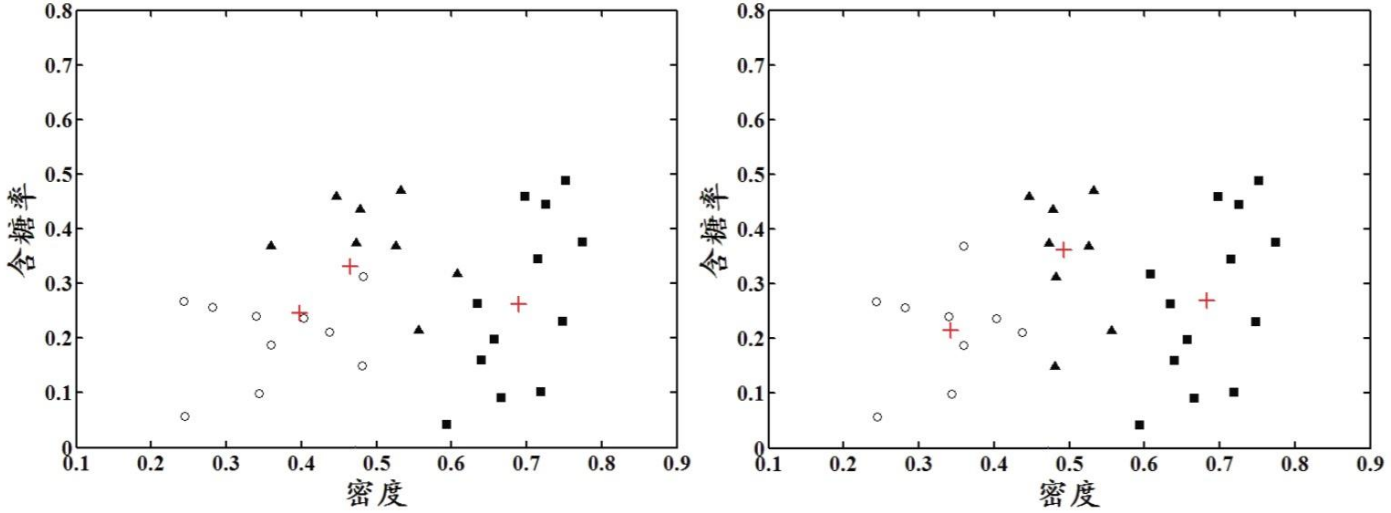
输出: 簇划分$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

聚类

- 聚类效果：



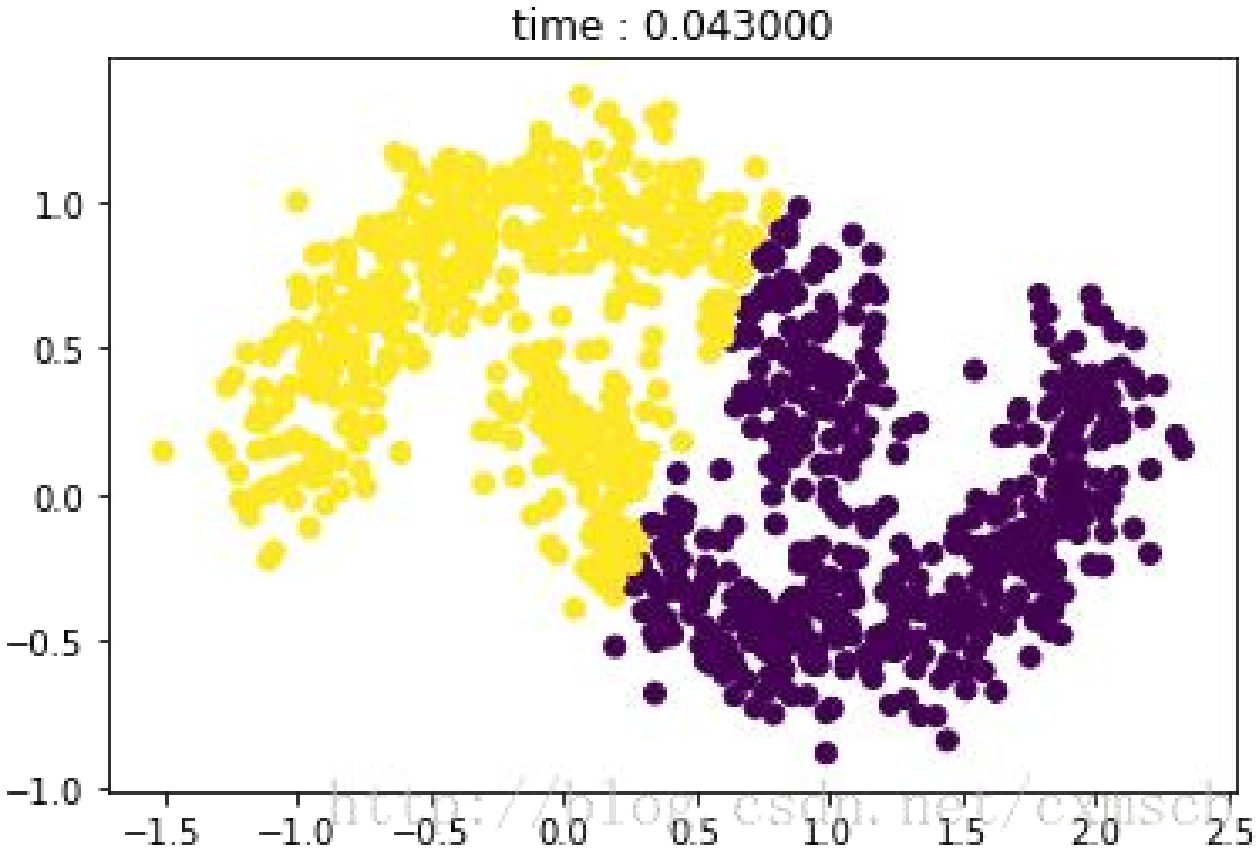
(a) 5轮迭代后　　(b) 10轮迭代后



(c) 20轮迭代后　　(d) 50轮迭代后

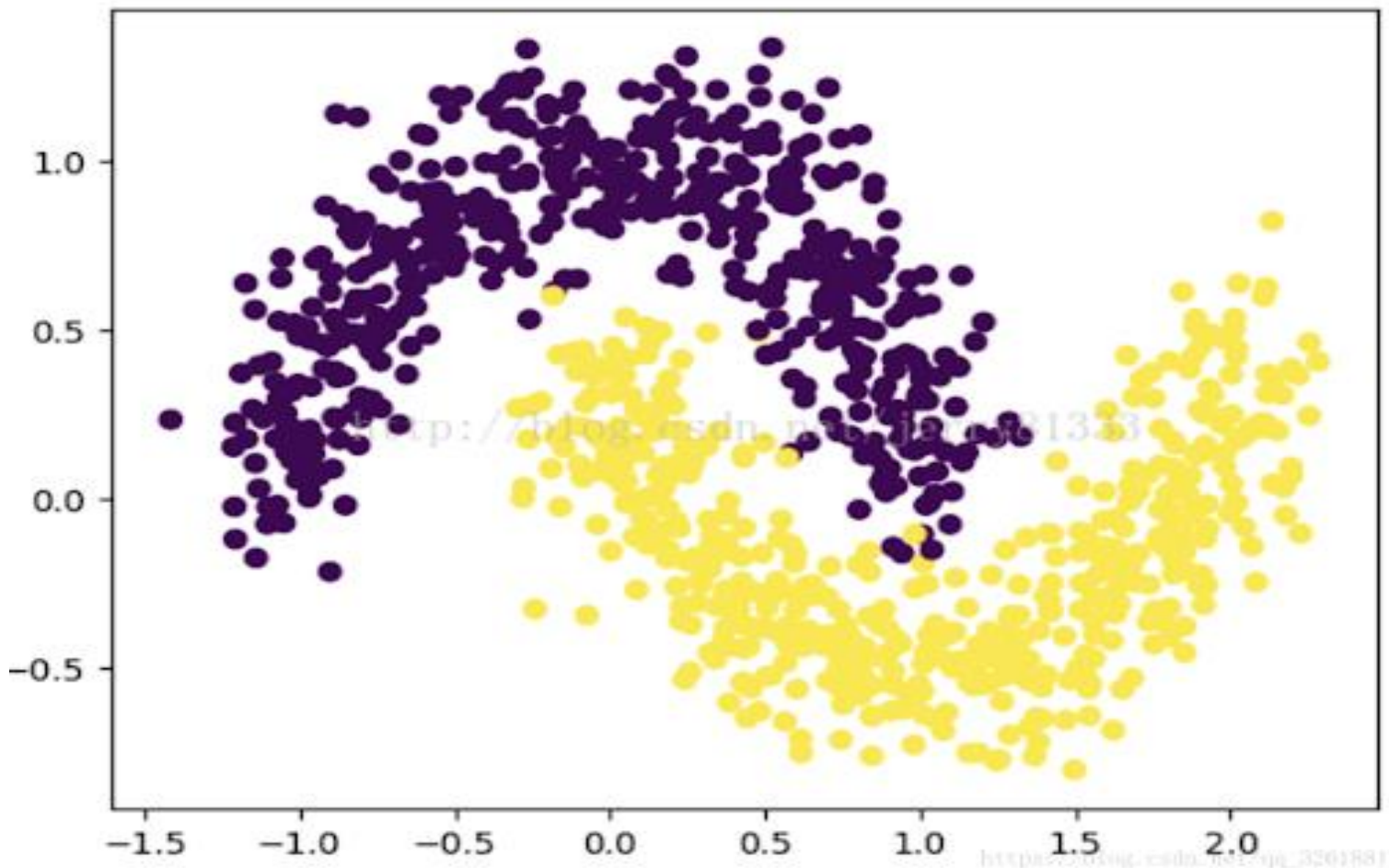
聚类

使用k均值方法对下图数据进行聚类，基于原型的聚类方法难以发掘到密度连接的信息，导致聚类结果同直观差异较大：



聚类

基于密度的聚类方法：从样本密度的角度考察样本的连接性，使密度相连的样本归结到一个簇，更符合直观认知：



聚类

- 密度聚类的定义

密度聚类也称为“基于密度的聚类”（density-based clustering)。

此类算法假设聚类结构能通过样本分布的紧密程度来确定。

通常情况下，密度聚类算法从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇来获得最终的聚类结果。

接下来介绍DBSCAN这一密度聚类算法。

- DBSCAN算法：基于一组“邻域”参数$(\epsilon, MinPts)$来刻画样本分布的紧密程度。
- 基本概念：
  - ϵ 邻域：对样本 $x_j \in D$，其ϵ 邻域包含样本集D 中与 x_j的距离不大于 ϵ 的样本；
  - 核心对象：若样本x_j 的ϵ邻域至少包含MinPts个样本，则该样本点为一个核心对象；
  - 密度直达：若样本 x_j 位于样本x_i的ϵ邻域中，且x_i是一个核心对象，则称样本 x_j由x_i密度直达；
  - 密度可达：对样本x_i与 x_j，若存在样本序列 $p_1, p_2, \cdots, p_n$，其中 $p_1 = x_i, p_n = x_j$ 且 p_{i+1}由 p_i密度直达，则该两样本密度可达；
  - 密度相连：对样本 x_i与x_j，若存在样本x_k 使得两样本均由 x_k密度可达，则称该两样本密度相连。

- 一个例子

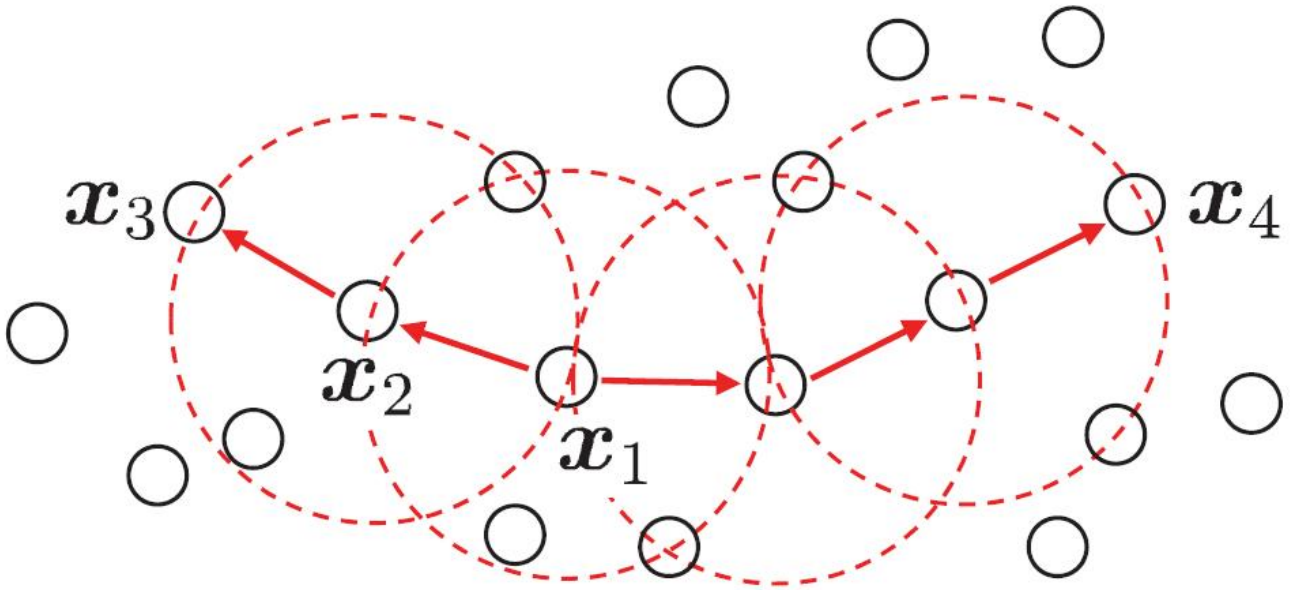
  令$MinPts = 3$，则虚线显示出ϵ领域。

  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。



- 对“簇”的定义

由密度可达关系导出的最大密度相连样本集合。

- 对“簇”的形式化描述

给定领域参数，簇是满足以下性质的非空样本子集：

连接性：$x_i \in C, x_j \in C \Rightarrow x_i$ 与 x_j 密度相连

最大性：$x_i \in C,$ x_i 与 x_j 密度可达 $\Rightarrow x_j \in C$

实际上，若x 为核心对象，由x 密度可达的所有样本组成的集合记为$X = \{x' \in D \mid x'$由x密度可达$\}$，则X为满足连接性与最大性的簇。

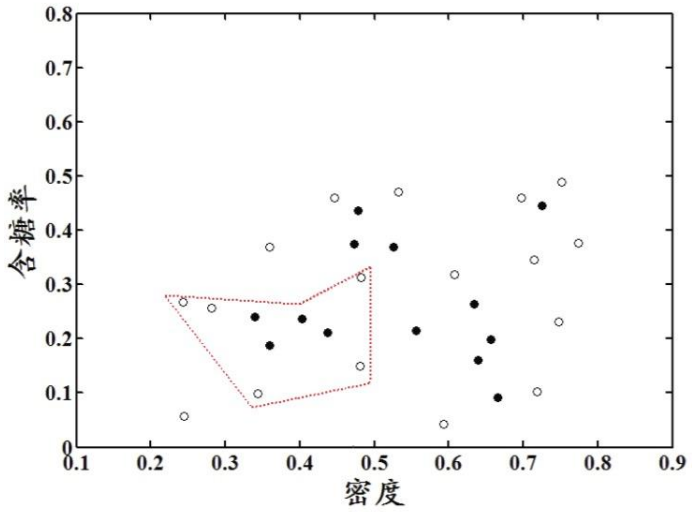
输入： 样本集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
邻域参数$(\epsilon,MinPts)$.

过程：

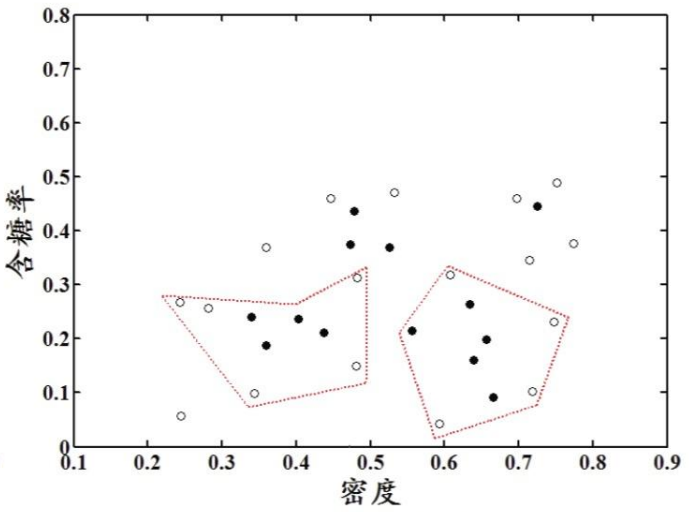
初始化核心对象集合：$\Omega=\emptyset$
for $j=1,\ldots,m$ **do**
  确定样本$\boldsymbol{x}_j$的ϵ-邻域$N_\epsilon(\boldsymbol{x}_j)$;
  if $|N_\epsilon(\boldsymbol{x}_j)|\geq MinPts$ **then**
    将样本$\boldsymbol{x}_j$加入核心对象集合：$\Omega=\Omega\bigcup\{\boldsymbol{x}_j\}$
  end if
end for
初始化聚类簇数：$k=0$
初始化未访问样本集合：$\Gamma=D$
while $\Omega\neq\emptyset$ **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 $Q\neq\emptyset$ **do**
    取出队列Q中的首个样本$\boldsymbol{q}$;
    if $|N_\epsilon(\boldsymbol{q})|\geq MinPts$ **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
return 簇划分结果

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

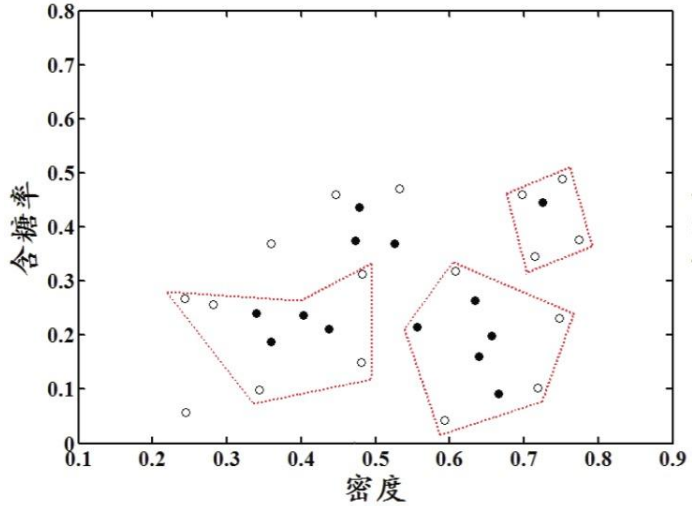
- 聚类效果：



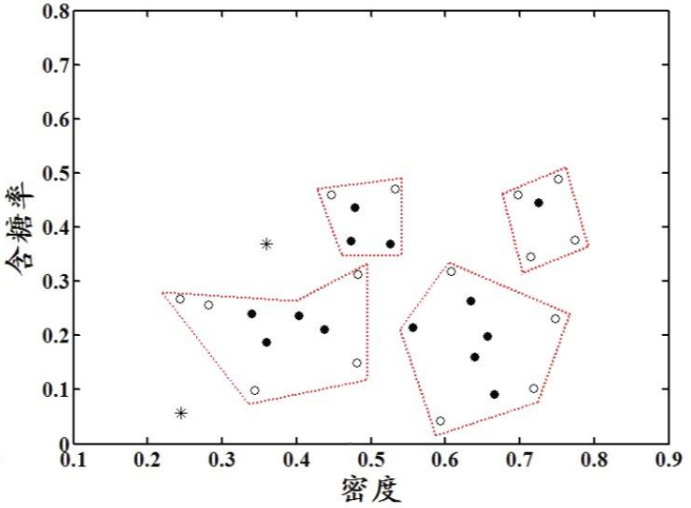
(a) 生成聚类簇C_1



(b) 生成聚类簇C_2



(c) 生成聚类簇C_3



(d) 生成聚类簇C_4

感 谢 观 看

统 计 机 器 学 习

主讲人：彭振华

数学与计算机学院

2026年