



机器学习（研究生）

聚类与 EM 算法

Clustering & the EM Algorithm

赵凯 上海大学

<https://kaizhao.net/teaching/shu-ml-grad>

本讲内容索引

01 聚类的度量与距离

1.1 聚类任务 · 1.2 性能度量 · 1.3 距离度量

02 原型聚类

2.1 k 均值 · 2.2 高斯混合模型（GMM）

03 密度聚类与层次聚类

3.1 密度聚类 · 3.2 层次聚类 · 3.3 方法比较

04 EM 算法

4.1 用途与隐变量 · 4.2 E 步与 M 步 ·
4.3 数学推导与收敛性

05 EM 求解 k 均值

5.1 E 步：估计簇归属 · 5.2 M 步：更新中心

06 小结

6.1 本讲要点 · 6.2 中英文术语对照 ·
6.3 课后作业

没有标记的数据

前几讲的训练样本都带标记 y ；更多时候只有样本本身

有监督学习（前五讲）

$$D = \{(x_1, y_1), \dots, (x_m, y_m)\}$$

学习从 x 到 y 的映射

无监督学习（本讲）

$$D = \{x_1, \dots, x_m\}, \text{ 标记未知}$$

揭示数据的内在性质与规律

- 无监督学习中研究最多、应用最广的任务是**聚类**（clustering）
- 聚类既可以单独使用，寻找数据的分布结构；也可以作为分类等任务的前驱，如先把用户聚成若干类

问题 没有标记，什么样的划分才算好？样本之间的“相似”如何计算？划分又如何求得？

记号 周志华《机器学习》第 9 章 9.1 节，p.197；常见的无监督学习任务还有密度估计、异常检测等。

本讲内容

Part 1 聚类

度量、距离与三类算法

01 聚类的度量与距离

- 1.1 聚类任务
- 1.2 性能度量
- 1.3 距离度量

02 原型聚类

- 2.1 k 均值
- 2.2 高斯混合模型 (GMM)

03 密度聚类与层次聚类

- 3.1 密度聚类 (DBSCAN)
- 3.2 层次聚类 (AGNES)
- 3.3 方法比较

Part 2 EM 算法

交替估计的思想与 k 均值例子

04 EM 算法

- 4.1 用途与隐变量
- 4.2 E 步与 M 步
- 4.3 数学推导与收敛性

05 EM 求解 k 均值

- 5.1 E 步：估计簇归属
- 5.2 M 步：更新中心

06 小结

- 6.1 本讲要点
- 6.2 中英文术语对照
- 6.3 课后作业

记号 西瓜数据集 4.0: 30 个样本，属性为密度与含糖率。周志华《机器学习》表 9.1, p.202。

符号约定

沿用前几讲的记号；新增与聚类有关的记号

数据	$D = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, m 个无标记样本; $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^\top \in \mathbb{R}^d$
簇划分	$\mathcal{C} = \{C_1, \dots, C_k\}$: $C_l \cap C_{l'} = \emptyset$ ($l \neq l'$), $\bigcup_{l=1}^k C_l = D$
簇标记	$\lambda_j \in \{1, \dots, k\}$: 样本 $\mathbf{x}_j$ 所属的簇, 即 $\mathbf{x}_j \in C_{\lambda_j}$
均值向量	$\boldsymbol{\mu}_i = \frac{1}{ C_i } \sum_{\mathbf{x} \in C_i} \mathbf{x}$: 簇 C_i 的中心
距离	$\text{dist}(\mathbf{x}_i, \mathbf{x}_j)$: 样本间的距离, 默认取欧氏距离

记号 向量均为列向量, $(a, b)^\top$ 表示二维列向量; $|C|$: 簇中的样本数; 本讲特征维数记 d (教材记 n)。

01

聚类的度量与距离

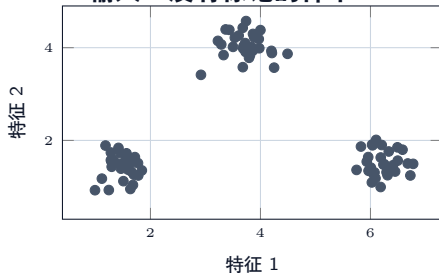
Clustering: Measures & Distances

1.1 聚类任务 · 1.2 性能度量 · 1.3 距离度量

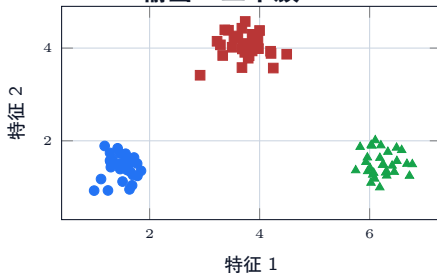
1.1 聚类任务

将相似样本归入同一个簇，使簇内紧密、簇间分离

输入：没有标记的样本



输出：三个簇



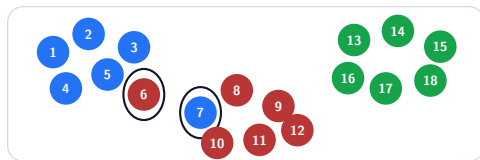
一个簇是一组相似样本；簇的实际含义由使用者解释。

记号 教学构造数据：三个明显分离的点群，共 90 个样本；两图坐标相同，颜色与点形表示簇。聚类任务定义见周志华《机器学习》9.1 节，p.197。

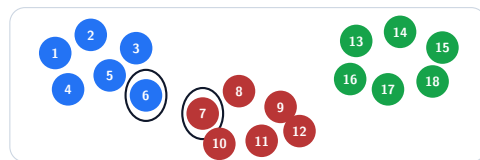
1.2 外部指标：需要参考模型

同一批 18 个样本：左为模型的聚类结果，右为参考模型给出的划分

模型划分 $\mathcal{C}$



参考划分 $\mathcal{C}^*$



$S = \text{Same}$ (同簇), $D = \text{Different}$ (异簇); 第一个字母看模型, 第二个看参考。

统计样本对	参考同簇 (S)	参考异簇 (D)
模型同簇 (S)	$a = SS = 35$, 如 (1, 2)	$b = SD = 10$, 如 (1, 7)
模型异簇 (D)	$c = DS = 10$, 如 (1, 6)	$d = DD = 98$, 如 (1, 8)

$$a + b + c + d = 35 + 10 + 10 + 98 = \binom{18}{2} = 153 \text{ 对}$$

记号 数字为样本编号, 颜色表示簇; 圈出的是划分发生差异的 6、7 号。每两个不同样本组成一对, 每对只计一次。参考可来自人工标注或专家划分; 图为教学构造示例。

外部指标：样本对计数与一致性

将“同簇”视为正类，四种计数对应二分类混淆矩阵

计数	对应	两种划分的判断
$a = 35$	TP	都认为同簇
$b = 10$	FP	模型同簇、参考异簇
$c = 10$	FN	模型异簇、参考同簇
$d = 98$	TN	都认为异簇

TP: True Positive (真正例)

FP: False Positive (假正例)

FN: False Negative (假反例)

TN: True Negative (真反例)

记号 JC: Jaccard Coefficient (Jaccard 系数); FMI: Fowlkes and Mallows Index (FM 指数); RI: Rand Index (Rand 指数)。周志华《机器学习》9.2 节, pp.198–199。

$$JC = \frac{a}{a + b + c} = \frac{35}{55} \approx 0.636$$

$$FMI = \sqrt{\frac{a}{a + b} \frac{a}{a + c}} = \frac{35}{45} \approx 0.778$$

$$RI = \frac{a + d}{a + b + c + d} = \frac{133}{153} \approx 0.869$$

三者均为**越大越好**；参考只用于评价。

内部指标：不需要参考模型

只根据数据和聚类结果，衡量簇内的紧密程度与簇间的分离程度

簇内：越紧密越好

$$\text{avg}(C) = \frac{2 \sum_{i < j} \text{dist}(\mathbf{x}_i, \mathbf{x}_j)}{|C|(|C| - 1)}$$

$$\text{diam}(C) = \max_{\mathbf{x}_i, \mathbf{x}_j \in C} \text{dist}(\mathbf{x}_i, \mathbf{x}_j)$$

簇间：越分离越好

$$d_{\min}(C_i, C_j) = \min_{\substack{\mathbf{x} \in C_i \\ \mathbf{z} \in C_j}} \text{dist}(\mathbf{x}, \mathbf{z})$$

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

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

DB 指数：越小越好

$$\text{DI} = \frac{\min_{i \neq j} d_{\min}(C_i, C_j)}{\max_l \text{diam}(C_l)}$$

Dunn 指数：越大越好

记号 DBI: Davies–Bouldin Index; DI: Dunn Index。avg 按簇内不同样本对求平均，要求 $|C| \geq 2$ ；上式假定分母非零。采用周志华《机器学习》式 (9.8)–(9.13) 的定义，p.199。

1.3 距离度量

距离度量的四条性质与闵可夫斯基距离（Minkowski distance）

四条基本性质

非负性： $\text{dist}(\boldsymbol{x}, \boldsymbol{z}) \geq 0$

同一性： $\text{dist}(\boldsymbol{x}, \boldsymbol{z}) = 0 \iff \boldsymbol{x} = \boldsymbol{z}$

对称性： $\text{dist}(\boldsymbol{x}, \boldsymbol{z}) = \text{dist}(\boldsymbol{z}, \boldsymbol{x})$

三角不等式（直递性）：

$\text{dist}(\boldsymbol{x}, \boldsymbol{z}) \leq \text{dist}(\boldsymbol{x}, \boldsymbol{w}) + \text{dist}(\boldsymbol{w}, \boldsymbol{z})$

闵可夫斯基距离

$$\text{dist}_{\text{mk}}(\boldsymbol{x}, \boldsymbol{z}) = \left(\sum_{u=1}^d |x_u - z_u|^p \right)^{1/p}$$

$p = 2$ 欧氏距离： $\|\boldsymbol{x} - \boldsymbol{z}\|_2$

$p = 1$ 曼哈顿距离： $\|\boldsymbol{x} - \boldsymbol{z}\|_1$

$p \rightarrow \infty$ 切比雪夫距离： $\max_u |x_u - z_u|$

$p \geq 1$ 时，闵可夫斯基距离满足四条基本性质。

记号 $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{w}$ 为样本向量； u 在求和式中为特征下标。周志华《机器学习》9.3 节，式 (9.14)–(9.20)，pp.199–200。

02

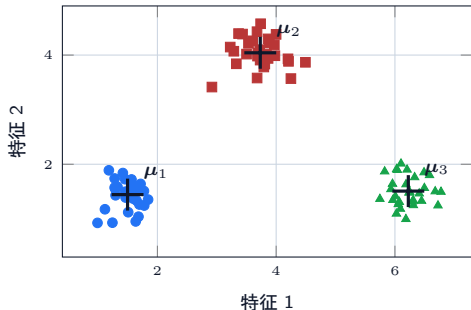
原型聚类

Prototype-Based Clustering

2.1 k 均值 · 2.2 高斯混合模型 (GMM)

原型聚类：用代表刻画簇

原型聚类（prototype-based clustering）：用代表性向量或分布参数刻画各个簇



原型（prototype）

对一个簇的概括与代表。

左图的 + 用位置概括一群样本。

两种原型表示

k 均值（ k -means）：

簇内样本的均值向量。

高斯混合聚类：

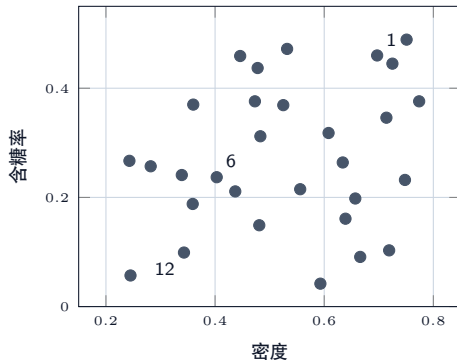
每个簇的高斯分布参数。

用样本与原型的关系确定归属。

记号 图为教学构造数据。“原型”指簇的代表； k 均值偏好紧凑、近球形的簇，“原型聚类”不等同于“圆形聚类”。均值向量可不对应任何已有样本。周志华《机器学习》9.4 节，p.202。

2.1 k 均值：西瓜样本与聚类任务

每个点是一只西瓜；用密度和含糖率，将 30 个样本划分为 $k=3$ 个簇



样本的两个属性

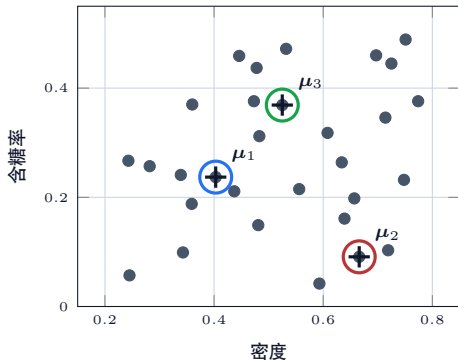
编号	密度	含糖率
1	0.697	0.460
6	0.403	0.237
12	0.343	0.099

输入：30 个二维向量。
输出：每个样本的簇标记。

记号 西瓜数据集 4.0：密度、含糖率；30 个样本的坐标始终不变。周志华《机器学习》表 9.1，p.202。 μ_i ：第 i 个簇的中心。

2.1 k 均值：1. 随机初始化中心

从数据中随机选择 $k = 3$ 个样本，作为三个原型的初始位置



一次随机选择的结果

$$\mu_1 = x_6 \\ (0.403, 0.237)^\top$$

$$\mu_2 = x_9 \\ (0.666, 0.091)^\top$$

$$\mu_3 = x_{25} \\ (0.525, 0.369)^\top$$

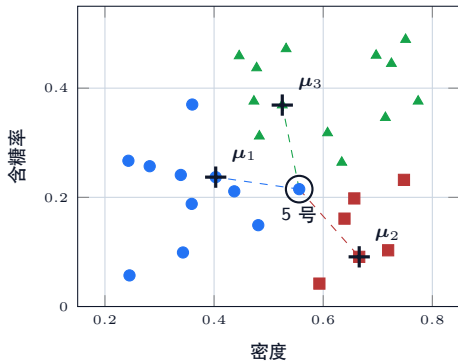
圈出：选中的样本。

+: 当前中心。

记号 西瓜数据集 4.0：密度、含糖率；30 个样本的坐标始终不变。周志华《机器学习》表 9.1, p.202。 μ_i ：第 i 个簇的中心。

2.1 k 均值：2. 第一次分配

固定三个中心，把每个样本分配给距离最近的中心



5 号样本的距离

到哪个中心	欧氏距离
μ_1	0.155
μ_2	0.166
μ_3	0.157

0.155 最小，故 $x_5 \in C_1$ 。

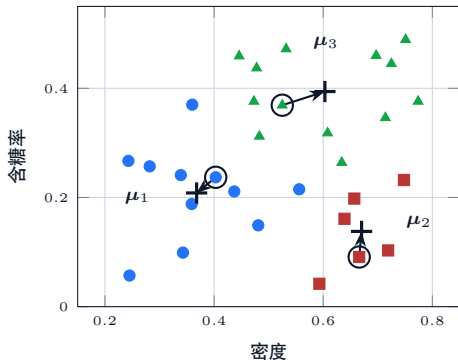
三个簇的样本数：

11、6、13。

记号 西瓜数据集 4.0：密度、含糖率；30 个样本的坐标始终不变。周志华《机器学习》表 9.1，p.202。 μ_i ：第 i 个簇的中心。

2.1 k 均值：3. 重新计算中心

固定簇成员，把每个中心移动到该簇所有样本的平均位置



簇 C_2 的 6 个成员

编号：9、13、14、16、17、21

$$\mu_2 = \begin{pmatrix} 4.022/6 \\ 0.827/6 \end{pmatrix} \approx \begin{pmatrix} 0.670 \\ 0.138 \end{pmatrix}$$

$$\mu_1 \approx (0.368, 0.208)^\top$$

$$\mu_3 \approx (0.603, 0.394)^\top$$

○：旧中心；+：新中心。

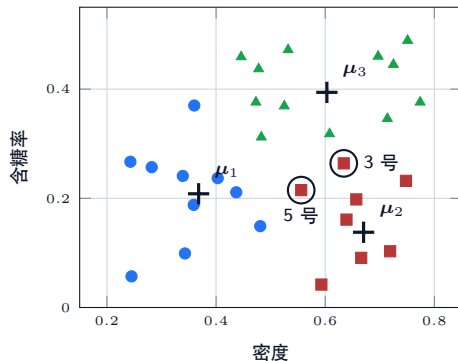
箭头：中心的移动。

点的颜色保持不变。

记号 西瓜数据集 4.0：密度、含糖率；30 个样本的坐标始终不变。周志华《机器学习》表 9.1，p.202。 μ_i ：第 i 个簇的中心。

2.1 k 均值：4. 重新分配样本

固定更新后的中心，重新比较距离；3、5 号样本转入 C_2



5 号样本的新距离

到哪个中心	欧氏距离
μ_1	0.188
μ_2	0.138
μ_3	0.185

5 号: $C_1 \rightarrow C_2$

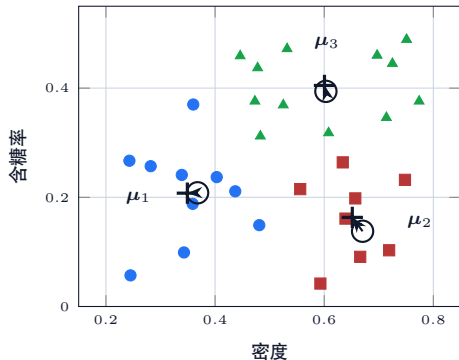
3 号: $C_3 \rightarrow C_2$

样本数: $(11, 6, 13) \rightarrow (10, 8, 12)$ 。

记号 西瓜数据集 4.0: 密度、含糖率; 30 个样本的坐标始终不变。周志华《机器学习》表 9.1, p.202。 μ_i : 第 i 个簇的中心。

2.1 k 均值：5. 再次计算中心

簇成员发生变化，再按新的成员计算三个均值向量



C_2 加入 3、5 号样本

$$\mu_2 = \begin{pmatrix} (4.022 + 0.634 + 0.556)/8 \\ (0.827 + 0.264 + 0.215)/8 \end{pmatrix}$$

$$\mu_2 = (0.6515, 0.16325)^\top$$

$$\mu_1 \approx (0.349, 0.208)^\top$$

$$\mu_3 \approx (0.601, 0.405)^\top$$

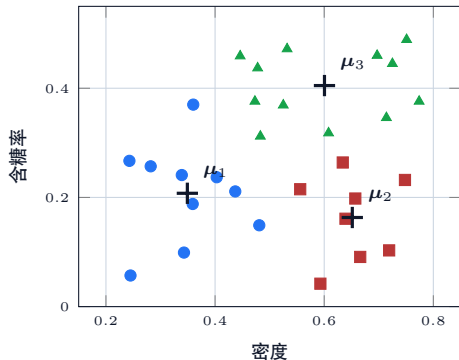
○: 旧中心; +: 新中心。

簇成员固定，中心重新计算。

记号 西瓜数据集 4.0: 密度、含糖率; 30 个样本的坐标始终不变。周志华《机器学习》表 9.1, p.202。 μ_i : 第 i 个簇的中心。

2.1 k 均值：6. 划分与中心稳定

用第二次更新的中心再次分配，30 个样本均保持原簇，中心也不再移动



最终划分

C_1 : 10 个样本

C_2 : 8 个样本

C_3 : 12 个样本

反复执行：
按最近中心划分
↓
按簇成员更新均值。

记号 西瓜数据集 4.0：密度、含糖率；30 个样本的坐标始终不变。周志华《机器学习》表 9.1，p.202。 μ_i ：第 i 个簇的中心。

2.1 k 均值的目标函数

让每个样本尽量接近代表其所在簇的均值向量

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

固定中心：分配样本

每个样本选择最近的中心。

$$\lambda_j = \arg \min_i \|x_j - \mu_i\|_2^2$$

固定划分：更新中心

簇内平方距离和在均值处最小。

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

两步都使 $\mathcal{L}$ 不增大；不同初值可能得到不同结果。

记号 周志华《机器学习》9.4.1 节，式 (9.24)，p.202；目标函数统一记 $\mathcal{L}$ 。交替更新通常得到局部最优解，不保证全局最优。

2.1 k 均值算法总结

输入样本集 D 与簇数 k ；输出簇划分 C 和均值向量

```
1: 随机选  $k$  个样本，作为初始中心  $\mu_1, \dots, \mu_k$ 
2: repeat
3:   清空各簇  $C_1, \dots, C_k$ 
4:   for 每个样本  $x_j$  do
5:      $\lambda_j \leftarrow \arg \min_i \|x_j - \mu_i\|_2^2$ 
6:     将  $x_j$  加入  $C_{\lambda_j}$ 
7:   end for
8:   for 每个非空簇  $C_i$  do
9:      $\mu_i \leftarrow \frac{1}{|C_i|} \sum_{x \in C_i} x$ 
10:  end for
11: until 簇划分不再变化
```

- 簇数 k 事先给定
- 成员与中心交替更新
- 常用不同初值多次运行

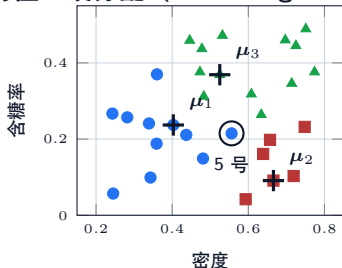
原型就是每一步使用和更新的均值向量。

记号 周志华《机器学习》9.4.1 节，图 9.2，p.203。距离相等时选编号最小的簇；空簇保留原中心；实际实现可另设最大迭代次数。

2.2 GMM: 从硬分配到软分配

高斯混合模型 (Gaussian Mixture Model, GMM): 用概率描述样本的簇归属

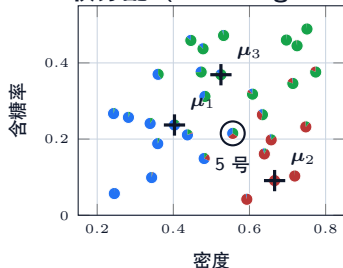
k 均值: 硬分配 (hard assignment)



5 号样本归入 C_1 : $(1, 0, 0)$ 。

每个样本只属于一个簇。

GMM: 软分配 (soft assignment)



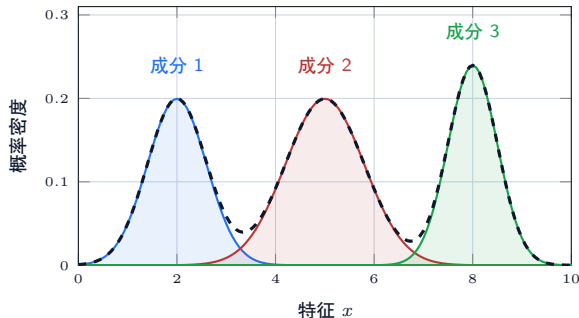
5 号样本: 35.7%、29.9%、34.4%。

为每个簇保留归属概率, 总和为 1。

记号 两图为同一组西瓜样本、相同初始中心; GMM 初始协方差为 $0.01I$, 混合系数为 $1/3$ 。小饼图三色面积表示归属概率; z_j 仍是单个未知簇编号, 软分配表示对该编号的不确定性。

2.2 GMM：多个高斯分布的加权组合

每个簇对应一个高斯分布；所有簇共同构成高斯混合模型



彩色实线：各成分的加权密度；黑色虚线：相加后的密度。

一个高斯成分的参数

μ_i ：中心位置。

σ_i^2 ：分布的宽窄。

α_i ：该成分的比例。

$$p(x) = \sum_{i=1}^k \alpha_i \mathcal{N}(x | \mu_i, \sigma_i^2)$$

$$\alpha_i \geq 0, \quad \sum_i \alpha_i = 1.$$

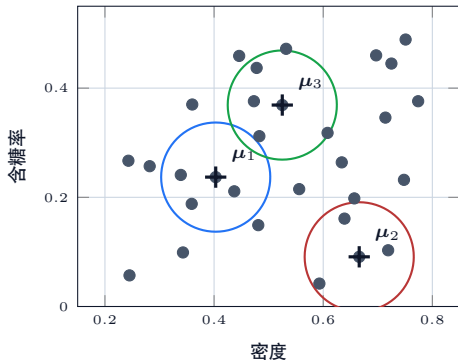
k 均值用均值代表簇；

GMM 用高斯分布代表簇。

记号 一维教学示意： $\alpha = (0.3, 0.4, 0.3)$ ， $\mu = (2, 5, 8)$ ， $\sigma = (0.6, 0.8, 0.5)$ 。多维时用均值向量 μ_i 与协方差矩阵 Σ_i 表示中心、尺度和方向。

2.2 GMM: 1. 初始化三个高斯成分

30 个西瓜样本、 $k = 3$ ；随机选三个样本作初始均值，并设定初始形状与比例



与 k 均值相同的初始中心

$$\mu_1 = x_6 = (0.403, 0.237)^\top$$

$$\mu_2 = x_9 = (0.666, 0.091)^\top$$

$$\mu_3 = x_{25} = (0.525, 0.369)^\top$$

初始形状: $\Sigma_i = 0.01I$ 。

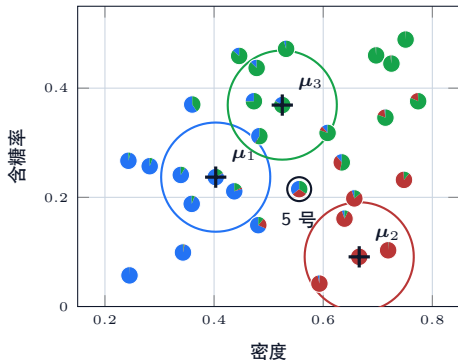
初始比例: $\alpha_i = 1/3$ 。

三个成分初始均为等大的圆。

记号 西瓜数据集 4.0, 30 个样本；+ 为均值，椭圆为马氏距离等于 1 的等密度线，非簇边界。小饼图的蓝、红、绿面积分别表示属于三个簇的概率。

2.2 GMM: 2. 第一次软分配

固定三个高斯成分，计算每个样本属于各簇的概率



5 号样本：加权密度归一化

簇	加权密度	归属概率
1	1.606	35.7%
2	1.343	29.9%
3	1.545	34.4%

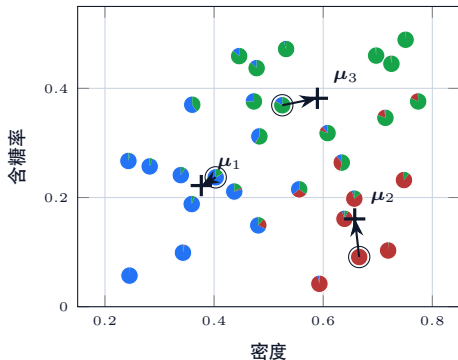
$$\gamma_{5i} = \frac{\alpha_i p(\mathbf{x}_5 | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_l \alpha_l p(\mathbf{x}_5 | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

每个样本都保留三个概率。

记号 西瓜数据集 4.0, 30 个样本；+ 为均值，椭圆为马氏距离等于 1 的等密度线，非簇边界。小饼图的蓝、红、绿面积分别表示属于三个簇的概率。

2.2 GMM: 3. 按概率重新计算中心

固定归属概率；每个样本按其属于该簇的概率，参与均值计算



成分 2 的加权平均

5 号的权重为 $\gamma_{52} \approx 0.299$ 。

有效样本数: $N_2 = \sum_j \gamma_{j2} \approx 6.910$ 。

$$\mu_2^{\text{new}} = \frac{\sum_j \gamma_{j2} \mathbf{x}_j}{N_2} = \begin{pmatrix} 4.545/6.910 \\ 1.111/6.910 \end{pmatrix}$$

$\mu_2^{\text{new}} \approx (0.658, 0.161)^\top$ 。

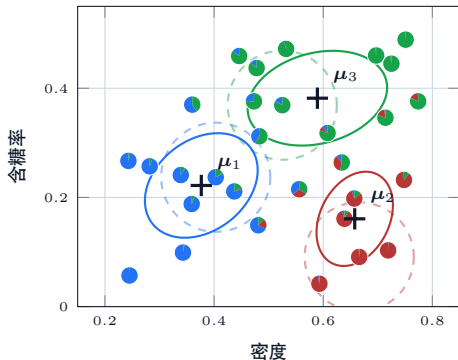
o: 旧中心; +: 新中心。

小饼图不变: 概率保持固定。

记号 西瓜数据集 4.0, 30 个样本; + 为均值, 椭圆为马氏距离等于 1 的等密度线, 非簇边界。小饼图的蓝、红、绿面积分别表示属于三个簇的概率。

2.2 GMM: 4. 更新形状与混合比例

继续使用同一组归属概率，计算加权协方差和各成分的比例



成分 2 的更新结果

$$\Sigma_2^{\text{new}} \approx \begin{pmatrix} 0.00490 & 0.00175 \\ 0.00175 & 0.00746 \end{pmatrix}$$

协方差决定椭圆的大小和方向。

$$\alpha_2^{\text{new}} = \frac{N_2}{30} = \frac{6.910}{30} \approx 0.230$$

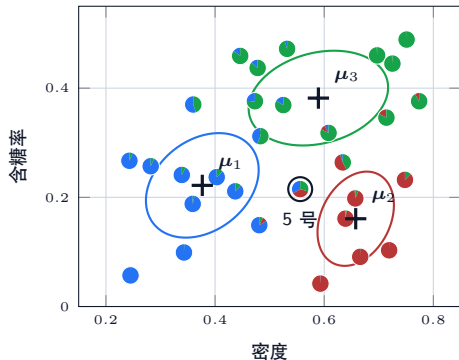
虚线：旧分布；实线：新分布。

概率固定，分布参数更新。

记号 西瓜数据集 4.0, 30 个样本；+ 为均值，椭圆为马氏距离等于 1 的等密度线，非簇边界。小饼图的蓝、红、绿面积分别表示属于三个簇的概率。

2.2 GMM: 5. 重新计算归属概率

固定更新后的三个高斯成分，再为每个样本计算三个归属概率



5 号样本的概率变化

簇	第一次	第二次
1	35.7%	33.4%
2	29.9%	37.2%
3	34.4%	29.4%

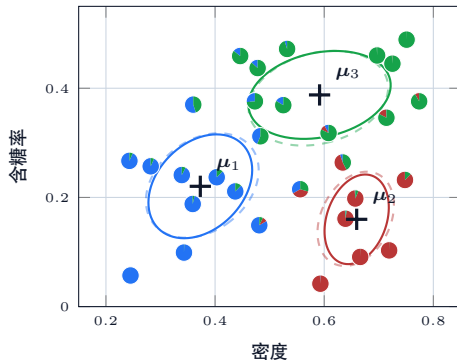
最大概率从簇 1 变为簇 2。

三个概率全部保留，
继续用于下一次加权更新。

记号 西瓜数据集 4.0, 30 个样本；+ 为均值，椭圆为马氏距离等于 1 的等密度线，非簇边界。小饼图的蓝、红、绿面积分别表示属于三个簇的概率。

2.2 GMM: 6. 再次更新高斯成分

按新的归属概率，重新计算均值、协方差和混合比例；继续交替迭代



成分 2 再次更新

$$N_2 = \sum_j \gamma_{j2} \approx 7.096。$$

$$\mu_2 \approx (0.660, 0.160)^\top。$$

$$\alpha_2 = N_2/30 \approx 0.237。$$

概率更新 → 参数更新

反复进行，直到变化足够小。

虚线：上轮分布；实线：本轮分布。

记号 西瓜数据集 4.0, 30 个样本；+ 为均值，椭圆为马氏距离等于 1 的等密度线，非簇边界。小饼图的蓝、红、绿面积分别表示属于三个簇的概率。

2.2 GMM：概率模型

高斯混合模型（Gaussian Mixture Model, GMM）：用分布表示簇

每个簇用一个高斯分布刻画：

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

高斯混合分布：

$$p_{\mathcal{M}}(\mathbf{x}) = \sum_{i=1}^k \alpha_i p(\mathbf{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

混合系数 $\alpha_i > 0$, $\sum_i \alpha_i = 1$

样本的生成过程

1. 按先验 $\alpha_1, \dots, \alpha_k$ 选择一个混合成分 $z_j \in \{1, \dots, k\}$
2. 从该成分的高斯分布 $p(\mathbf{x} | \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})$ 采样，得到 $\mathbf{x}_j$

z_j 决定了样本属于哪个簇，但取值未知：它是一个隐变量。

记号 $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ ：第 i 个混合成分的均值向量与协方差矩阵； $|\boldsymbol{\Sigma}|$ ：行列式。周志华《机器学习》第 9 章 9.4.3 节，式 (9.28)–(9.29)，pp.206–207。

2.2 GMM: 后验概率与簇划分

由贝叶斯公式，求样本 $\mathbf{x}_j$ 由第 i 个混合成分生成的后验概率

$$\gamma_{ji} = p_{\mathcal{M}}(z_j = i | \mathbf{x}_j) = \frac{P(z_j = i) p_{\mathcal{M}}(\mathbf{x}_j | z_j = i)}{p_{\mathcal{M}}(\mathbf{x}_j)} = \frac{\alpha_i p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^k \alpha_l p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

$$\lambda_j = \arg \max_{i \in \{1, \dots, k\}} \gamma_{ji}$$

- 先验 $P(z_j = i) = \alpha_i$
- 簇划分由原型对应的后验概率确定（第五讲的最大后验判定）

问题 参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}$ 未知。用第五讲的极大似然估计求解，会遇到什么困难？

记号 周志华《机器学习》第 9 章 9.4.3 节，式 (9.30)–(9.31)，p.208。

2.2 GMM：极大似然估计：均值与协方差

最大化对数似然 $LL(D) = \sum_{j=1}^m \ln \left(\sum_{i=1}^k \alpha_i p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$

对数里面是一个和：无法像单个高斯那样直接得到闭式解。令 $\partial LL / \partial \boldsymbol{\mu}_i = 0$ ：

$$\sum_{j=1}^m \underbrace{\frac{\alpha_i p(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_l \alpha_l p(\mathbf{x}_j | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}}_{\gamma_{ji}} (\mathbf{x}_j - \boldsymbol{\mu}_i) = 0 \quad \implies \quad \boldsymbol{\mu}_i = \frac{\sum_j \gamma_{ji} \mathbf{x}_j}{\sum_j \gamma_{ji}}$$

类似地，令 $\partial LL / \partial \boldsymbol{\Sigma}_i = 0$ ：

$$\boldsymbol{\Sigma}_i = \frac{\sum_j \gamma_{ji} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^\top}{\sum_j \gamma_{ji}}$$

均值与协方差是加权估计；权重 γ_{ji} 是依赖当前参数的后验概率。

记号 周志华《机器学习》第9章 9.4.3节，式(9.32)–(9.35)，pp.208–209；ln：自然对数。

2.2 GMM：算法总结

输入样本集与成分数 k ；输出各成分参数以及每个样本的归属概率

- 1: 初始化 $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1}^k$
- 2: **repeat**
- 3: 固定参数，计算每个样本的归属概率 γ_{ji}
- 4: 计算有效样本数 $N_i = \sum_j \gamma_{ji}$
- 5: 更新加权均值 μ_i 与协方差 Σ_i
- 6: 更新混合系数 $\alpha_i = N_i/m$
- 7: **until** 对数似然的增量小于阈值

软分配输出

每个样本保留 k 个概率：

$$(\gamma_{j1}, \dots, \gamma_{jk})$$

需要最终簇编号时

$$\lambda_j = \arg \max_i \gamma_{ji}$$

选择归属概率最大的簇。

样本归属概率与高斯分布参数相互依赖，通过交替更新共同估计。

记号 周志华《机器学习》9.4.3 节，pp.208–209；Stanford CS229：Mixtures of Gaussians。实现中还需限制迭代次数，并为协方差的特征值设置正下界。

03

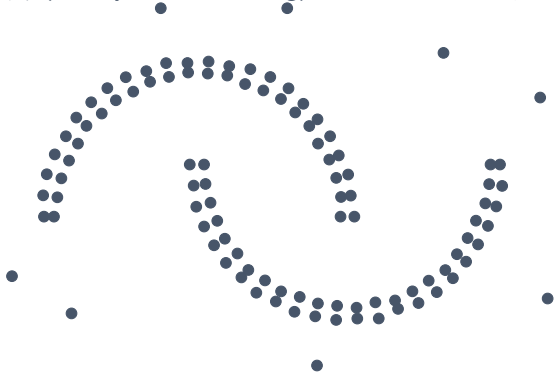
密度聚类与层次聚类

Density & Hierarchical Clustering

3.1 密度聚类 · 3.2 层次聚类 · 3.3 方法比较

3.1 密度聚类：相连的密集区域

密度聚类 (density-based clustering)：将相连的密集区域组成簇，以稀疏区域分隔簇

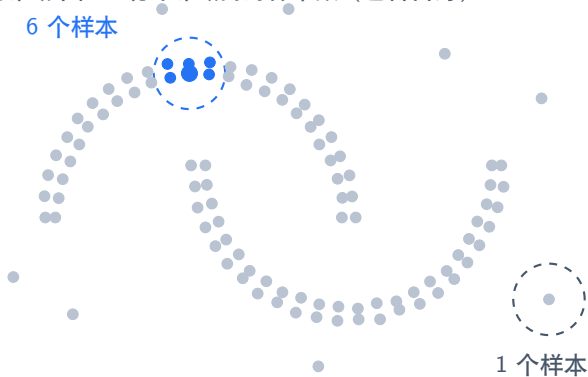


- **局部密集**：邻近样本多
- **逐段相连**：形成弯曲簇
- **稀疏、孤立样本**可作为噪声

记号 教学构造数据：104 个二维样本；横、纵坐标均为无量纲特征。

DBSCAN: 局部邻域与核心对象

固定邻域半径，统计邻域内的样本数（包含自身）



邻域半径: 0.24

最少样本数: 5

密集邻域的中心样本
称为**核心对象**。

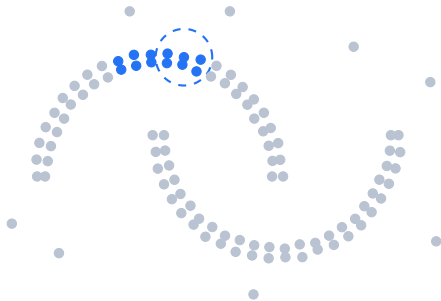
孤立样本的邻域不够密集。

记号 DBSCAN: Density-Based Spatial Clustering of Applications with Noise; 数据为教学构造。两个虚线圆的半径相同。

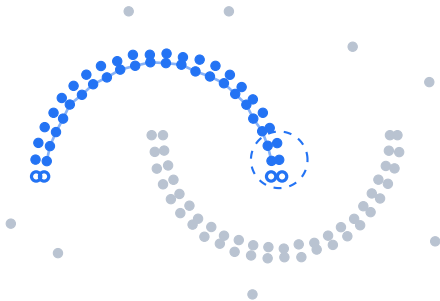
DBSCAN: 沿核心对象扩展

把邻域样本加入当前簇；遇到新的核心对象，再继续搜索其邻域

扩展初期：12 个样本



扩展完成：48 个样本

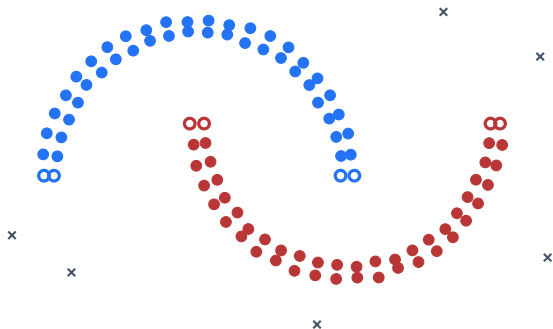


核心对象继续扩展；空心的边界样本加入簇后，搜索在该样本处停止。

记号 同一组构造数据；分别处理 8、44 个核心对象后的状态。半径 0.24，最少 5 个样本；细线连接相邻核心对象。

DBSCAN: 簇与噪声

再从未访问的核心对象出发，生成下一个簇



2 个簇

每簇 48 个样本

8 个噪声样本

未被任何核心对象的邻域覆盖

簇数由数据和邻域参数共同决定。

记号 教学构造数据；半径 0.24，最少 5 个样本。实心：核心对象；空心：边界样本；x：噪声。

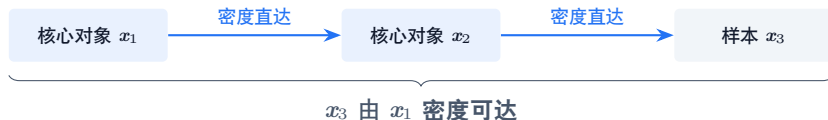
DBSCAN: 邻域与可达关系

$$N_{\epsilon}(\mathbf{x}) = \{\mathbf{z} \in D \mid \text{dist}(\mathbf{x}, \mathbf{z}) \leq \epsilon\} \quad \epsilon: \text{邻域半径}$$

核心对象 $|N_{\epsilon}(\mathbf{x})| \geq \text{MinPts}$

边界样本 非核心对象，位于某个核心对象的邻域内

噪声 不属于任何核心对象的邻域



两个样本可由同一核心对象密度可达，则二者密度相连。

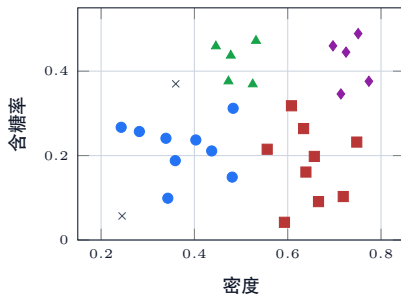
记号 MinPts : 核心对象邻域所需的最少样本数（含自身）。每一箭头的终点位于起点的 ϵ -邻域内。周志华《机器学习》9.5 节，式 (9.39)–(9.40)。

DBSCAN 算法与示例

1. 找出所有核心对象
2. 选种子，扩展至整个簇
3. 从剩余核心对象出发，重复扩展

$$\epsilon = 0.11, \quad MinPts = 5$$

4 个簇；11、15 号样本为噪声。

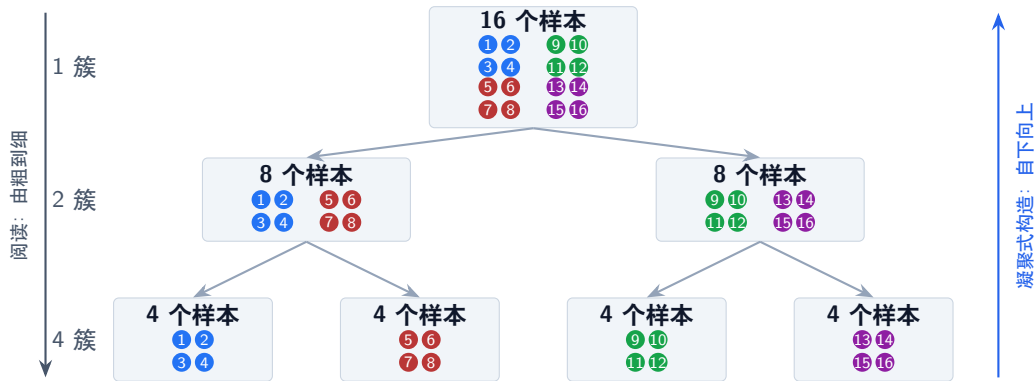


核心对象全部访问后停止；簇逐个生成，无需预设簇数。

记号 西瓜数据集 4.0；× 为噪声。以 8 号样本为首个种子，对应蓝色簇。周志华《机器学习》9.5 节，图 9.9–9.10，pp.213–214。

3.2 层次聚类：三个层次的簇划分

层次聚类（hierarchical clustering）：以树形结构表示不同粒度的划分；每个框表示一个簇



记号 教学构造：编号追踪同一样本。凝聚嵌套（AGglomerative NESTing, AGNES）从单样本簇开始合并；图示省略中间层次。

AGNES：自下向上的合并

每个样本为一个初始簇

合并距离最近的两个簇

重复至剩下 k 个簇

簇间距离的三种定义

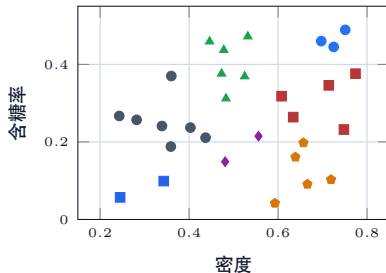
单链接 $d_{\min} = \min_{x \in C_i, z \in C_j} \text{dist}(x, z)$

全链接 $d_{\max} = \max_{x \in C_i, z \in C_j} \text{dist}(x, z)$

均链接 $d_{\text{avg}} = \frac{\sum_{x \in C_i} \sum_{z \in C_j} \text{dist}(x, z)}{|C_i| |C_j|}$

每次合并，簇数减少 1。

全链接， $k = 7$



记号 西瓜数据集 4.0；周志华《机器学习》9.6 节，图 9.11–9.13，pp.214–217。DIANA（DIvisive ANALysis）采用自顶向下的分拆策略。

3.3 三类聚类方法的比较

对“簇”的假设不同，迭代过程中簇的变化方式也不同

	原型聚类	密度聚类	层次聚类
对簇的假设	能用一组原型刻画	样本分布的紧密程度	不同层次上的划分
典型算法	k 均值、高斯混合聚类	DBSCAN	AGNES (DIANA)
每一步	更新归属，再更新原型参数	由一个核心对象生成一个簇	合并距离最近的两个簇
簇个数	始终为 k ，事先给定	由数据与邻域参数共同决定	每步减 1: $m \rightarrow m - 1 \rightarrow \dots \rightarrow k$
簇中心	每轮重算，移动至稳定	无簇中心	无簇中心
簇成员	样本可在簇间转移	簇生成后不变	只合并，不拆分
停止条件	参数与归属趋于稳定	核心对象全部访问	簇数达到 k
簇的形状	凸形（椭球）为主	任意形状	取决于簇间距离

记号 周志华《机器学习》第 9 章 9.4–9.7 节，pp.202–219； k 均值类算法仅在凸形簇结构上效果较好。

04

EM 算法

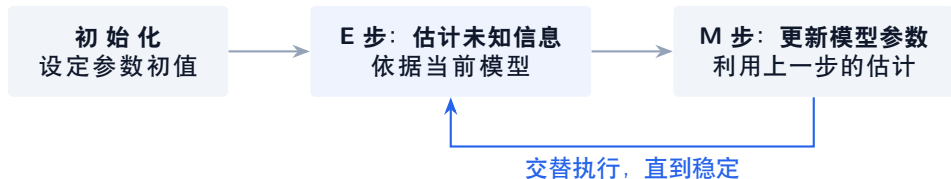
The EM Algorithm

4.1 用途与隐变量 · 4.2 E 步与 M 步 · 4.3 数学推导与收敛性

4.1 EM：部分信息未知时的参数估计

期望最大化（Expectation–Maximization, EM）：一种交替迭代的参数学习方法

用途 利用已观测的数据，在部分信息未知（隐变量）时估计模型参数。



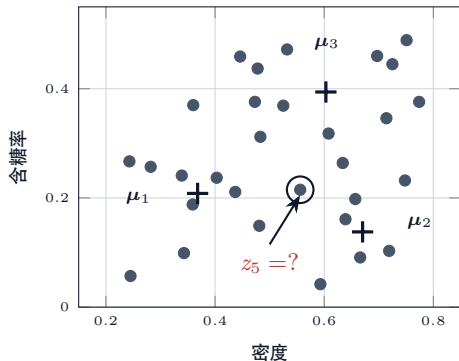
聚类中的例子

样本特征已观测到；样本归属与簇的参数需要估计。

记号 E: Expectation (期望); M: Maximization (最大化)。EM 常用于混合模型、缺失数据等问题。周志华《机器学习》7.6 节, pp.162–163。

k 均值中的隐变量 z

z_j 表示第 j 个样本所属的簇；样本特征可见，簇归属需要推断



已观测数据 $X = \{x_j\}$

密度、含糖率：

$$x_5 = (0.556, 0.215)^T$$

隐变量 $Z = \{z_j\}$

每个样本的簇编号。

$k = 3$ 时, $z_5 \in \{1, 2, 3\}$ 。

模型参数 $\Theta = \{\mu_i\}$

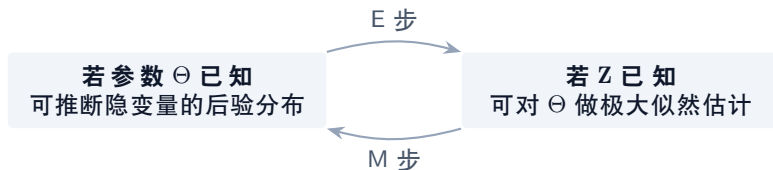
三个簇的中心；图中 + 为当前估计。

隐变量 (latent variable):
未直接观测到的变量。

记号 圈出的是 5 号西瓜；数据来自西瓜数据集 4.0。用 z_j 表示未知簇归属，对它作硬指派得到前文的簇标记 λ_j 。

4.2 EM 算法的思想

E 步估计隐变量的分布，M 步利用该分布更新模型参数



以初值 Θ^0 为起点，交替执行两步直到收敛：

E 步 (Expectation)：以 Θ^t 推断隐变量分布 $P(\mathbf{Z} | \mathbf{X}, \Theta^t)$ ，求对数似然关于 $\mathbf{Z}$ 的期望

$$Q(\Theta | \Theta^t) = \mathbb{E}_{\mathbf{Z} | \mathbf{X}, \Theta^t} LL(\Theta | \mathbf{X}, \mathbf{Z})$$

M 步 (Maximization)：寻找使完整数据对数似然的期望最大的参数

$$\Theta^{t+1} = \arg \max_{\Theta} Q(\Theta | \Theta^t)$$

记号 $LL(\Theta | \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} | \Theta)$ ； Θ^t ：第 t 轮参数。连续隐变量时求和换为积分。周志华《机器学习》7.6 节，p.163。

4.3 数学推导：对数似然的下界

观测数据的对数似然 $LL(\Theta | \mathbf{X}) = \ln P(\mathbf{X} | \Theta)$ ；对任意分布 $q(\mathbf{Z})$ 构造下界

$$LL(\Theta | \mathbf{X}) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{P(\mathbf{X}, \mathbf{Z} | \Theta)}{q(\mathbf{Z})} \geq \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{P(\mathbf{X}, \mathbf{Z} | \Theta)}{q(\mathbf{Z})}}_{F(q, \Theta)}$$

下界与对数似然之差

$$\begin{aligned} LL(\Theta | \mathbf{X}) - F(q, \Theta) \\ = \text{KL}(q \| P(\mathbf{Z} | \mathbf{X}, \Theta)) \geq 0 \end{aligned}$$

q 等于当前后验时，KL 散度为 0，
下界等于对数似然。

下界的分解

$$\begin{aligned} F(q, \Theta) &= \sum_{\mathbf{Z}} q \ln P(\mathbf{X}, \mathbf{Z} | \Theta) \\ &\quad - \sum_{\mathbf{Z}} q \ln q \end{aligned}$$

取 $q = P(\mathbf{Z} | \mathbf{X}, \Theta^t)$ ：
第一项为 $Q(\Theta | \Theta^t)$ ；
第二项与 Θ 无关。

记号 Jensen 不等式用于凹函数 $\ln$ ；上式假定联合概率非零处 $q(\mathbf{Z}) > 0$ 。KL：Kullback-Leibler 散度； Q 为完整数据对数似然的后验期望。Bishop (2006) §9.4，式 (9.70)。

4.3 数学推导：E 步的下界更新

固定参数 Θ^t ，将 q 更新为当前隐变量后验；观测数据的似然保持不变

1. 参数固定，观测似然不变

$$LL_{\text{E 步之后}} = LL_{\text{E 步之前}} = \ln P(\mathbf{X} | \Theta^t)$$

2. 取后验分布，让下界贴紧似然

$$q^t(\mathbf{Z}) = P(\mathbf{Z} | \mathbf{X}, \Theta^t) \implies F(q^t, \Theta^t) = LL(\Theta^t | \mathbf{X})$$

3. 下界的增量非负

$$F(q^t, \Theta^t) - F(q^{\text{old}}, \Theta^t) = \text{KL}(q^{\text{old}} \| q^t) \geq 0$$

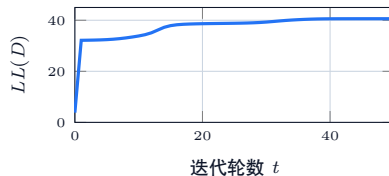
记号 q^{old} ：E 步更新前的分布； q^t ：本轮 E 步得到的后验分布。E 步改变 q ，不改变 Θ^t ，因此提高的是下界，而观测似然不变。Bishop (2006) §9.4。

4.3 数学推导：M 步的似然不减性

固定 E 步得到的 q^t ，提高下界；利用起点处的相等关系，保证观测似然不减

$$\begin{aligned} LL(\Theta^{t+1} | \mathbf{X}) &\geq F(q^t, \Theta^{t+1}) && \text{新参数处：似然不低于下界} \\ &\geq F(q^t, \Theta^t) && \text{M 步：提高固定 } q^t \text{ 下的下界} \\ &= LL(\Theta^t | \mathbf{X}) && \text{E 步已使起点处下界贴紧} \end{aligned}$$

- E 步：观测似然不变
- M 步：观测似然不减
- 似然有上界时，似然值单调收敛



西瓜数据集 4.0 上高斯混合聚类 ($k=3$) 的对数似然

记号 M 步最大化 $F(q^t, \Theta)$ ，等价于最大化 $Q(\Theta | \Theta^t)$ 。似然值收敛不保证全局最优。自由协方差 GMM 的似然可能无界，需限制协方差；右图 $t=0$ 为初值。Bishop (2006) §9.2.1、§9.4。

05

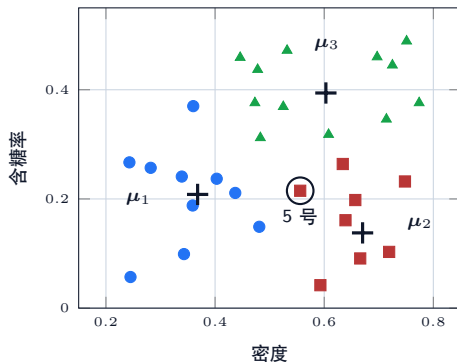
EM 求解 k 均值

Hard-assignment EM for k -means

5.1 E 步：估计簇归属 · 5.2 M 步：更新中心

5.1 k 均值的 E 步：估计簇归属 z

固定当前中心，将每个样本分配给最近的中心，得到 z_j 的估计



固定：中心 μ_1, μ_2, μ_3

更新：簇归属 $z_1, \dots, z_{30}$

5 号样本到三个中心的距离：

μ_1	μ_2	μ_3
0.188	0.138	0.185

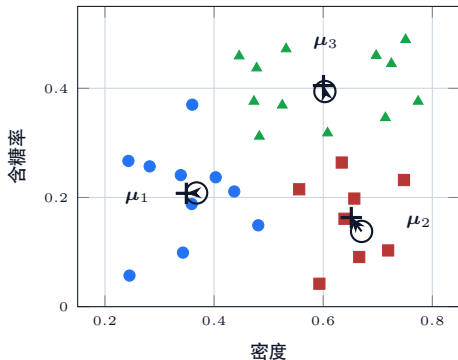
最近的是 μ_2 ：
估计 $z_5 = 2$ ，划入第 2 簇。

图中：中心不动，点按归属着色。

记号 这里采用硬指派 EM 的视角：每个样本选一个簇。标准 EM 的 E 步计算隐变量的后验分布，并据此求期望； k 均值只保留一个簇归属。

5.2 k 均值的 M 步: 更新中心 μ

固定 E 步得到的簇归属, 把每个中心移到该簇样本的平均位置



固定: 簇归属 $z_1, \dots, z_{30}$

更新: 中心 μ_1, μ_2, μ_3

例如: 对归入第 2 簇的 8 个样本取平均, 得到

$$\mu_2 = (0.6515, 0.16325)^T$$

图中: 点的颜色不变, 中心移动。

○: 旧中心; +: 新中心。

E 步估计 z , M 步更新 μ ;
反复交替, 直到稳定。

记号 西瓜数据集 4.0, 同一轮 E、M 步; 均值使簇内平方距离和最小。 k 均值与 EM 的联系见周志华《机器学习》7.7 节, p.165; 9.7 节, p.218。

06

小结

Summary

6.1 本讲要点 · 6.2 中英文术语对照 · 6.3 课后作业

6.1 本讲要点

性能度量	外部指标 JC、FMI、RI（越大越好）；内部指标 DBI（越小越好）、DI（越大越好）
距离	四条基本性质；闵可夫斯基距离（ $p = 2$ 欧氏， $p = 1$ 曼哈顿）
原型聚类	k 均值、高斯混合；簇数事先给定，成员与中心迭代调整
密度与层次	DBSCAN 的簇数由数据与邻域参数决定；AGNES 每次合并最近的两个簇
EM 算法	E 步推断隐变量分布，M 步最大化完整数据对数似然的期望；对数似然单调不减
联系	k 均值的硬指派 EM 视角：E 步估计簇归属 z ，M 步更新中心 μ

6.2 中英文术语对照

英文全称	缩写	中文
Jaccard Coefficient	JC	Jaccard 系数
Fowlkes and Mallows Index	FMI	FM 指数
Rand Index	RI	Rand 指数
Davies–Bouldin Index	DBI	DB 指数
Dunn Index	DI	Dunn 指数
True Positive	TP	真正例
False Positive	FP	假正例
False Negative	FN	假反例
True Negative	TN	真反例

英文全称	缩写	中文
Gaussian Mixture Model	GMM	高斯混合模型
Expectation–Maximization	EM	期望最大化
Density-Based Spatial Clustering of Applications with Noise	DBSCAN	基于密度的带噪声空间聚类
AGglomerative NESTing	AGNES	聚合嵌套
DIvisive ANAlysis	DIANA	分拆分析

6.3 课后作业

1. EM 算法的推导

写出含隐变量模型的对数边际似然 $LL(\Theta | \mathbf{X})$ ，用 Jensen 不等式构造其下界 $F(q, \Theta)$ ，说明等号成立的条件，并证明 EM 的每一轮迭代都有 $LL(\Theta^{t+1} | \mathbf{X}) \geq LL(\Theta^t | \mathbf{X})$ 。

2. 高斯混合的 M 步

由 $\partial LL(D)/\partial \mu_i = 0$ 推出 $\mu_i = \sum_j \gamma_{ji} \mathbf{x}_j / \sum_j \gamma_{ji}$ ；用拉格朗日乘子法推出 $\alpha_i = \frac{1}{m} \sum_j \gamma_{ji}$ 。

3. 编程

在西瓜数据集 4.0 上实现 k 均值，比较三组不同的 k 值与初始均值向量，讨论什么样的初值有利于得到好的结果（周志华《机器学习》第 9 章习题 9.4，p.220）。