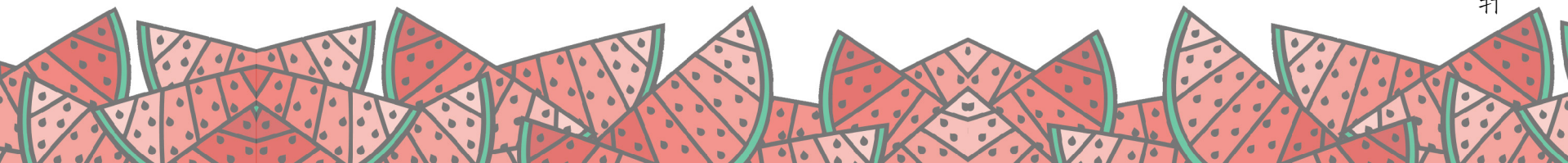
周志华 著

MACHINE LEARNING

机器学习

清华大学出版社

本章课件致谢：
霍轩



章节目录

- 贝叶斯决策论
- 极大似然估计
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网
- **EM算法**

EM算法

- “不完整”的样本：西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”，则训练样本的“根蒂”属性变量值未知，如何计算？

- 未观测的变量称为“隐变量”（latent variable）。令 $\mathbf{X}$ 表示已观测变量集，$\mathbf{Z}$ 表示隐变量集，若预对模型参数 Θ 做极大似然估计，则应最大化对数似然函数

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \qquad (7.34)$$

EM算法

EM算法流程

输入：观察数据 $x = (x^{(1)}, x^{(2)}, \ldots x^{(m)})$，联合分布 $p(x, z|\theta)$，条件分布 $p(z|x, \theta)$，极大迭代次数 J 。

1) 随机初始化模型参数 θ 的初值 θ^0

2) for j from 1 to J :

- E步：计算联合分布的条件概率期望：

$$Q_i(z^{(i)}) := P(z^{(i)}|x^{(i)}, \ \theta))$$

- M步：极大化 $L(\theta)$,得到 θ ：

$$\theta := arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) logP(x^{(i)}, \ z^{(i)}|\theta)$$

- 重复E、M步骤直到 θ 收敛

第九章：聚类

章节目录

- **聚类任务**
- 性能度量
- 距离计算
- K均值聚类
- 高斯混合聚类

聚类任务

- 聚类：经典的无监督学习方法，无监督学习的目标是通过对无标记训练样本的学习，发掘和揭示数据集本身潜在的结构与规律，即不依赖于训练数据集的类标记信息。聚类则是试图将数据集的样本划分为若干个互不相交的类簇，从而每个簇对应一个潜在的类别。
- 聚类过程仅能自动形成簇结构，簇说对应的概念语义需要使用者来把我和命名。

聚类任务

- 聚类既可以作为一个单独过程，用于寻找数据内在的分布结构；也可以作为分类等任务的前驱过程。聚类直观上来说是将相似的样本聚在一起，从而形成一个类簇（cluster）。那首先的问题是如何来度量相似性（similarity measure）呢？这便是距离度量，在生活中我们说差别小则相似，对应到多维样本，每个样本可以对应于高维空间中的一个数据点，若它们的距离相近，我们便可以称它们相似。那接着如何来评价聚类结果的好坏呢？这便是性能度量，性能度量为评价聚类结果的好坏提供了一系列有效性指标。

章节目录

- 聚类任务
- **性能度量**
- 距离计算
- K均值聚类
- 高斯混合聚类

性能度量

- 聚类的性能度量又叫“有效性指标”；
  - 簇内相似度：越高越好；
  - 簇间相似度：越低越好；
- **外部指标**：将聚类结果与某个“参考模型”进行比较；如：Jaccard系数、FM指数、Rand指数等；
- **内部指标**：直接考察聚类结果而不利于任何参考模型；如：DB指数、Dunn指数。

章节目录

- 聚类任务
- 性能度量
- **距离计算**
- K均值聚类
- 高斯混合聚类

距离计算

- 距离度量dist(x,y)需要满足的一些基本性质：
  - 非负性： dist(x,y) ≥ 0；
  - 同一性： dist(x,y) = 0当且仅当x = y ；
  - 对称性： dist(x,y) = dist(y,x) ；
  - 直递性： dist(x,y) ≤ dist(x,z) + dist(z,y) ；
- **常用距离度量**：
  - 闵可夫斯基距离(Minkowski distance)；
  - 欧氏距离(Euclidean distance)；
  - 曼哈顿距离(Manhattan distance)；

章节目录

- 聚类任务
- 性能度量
- 距离计算
- **K均值聚类**
- 高斯混合聚类

K均值聚类

给定样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$, “$k$ 均值” (k-means)算法针对聚类所得簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ 最小化平方误差

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} ||\boldsymbol{x} - \boldsymbol{\mu}_i||_2^2 , \tag{9.24}$$

其中 $\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$ 是簇 C_i 的均值向量. 直观来看, 式(9.24) 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.

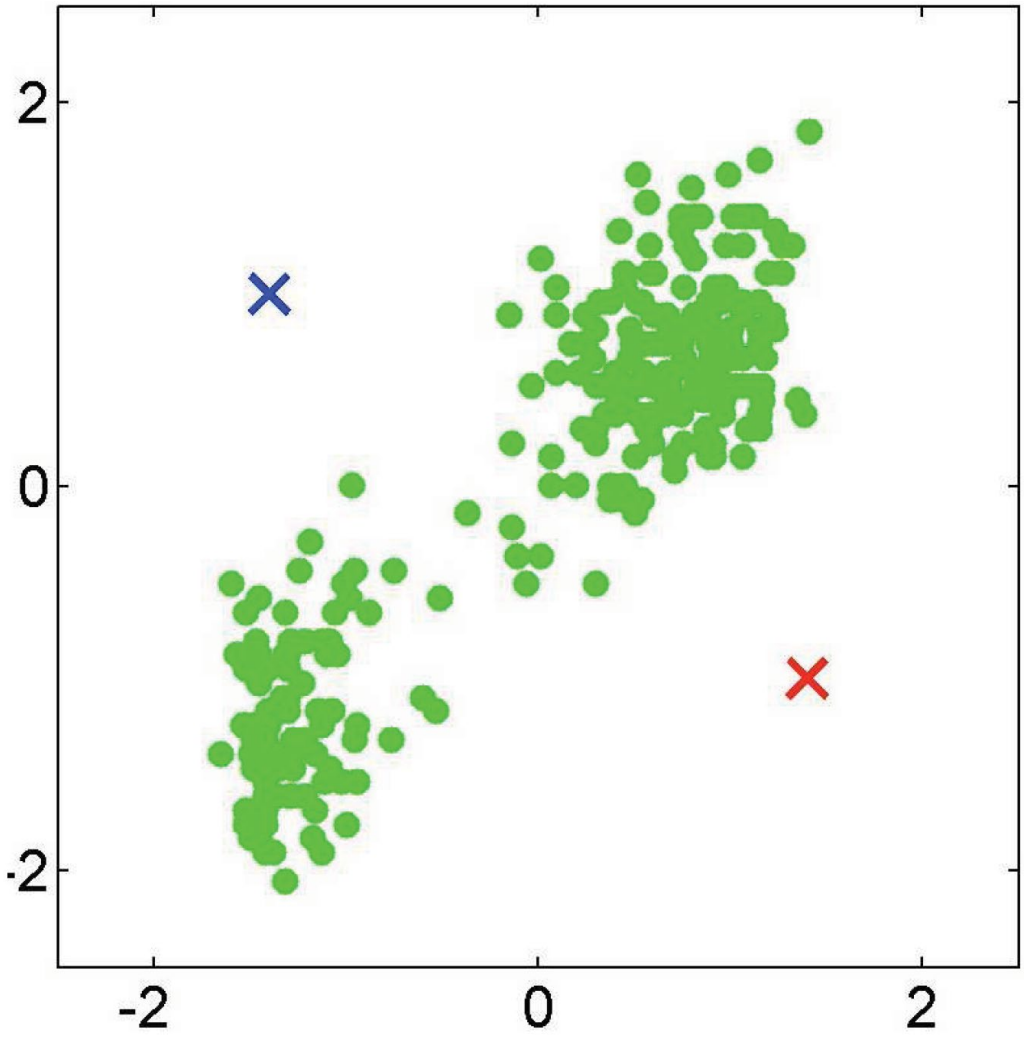
过程:

1: 从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$
2: **repeat**
3: 令 $C_i = \varnothing$ $(1 \leqslant i \leqslant k)$
4: **for** $j = 1, 2, \ldots, m$ **do**
5: 计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i$ $(1 \leqslant i \leqslant k)$ 的距离: $d_{ji} = ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||_2$;
6: 根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j = \arg\min_{i \in \{1,2,\ldots,k\}} d_{ji}$;
7: 将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{\boldsymbol{x}_j\}$;
8: **end for**
9: **for** $i = 1, 2, \ldots, k$ **do**
10: 计算新均值向量: $\boldsymbol{\mu}'_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$;
11: **if** $\boldsymbol{\mu}'_i \neq \boldsymbol{\mu}_i$ **then**
12: 将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}'_i$
13: **else**
14: 保持当前均值向量不变
15: **end if**
16: **end for**
17: **until** 当前均值向量均未更新

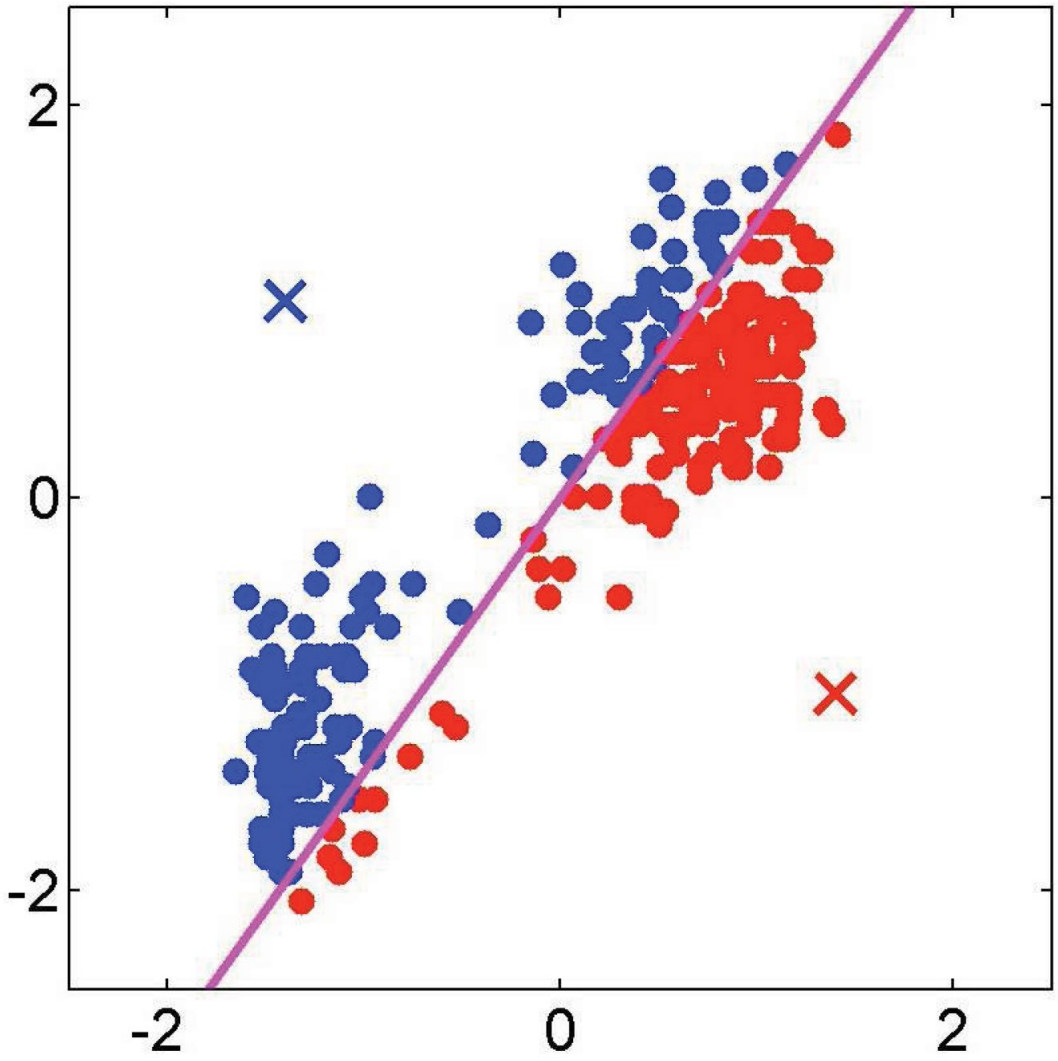
输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

为避免运行时间过长, 通常设置一个最大运行轮数或最小调整幅度阈值, 若达到最大轮数或调整幅度小于阈值, 则停止运行.

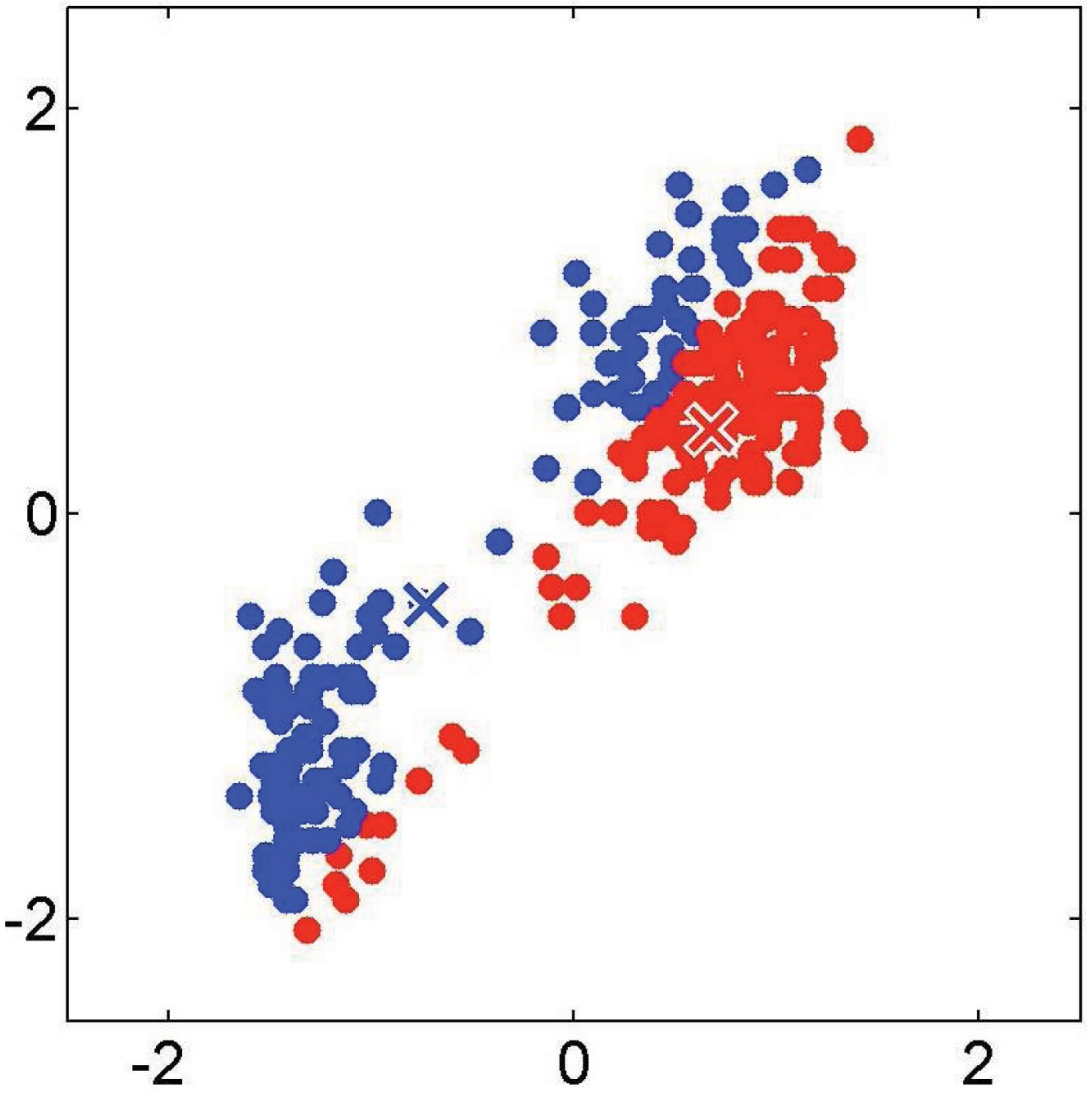
K均值聚类



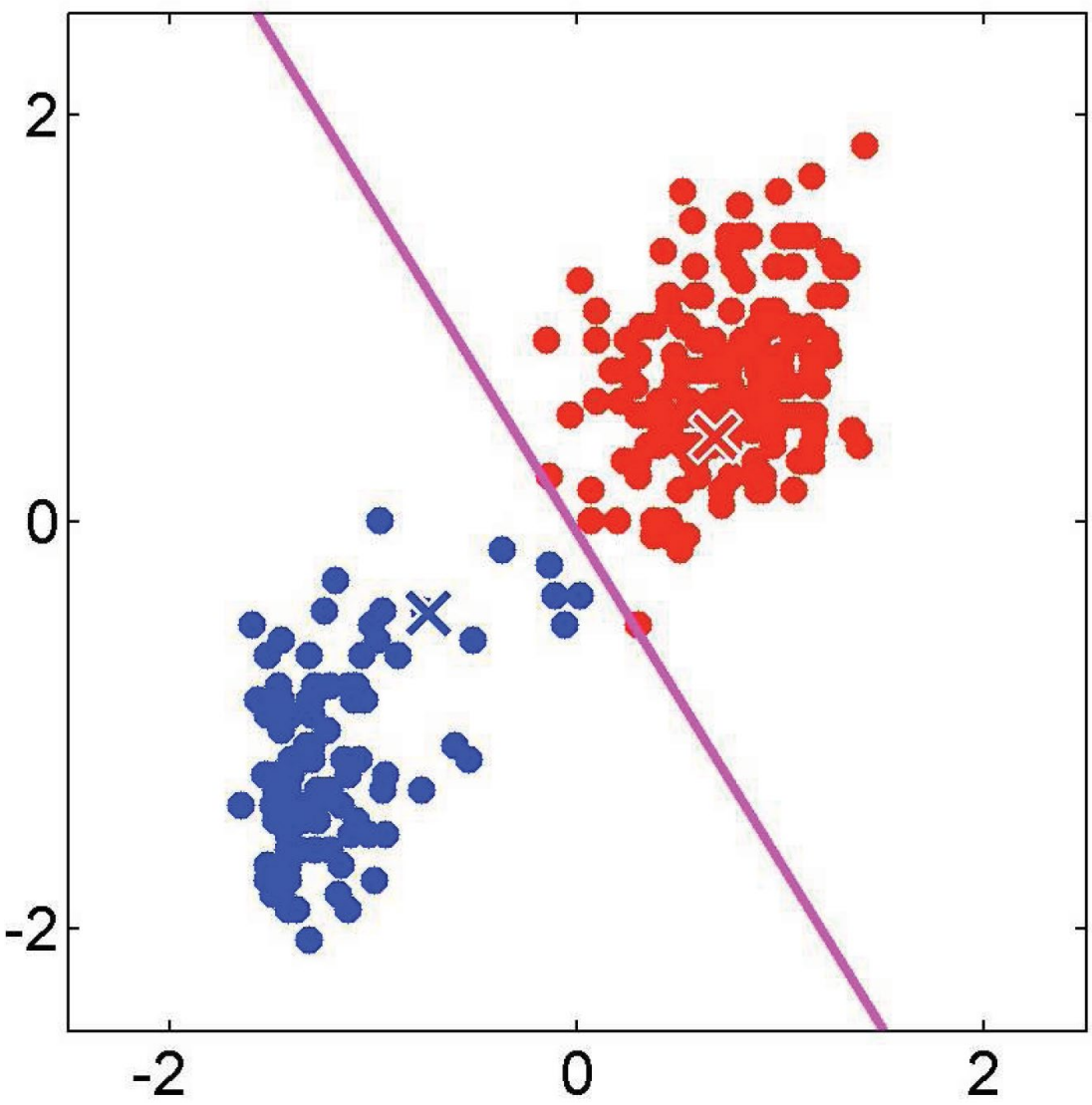
K均值聚类



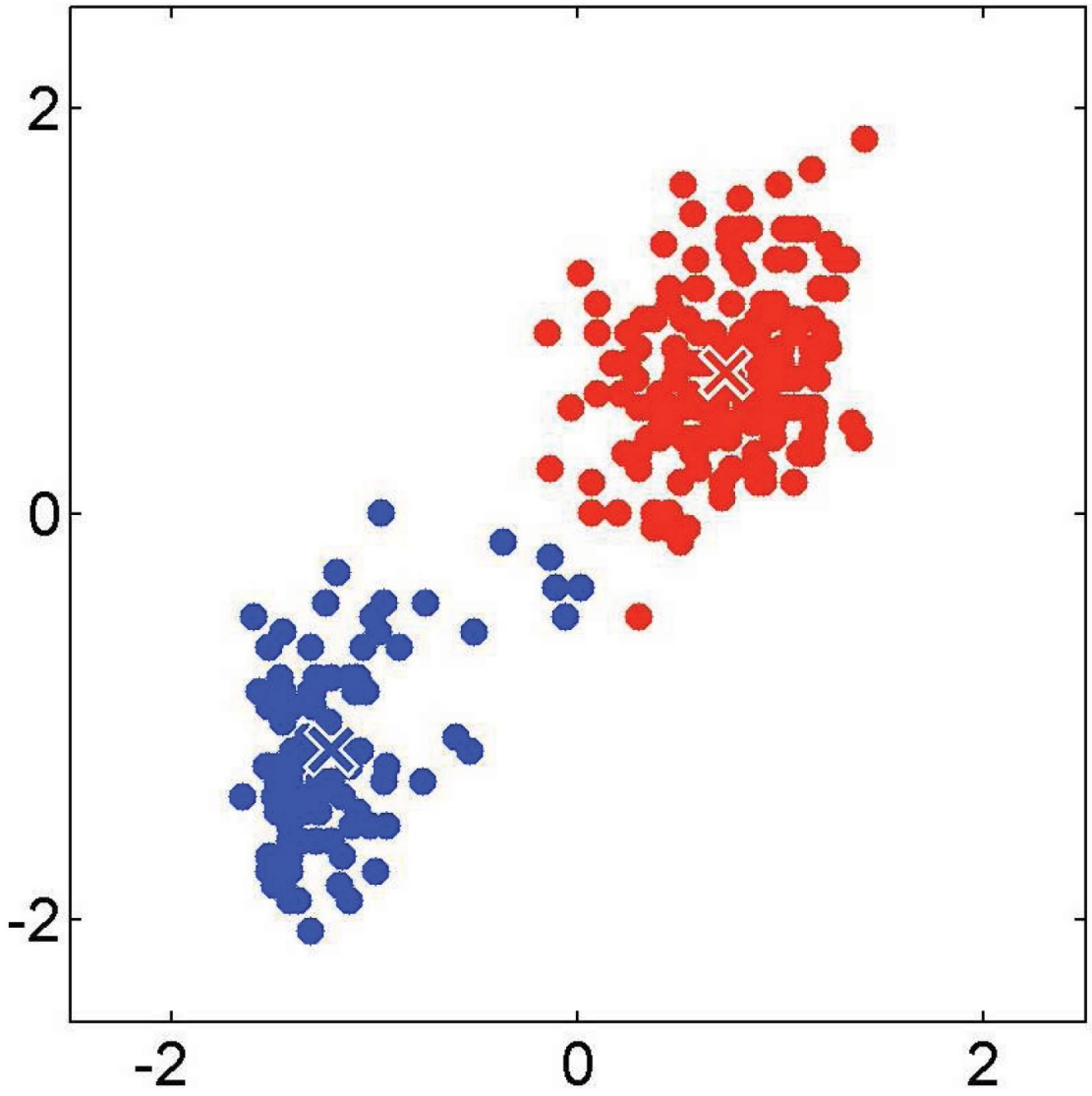
K均值聚类



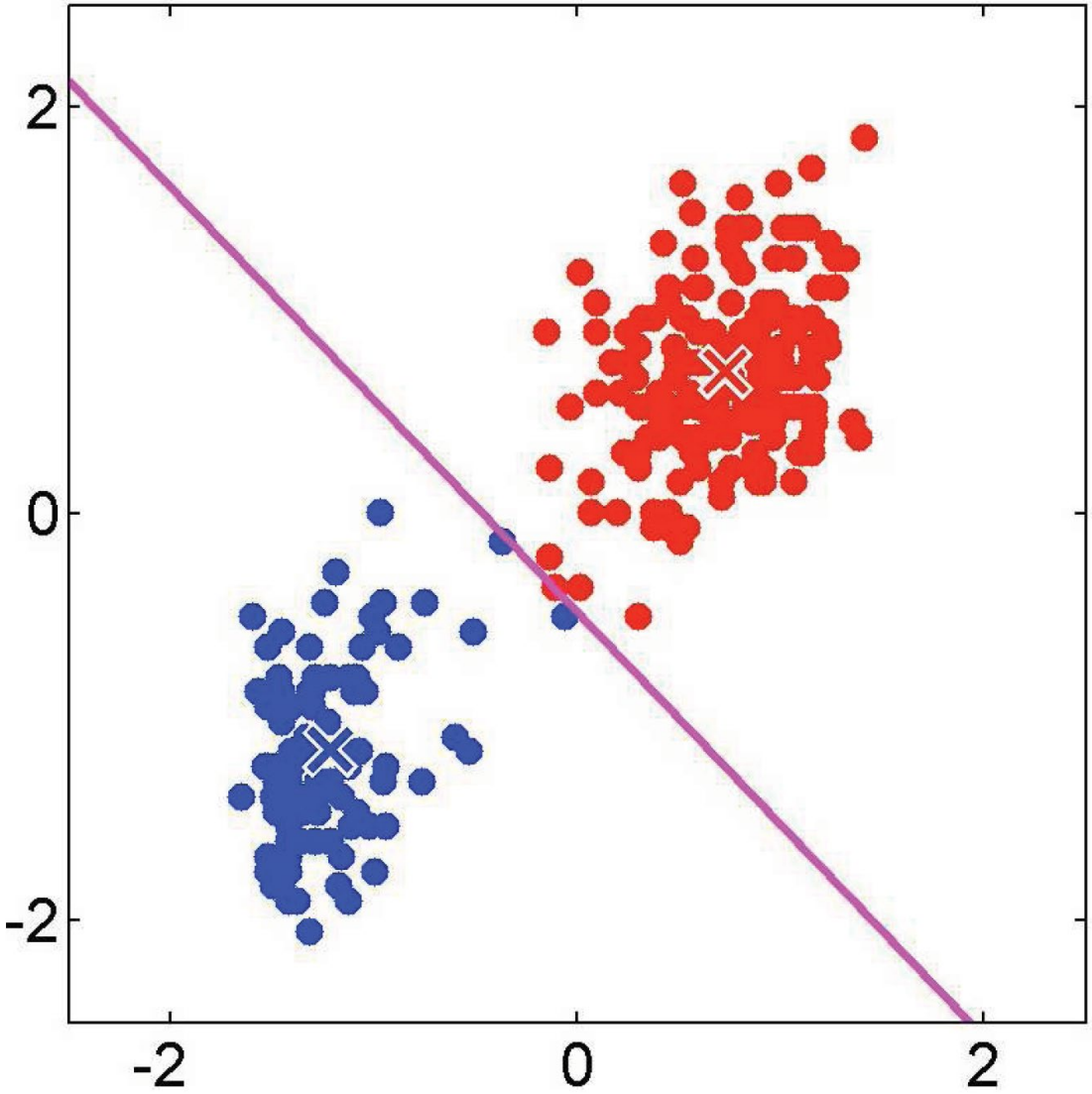
K均值聚类



K均值聚类



K均值聚类



章节目录

- 聚类任务
- 性能度量
- 距离计算
- K均值聚类
- **高斯混合聚类**

高斯混合聚类

- （多元）高斯分布的定义：

对 n 维样本空间 $\mathcal{X}$ 中的随机向量 $\boldsymbol{x}$，若 $\boldsymbol{x}$ 服从高斯分布，其概率密度函数为

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{9.28}$$

记为 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$\boldsymbol{\Sigma}$: 对称正定矩阵;
$|\boldsymbol{\Sigma}|$: $\boldsymbol{\Sigma}$ 的行列式;
$\boldsymbol{\Sigma}^{-1}$: $\boldsymbol{\Sigma}$ 的逆矩阵.

其中 $\boldsymbol{\mu}$ 是 n 维均值向量，$\boldsymbol{\Sigma}$ 是 $n \times n$ 的协方差矩阵.

- 高斯混合分布：

$$p_{\mathcal{M}}(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{9.29}$$

$p_{\mathcal{M}}(\cdot)$ 也是概率密度函数，$\int p_{\mathcal{M}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$.

该分布共由 k 个混合成分组成，每个混合成分对应一个高斯分布. 其中 $\boldsymbol{\mu}_i$ 与 $\boldsymbol{\Sigma}_i$ 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”(mixture coefficient)，$\sum_{i=1}^{k} \alpha_i = 1$.

高斯混合聚类

- （多元）高斯分布的定义：

对 n 维样本空间 $\mathcal{X}$ 中的随机向量 $\boldsymbol{x}$，若 $\boldsymbol{x}$ 服从高斯分布，其概率密度函数为

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{9.28}$$

记为 $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$\boldsymbol{\Sigma}$: 对称正定矩阵;
$|\boldsymbol{\Sigma}|$: $\boldsymbol{\Sigma}$ 的行列式;
$\boldsymbol{\Sigma}^{-1}$: $\boldsymbol{\Sigma}$ 的逆矩阵.

其中 $\boldsymbol{\mu}$ 是 n 维均值向量，$\boldsymbol{\Sigma}$ 是 $n \times n$ 的协方差矩阵.

- 高斯混合分布：

$$p_{\mathcal{M}}(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{9.29}$$

$p_{\mathcal{M}}(\cdot)$ 也是概率密度函数，$\int p_{\mathcal{M}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$.

该分布共由 k 个混合成分组成，每个混合成分对应一个高斯分布. 其中 $\boldsymbol{\mu}_i$ 与 $\boldsymbol{\Sigma}_i$ 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”(mixture coefficient)，$\sum_{i=1}^{k} \alpha_i = 1$.

高斯混合聚类

该分布共由 k 个混合成分组成, 每个混合成分对应一个高斯分布. 其中 $\boldsymbol{\mu}_i$ 与 $\boldsymbol{\Sigma}_i$ 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”(mixture coefficient), $\sum_{i=1}^{k} \alpha_i = 1$.

假设样本的生成过程由高斯混合分布给出: 首先, 根据 $\alpha_1, \alpha_2, \ldots, \alpha_k$ 定义的先验分布选择高斯混合成分, 其中 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.

若训练集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ 由上述过程生成, 令随机变量 $z_j \in \{1, 2, \ldots, k\}$ 表示生成样本 $\boldsymbol{x}_j$ 的高斯混合成分, 其取值未知. 显然, z_j 的先验概率 $P(z_j = i)$ 对应于 α_i $(i = 1, 2, \ldots, k)$. 根据贝叶斯定理, z_j 的后验分布对应于

$$
\begin{aligned}
p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j) &= \frac{P(z_j = i) \cdot p_{\mathcal{M}}(\boldsymbol{x}_j \mid z_j = i)}{p_{\mathcal{M}}(\boldsymbol{x}_j)} \\
&= \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} .
\end{aligned} \tag{9.30}
$$

高斯混合聚类

换言之, $p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j)$ 给出了样本 $\boldsymbol{x}_j$ 由第 i 个高斯混合成分生成的后验概率. 为方便叙述, 将其简记为 γ_{ji} $(i = 1, 2, \ldots, k)$.

当高斯混合分布(9.29)已知时, 高斯混合聚类将把样本集 D 划分为 k 个簇 $C = \{C_1, C_2, \ldots, C_k\}$, 每个样本 $\boldsymbol{x}_j$ 的簇标记 λ_j 如下确定:

$$\lambda_j = \underset{i \in \{1,2,\ldots,k\}}{\arg\max} \ \gamma_{ji} \ . \tag{9.31}$$

因此, 从原型聚类的角度来看, 高斯混合聚类是采用概率模型(高斯分布)对原型进行刻画, 簇划分则由原型对应后验概率确定.

那么, 对于式(9.29), 模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$ 如何求解呢? 显然, 给定样本集 D, 可采用极大似然估计, 即最大化(对数)似然

$$\begin{aligned} LL(D) &= \ln \left(\prod_{j=1}^{m} p_{\mathcal{M}}(\boldsymbol{x}_j) \right) \\ &= \sum_{j=1}^{m} \ln \left(\sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) \ , \end{aligned} \tag{9.32}$$

高斯混合聚类

常采用 EM 算法进行迭代优化求解. 下面我们做一个简单的推导.

若参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$ 能使式(9.32)最大化, 则由 $\frac{\partial LL(D)}{\partial \boldsymbol{\mu}_i} = 0$ 有

$$\sum_{j=1}^{m} \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) = 0 , \tag{9.33}$$

由式(9.30)以及 $\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j)$, 有

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{m} \gamma_{ji} \boldsymbol{x}_j}{\sum_{j=1}^{m} \gamma_{ji}} , \tag{9.34}$$

即各混合成分的均值可通过样本加权平均来估计, 样本权重是每个样本属于该成分的后验概率. 类似的, 由 $\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = 0$ 可得

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}}{\sum_{j=1}^{m} \gamma_{ji}} . \tag{9.35}$$

高斯混合聚类

对于混合系数 α_i, 除了要最大化 $LL(D)$, 还需满足 $\alpha_i \geqslant 0$, $\sum_{i=1}^{k} \alpha_i = 1$. 考虑 $LL(D)$ 的拉格朗日形式

$$LL(D) + \lambda \left(\sum_{i=1}^{k} \alpha_i - 1 \right), \tag{9.36}$$

其中 λ 为拉格朗日乘子. 由式(9.36)对 α_i 的导数为 0, 有

$$\sum_{j=1}^{m} \frac{p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} + \lambda = 0 , \tag{9.37}$$

两边同乘以 α_i, 对所有混合成分求和可知 $\lambda = -m$, 有

$$\alpha_i = \frac{1}{m} \sum_{j=1}^{m} \gamma_{ji} , \tag{9.38}$$

即每个高斯成分的混合系数由样本属于该成分的平均后验概率确定.

高斯混合聚类

由上述推导即可获得高斯混合模型的 EM 算法: 在每步迭代中, 先根据当前参数来计算每个样本属于每个高斯成分的后验概率 γ_{ji} (E步), 再根据式(9.34)、(9.35)和(9.38)更新模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$ (M 步).

高斯混合聚类

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 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$
repeat
  for $j = 1, 2, \ldots, m$ **do**
    根据式(9.30)计算 $\boldsymbol{x}_j$ 由各混合成分生成的后验概率, 即
    $\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j)$ $(1 \leqslant i \leqslant k)$
  end for
  for $i = 1, 2, \ldots, k$ **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: $\boldsymbol{\Sigma}_i' = \frac{\sum_{j=1}^{m} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i')(\boldsymbol{x}_j - \boldsymbol{\mu}_i')^{\mathrm{T}}}{\sum_{j=1}^{m} \gamma_{ji}}$;
    计算新混合系数: $\alpha_i' = \frac{\sum_{j=1}^{m} \gamma_{ji}}{m}$;
  end for
  将模型参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant k\}$ 更新为 $\{(\alpha_i', \boldsymbol{\mu}_i', \boldsymbol{\Sigma}_i') \mid 1 \leqslant i \leqslant k\}$
until 满足停止条件
$C_i = \varnothing$ $(1 \leqslant i \leqslant k)$
for $j = 1, 2, \ldots, m$ **do**
  根据式(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

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

EM 算法的 E步.

EM 算法的 M步.

例如达到最大迭代轮数.

Task

- 完成高斯混合模型聚类代码，分为以下模块：
- （1）预处理数据（读取，数据类型）；
- （2） option初始化（最大迭代次数，ε）；
- （3）聚类中心初始化（随机生成k个）；
- （4）参数初始化（α, μ, Σ）；
- （5）EM算法；
- （6）评价指标（自选2~3个）；
- （7）聚类结果可视化；
- 要求：不直接使用python库中的函数。