

# 22

Expectation Maximization

## 最大期望算法

迭代优化两步走：E 步，M 步；最大化对数似然函数

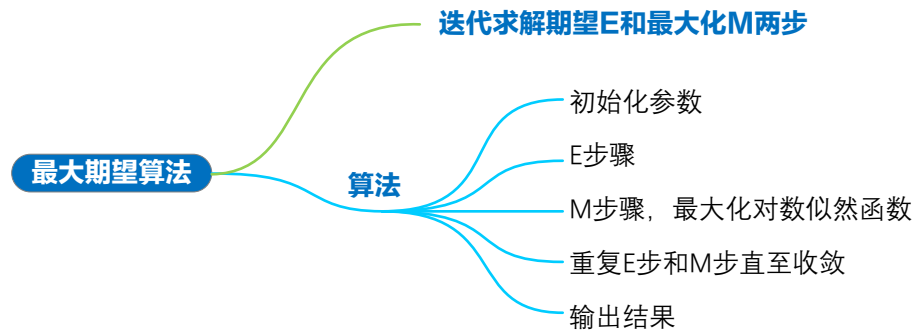


我解决的每个问题，都变成了定理法则；它们都被拿去解决更多的问题。

*Each problem that I solved became a rule, which served afterwards to solve other problems.*

—— 勒内·笛卡尔 (René Descartes) | 法国哲学家、数学家、物理学家 | 1596 ~ 1650





本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

## 22.1 最大期望

求解高斯混合模型 (Gaussian Mixture Model, GMM) 绕不开 **EM 算法**，即**最大期望算法** (Expectation Maximization, EM)。EM 算法是一种迭代算法，其核心思想是在不完全观测的情况下，通过已知的观测数据来估计模型参数。

上一章介绍的高斯混合模型核心思想是，叠加若干高斯分布来描述样本数据分布。一元高斯分布有两个重要参数，均值和均方差；而多元高斯分布则通过质心和协方差来描述。除此，我们还需要知道每个高斯分布分量的贡献，即先验概率值。遗憾的是，这几个参数不能通过解析方法求解。

本章介绍的最大期望算法正是求解高斯混合模型参数的方法。

### E 步、M 步

EM 算法是一个收敛迭代过程。EM 算法两个步骤交替进行迭代：

- ▶ 第一步 (即所谓 E 步)，利用当前参数  $\theta$  计算期望值，并计算对数似然函数  $L(\theta)$ ；根据当前参数估计值计算每个数据点属于每个高斯分布的后验概率，即每个数据点在每个簇中的权重。
- ▶ 第二步 (即所谓 M 步)，在第一步基础上最大化，并更新参数  $\theta$ ；根据上一步中计算得到的后验概率重新估计每个高斯分布的均值、方差和系数，并更新参数估计值。

EM 算法不断迭代这两个步骤，直到收敛为止。在 GMM 中，EM 算法的收敛条件可以是参数变化的阈值或者似然函数的收敛。

## 22.2 E 步：最大化期望

本节以单一特征样本数据为例，可视化最大期望算法迭代过程。观察发现数据应该被分为两簇，设定  $K = 2$ 。

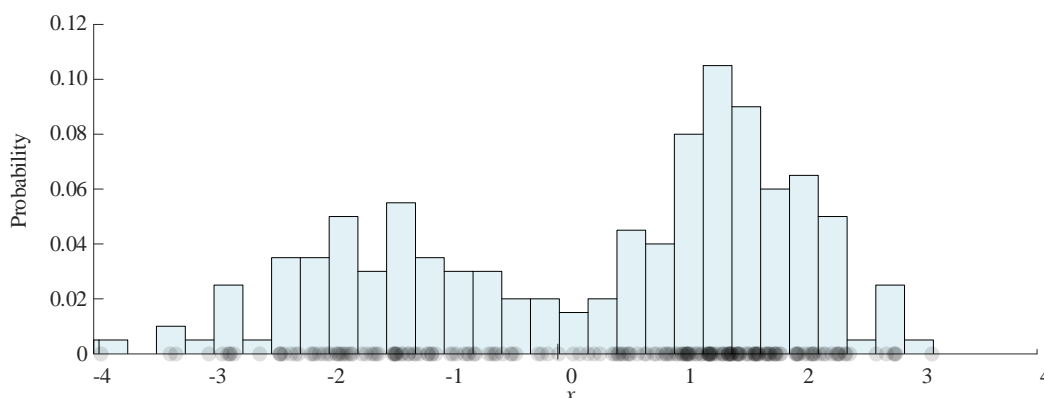


图 1. 一维样本待聚类样本数据

### 初始化

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

利用一元高斯分布叠加，首先初始化参数  $\theta$ ：

$$\theta^{(0)} = \{\alpha_1^{(0)}, \alpha_2^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{(0)}, \sigma_2^{(0)}\} \quad (1)$$

上角标  $^{(i)}$  代表当前迭代次数， $^{(0)}$  代表迭代初始。

选定初始化参数  $\theta$  具体数值如下：

$$\begin{cases} \alpha_1^{(0)} = p_Y(C_1, \theta^{(0)}) = 0.5, & \alpha_2^{(0)} = p_Y(C_2, \theta^{(0)}) = 0.5 \\ \mu_1^{(0)} = -0.05, & \mu_2^{(0)} = 0.05 \\ \sigma_1^{(0)} = \sigma_2^{(0)} = 1 \end{cases} \quad (2)$$

$\alpha_1$  和  $\alpha_2$  代表两个不同高斯分布对  $f_X(x)$  的贡献。

$\mu_1$  和  $\mu_2$  为期望值，描述两个正态分布质心位置。

$\sigma_1$  和  $\sigma_2$  为标准差，刻画正态分布离散程度。

## 似然概率

通过 (2) 给出六个参数，利用高斯分布估算得到  $f_{X|Y}(x | C_1, \theta^{(0)})$  和  $f_{X|Y}(x | C_2, \theta^{(0)})$  的两个似然概率 PDF，具体如下：

$$\begin{cases} f_{X|Y}(x | C_1, \theta^{(0)}) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x - \mu_1}{\sigma_1}\right)^2\right)}{\sigma_1 \sqrt{2\pi}} = \frac{\exp\left(-\frac{1}{2}(x + 0.05)^2\right)}{\sqrt{2\pi}} \\ f_{X|Y}(x | C_2, \theta^{(0)}) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x - \mu_2}{\sigma_2}\right)^2\right)}{\sigma_2 \sqrt{2\pi}} = \frac{\exp\left(-\frac{1}{2}(x - 0.05)^2\right)}{\sqrt{2\pi}} \end{cases} \quad (3)$$

图 2 所示为初始化参数对应的初始化参数对应的  $f_{X|Y}(x | C_1)$  和  $f_{X|Y}(x | C_2)$  图像。

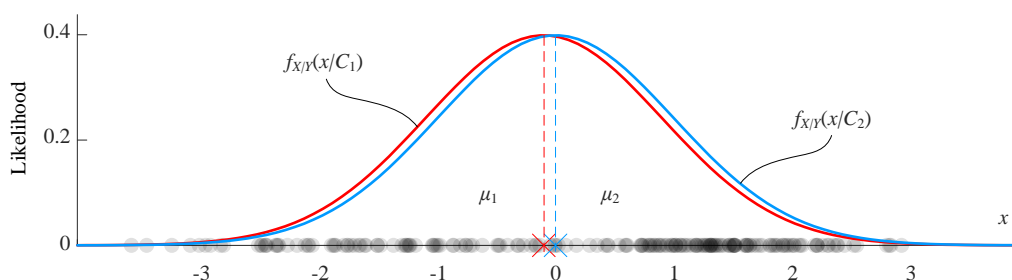


图 2. 初始化参数  $\theta^{(0)}$  对应的  $f_{X|Y}(x | C_1)$  和  $f_{X|Y}(x | C_2)$  图像

## 证据因子

下一步，估算概率密度函数  $f_X(x | \theta^{(0)})$ ：

$$\begin{aligned}
 f_X(x|\theta^{(0)}) &= f_{X,Y}(x, C_1, \theta^{(0)}) + f_{X,Y}(x, C_2, \theta^{(0)}) \\
 &= p_Y(C_1, \theta^{(0)}) f_{X|Y}(x|C_1, \theta^{(0)}) + p_Y(C_2, \theta^{(0)}) f_{X|Y}(x|C_2, \theta^{(0)})
 \end{aligned} \quad (4)$$

将 (2) 和 (3) 代入 (4)，整理得到：

$$f_X(x|\theta^{(0)}) = \frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x+0.05)^2\right)}{\sqrt{2\pi}} + \frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x-0.05)^2\right)}{\sqrt{2\pi}} \quad (5)$$

图 3 展示的是这一轮迭代  $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$  和  $f_X(x)$  结果图像。

根据本书第 9 章有关朴素贝叶斯分类介绍的内容，图 3 所示  $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$  曲线高度可以判断当前条件下数据聚类结果。图 3 中横轴数据点颜色代表本轮预测聚类结果。

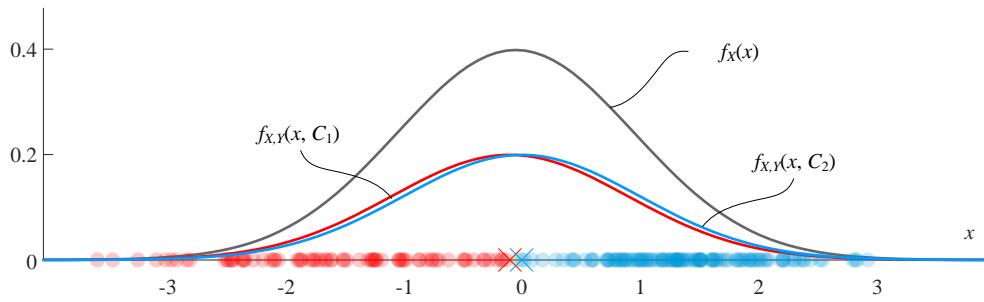


图 3. 初始化参数计算得到  $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$  和  $f_X(x)$

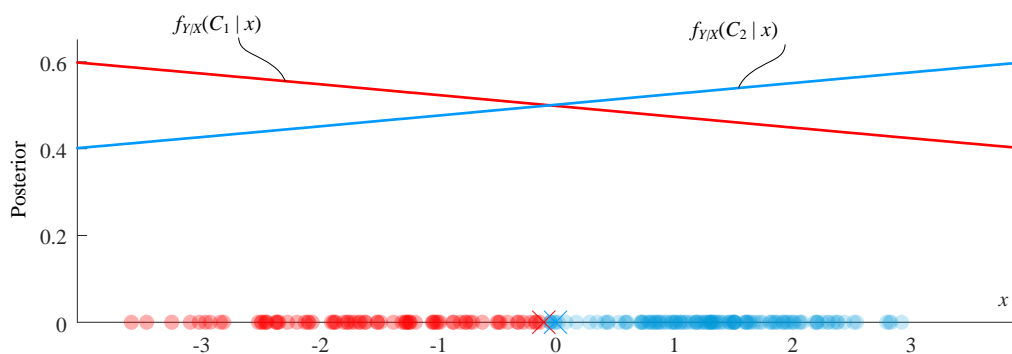
## 后验概率

根据贝叶斯定理，计算后验概率  $f_{Y|X}(C_1|x, \theta^{(0)})$  和  $f_{Y|X}(C_2|x, \theta^{(0)})$ ：

$$\begin{cases}
 f_{Y|X}(C_1|x, \theta^{(0)}) = \frac{p_Y(C_1, \theta^{(0)}) f_{X|Y}(x|C_1, \theta^{(0)})}{f_X(x|\theta^{(0)})} = \frac{\frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x+0.05)^2\right)}{\sqrt{2\pi}}}{\frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x+0.05)^2\right)}{\sqrt{2\pi}} + \frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x-0.05)^2\right)}{\sqrt{2\pi}}} \\
 f_{Y|X}(C_2|x, \theta^{(0)}) = \frac{p_Y(C_2, \theta^{(0)}) f_{X|Y}(x|C_2, \theta^{(0)})}{f_X(x|\theta^{(0)})} = \frac{\frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x-0.05)^2\right)}{\sqrt{2\pi}}}{\frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x+0.05)^2\right)}{\sqrt{2\pi}} + \frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(x-0.05)^2\right)}{\sqrt{2\pi}}}
 \end{cases} \quad (6)$$

图 4 给出初始参数条件下后验概率  $f_{Y|X}(C_1|x)$  和  $f_{Y|X}(C_2|x)$  随  $x$  变化。对于任意一点  $x$ ，下式成立：

$$f_{Y|X}(C_1|x) + f_{Y|X}(C_2|x) = 1 \quad (7)$$

图 4. 初始化参数计算得到后验概率  $f_{Y|X}(C_1 | x)$  和  $f_{Y|X}(C_2 | x)$ 

后验概率大小代表成员值，某一点不同簇后验值区分越大，分类才越有理有据。如果不同簇后验值区分不大，据此得到的分类预测则显得很牵强。因此，迭代优化还需要继续。

## 22.3 M 步：最大化似然概率

下一步是 EM 算法中非常重要的环节——更新参数、最大化似然概率。对于迭代 EM 算法，这便是 M 步。

### 先验概率

更新参数  $\alpha_1$  和  $\alpha_2$ ：

$$\begin{cases} \alpha_1^{(1)} = \frac{\sum_{i=1}^n f_{Y|X}(C_1 | x^{(i)}, \theta^{(0)})}{n} = 0.49379 \\ \alpha_2^{(1)} = \frac{\sum_{i=1}^n f_{Y|X}(C_2 | x^{(i)}, \theta^{(0)})}{n} = 0.50621 \end{cases} \quad (8)$$

$\alpha_1$  和  $\alpha_2$  相当于数据聚类比例。可以这样理解上式，一共有  $n$  个数据点，每个点有  $1/n$  的投票权。对二聚类问题， $1/n$  要分成两份，分别给  $C_1$  和  $C_2$ 。每个点的后验概率决定比例分配。

整理 (8) 可以得到如下等式：

$$\begin{cases} n\alpha_1^{(1)} = \sum_{i=1}^n f_{Y|X}(C_1 | x^{(i)}, \theta^{(0)}) \\ n\alpha_2^{(1)} = \sum_{i=1}^n f_{Y|X}(C_2 | x^{(i)}, \theta^{(0)}) \end{cases} \quad (9)$$

### 均值

利用当前每个样本数据估算得到的后验概率/成员值，更新  $\mu_1$  和  $\mu_2$ ：

$$\left\{ \begin{aligned} \mu_1^{(1)} &= \frac{\sum_{i=1}^n \underbrace{\left\{ f_{Y|X} \left( C_1 | x^{(i)}, \theta^{(0)} \right) \cdot x^{(i)} \right\}}_{\text{Membership score}}}{\sum_{i=1}^n f_{Y|X} \left( C_1 | x^{(i)}, \theta^{(0)} \right)} = \frac{\sum_{i=1}^n \left\{ f_{Y|X} \left( C_1 | x^{(i)}, \theta^{(0)} \right) \cdot x_i \right\}}{n\alpha_1^{(1)}} = 0.11073 \\ \mu_2^{(1)} &= \frac{\sum_{i=1}^n \underbrace{\left\{ f_{Y|X} \left( C_2 | x^{(i)}, \theta^{(0)} \right) \cdot x^{(i)} \right\}}_{\text{Membership score}}}{\sum_{i=1}^n f_{Y|X} \left( C_2 | x^{(i)}, \theta^{(0)} \right)} = \frac{\sum_{i=1}^n \left\{ f_{Y|X} \left( C_2 | x^{(i)}, \theta^{(0)} \right) \cdot x^{(i)} \right\}}{n\alpha_2^{(1)}} = 0.38248 \end{aligned} \right. \quad (10)$$

上式相当于求加权均值。后验概率/成员值相当于样本数据从属于不同聚类的权重。

## 标准差

同理，求加权方法，更新  $\sigma_1$  和  $\sigma_2$ ：

$$\left\{ \begin{aligned} \sigma_1^{(1)} &= \sqrt{\frac{\sum_{i=1}^n \underbrace{\left\{ f_{Y|X} \left( C_1 | x^{(i)}, \theta^{(0)} \right) \cdot \left( x^{(i)} - \mu_1^{(1)} \right)^2 \right\}}_{\text{Membership score}}}{N\alpha_1^{(1)}}} = 2.8303 \\ \sigma_2^{(1)} &= \sqrt{\frac{\sum_{i=1}^n \underbrace{\left\{ f_{Y|X} \left( C_2 | x^{(i)}, \theta^{(0)} \right) \cdot \left( x^{(i)} - \mu_2^{(1)} \right)^2 \right\}}_{\text{Membership score}}}{N\alpha_2^{(1)}}} = 2.5922 \end{aligned} \right. \quad (11)$$

## 全新参数

这样，我们便得到了一组全新的参数  $\theta^{(1)}$ ：

$$\left\{ \begin{aligned} \alpha_1^{(1)} &= p_Y(C_1) = 0.49379, & \alpha_2^{(1)} &= p_Y(C_2) = 0.50621 \\ \mu_1^{(1)} &= 0.11073, & \mu_2^{(1)} &= 0.38248 \\ \sigma_1^{(1)} &= 2.8303, & \sigma_2^{(1)} &= 2.5922 \end{aligned} \right. \quad (12)$$

## 证据因子

根据全概率公式，第  $i$  个数据点证据因子  $f_X(x^{(i)}, \theta)$  可以通过叠加联合概率得到：

$$\begin{aligned} \underbrace{f_X \left( x^{(i)}, \theta \right)}_{\text{Evidence}} &= \sum_{k=1}^K \underbrace{f_{X,Y} \left( x^{(i)}, C_k, \theta \right)}_{\text{Joint}} \\ &= \sum_{k=1}^K \underbrace{p_Y \left( C_k, \theta \right)}_{\text{Prior}} \underbrace{f_{X|Y} \left( x^{(i)} | C_k, \theta \right)}_{\text{Likelihood}} \end{aligned} \quad (13)$$

## 对数似然函数

构造**对数似然函数** (log likelihood function)  $L(\theta)$ ，如下：

$$L(\theta) = \ln \underbrace{\left[ \prod_{i=1}^n f_X(x^{(i)}, \theta) \right]}_{\text{Likelihood function}} = \sum_{i=1}^n \left[ \ln f_X(x^{(i)}, \theta) \right] \quad (14)$$

对数似然函数  $L(\theta)$  就是样本数据证据因子之积，再求对数。

取对数的叫做对数似然函数，而不做对数处理的叫做**似然函数** (likelihood function)。白话说，这里的“似然”指的是“可能性”。



对于似然函数陌生的同学可以参考《统计至简》第 16、20 章。

不管是似然函数，还是对数似然函数，反映的都是在特定参数  $\theta$  取值下，当前样本集合的可能性。

将 (13) 代入 (14) 可以得到：

$$L(\theta) = \sum_{i=1}^n \left\{ \ln \left[ \underbrace{\sum_{k=1}^K p_Y(C_k, \theta)}_{\text{Prior}} \underbrace{f_{X|Y}(x^{(i)} | C_k, \theta)}_{\text{Likelihood}} \right] \right\} \quad (15)$$

对于本例二聚类问题，对数似然函数值可以通过下式计算获得：

$$L(\theta^{(1)}) = \sum_{i=1}^n \left\{ \ln \left[ \underbrace{p_Y(C_1, \theta^{(1)})}_{\text{Prior}} \underbrace{f_{X|Y}(x^{(i)} | C_1, \theta^{(1)})}_{\text{Likelihood}} + \underbrace{p_Y(C_2, \theta^{(1)})}_{\text{Prior}} \underbrace{f_{X|Y}(x^{(i)} | C_2, \theta^{(1)})}_{\text{Likelihood}} \right] \right\} \quad (16)$$

代入 (12) 列出的本轮参数以及样本数据，得到  $L(\theta^{(1)}) = -1.9104$ 。

下面便是重复 E 步和 M 步，直到满足收敛条件。

## 22.4 迭代过程

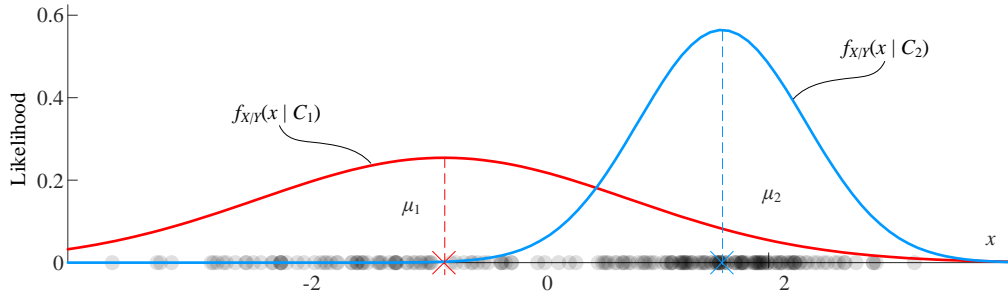
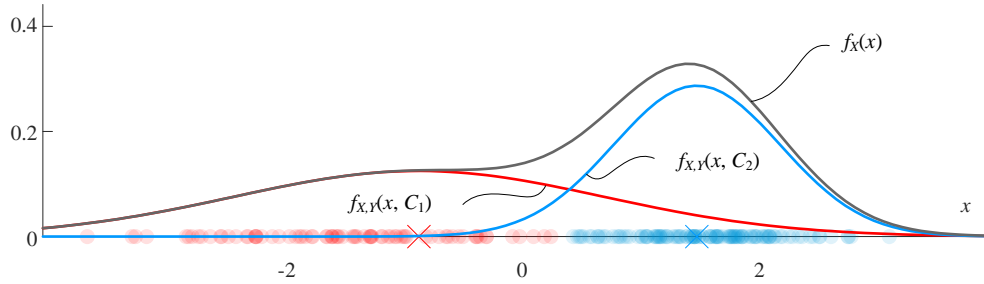
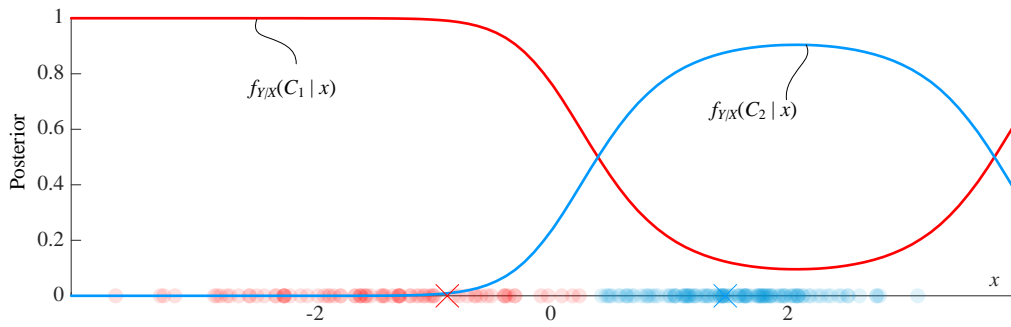
### 12 轮迭代

经过 12 轮迭代，参数  $\theta$  如下：

$$\begin{cases} \alpha_1^{(12)} = 0.49105, & \alpha_2^{(12)} = 0.50895 \\ \mu_1^{(12)} = -0.81597, & \mu_2^{(12)} = 1.5396 \\ \sigma_1^{(12)} = 2.4602, & \sigma_2^{(12)} = 0.49993 \end{cases} \quad (17)$$

图 5 到图 7 给出第 12 轮迭代结果。本轮对数似然函数值  $L(\theta^{(12)}) = -1.7344$ 。



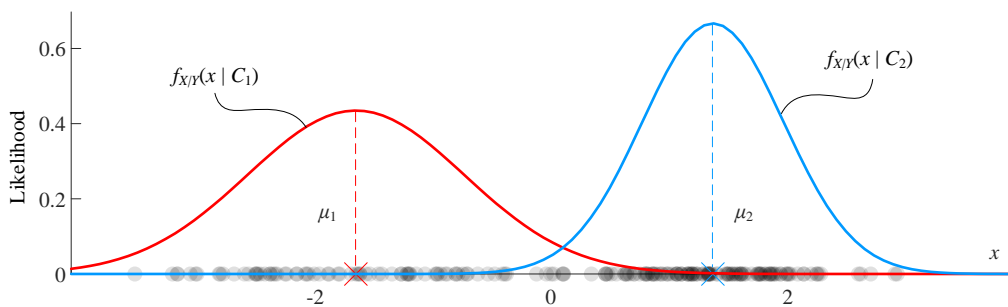
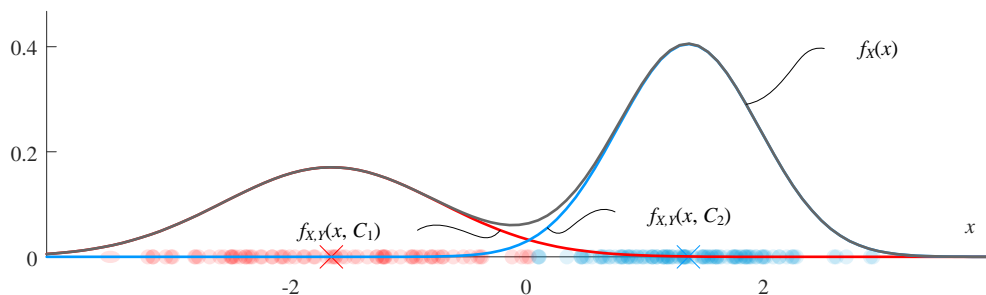
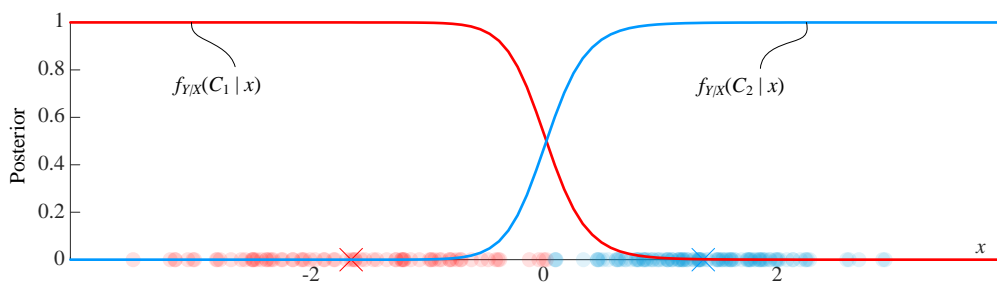
图 5. 经过 12 轮迭代参数对应的似然概率  $f_{X|Y}(x | C_1)$  和  $f_{X|Y}(x | C_2)$ 图 6. 经过 12 轮迭代参数对应的  $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$  和  $f_X(x)$ 图 7. 经过 12 轮迭代参数对应的后验概率  $f_{Y|X}(C_1 | x)$  和  $f_{Y|X}(C_2 | x)$ 

### 36 轮迭代

经过 36 轮迭代，得到的参数  $\theta$  如下：

$$\begin{cases} \alpha_1^{(36)} = 0.410, & \alpha_2^{(36)} = 0.590 \\ \mu_1^{(36)} = -1.325, & \mu_2^{(36)} = 1.493 \\ \sigma_1^{(36)} = 1.329, & \sigma_2^{(36)} = 0.364 \end{cases} \quad (18)$$

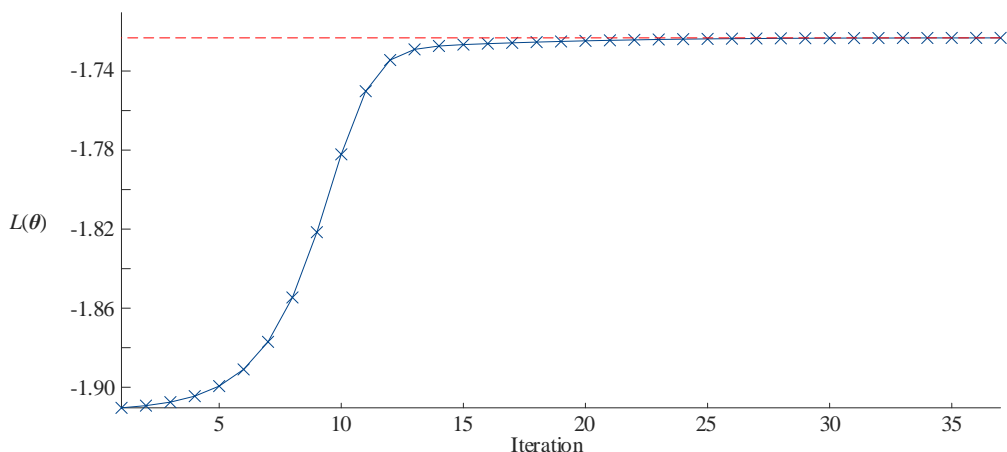
图 8 到图 10 所示为经过 36 轮迭代得到的结果。本轮对数似然函数值  $L(\theta^{(36)}) = -1.7232$ 。

图 8. 经过 36 轮迭代参数对应的似然概率  $f_{X|Y}(x | C_1)$  和  $f_{X|Y}(x | C_2)$ 图 9. 经过 36 轮迭代参数对应的  $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$  和  $f_X(x)$ 图 10. 经过 36 轮迭代参数对应的  $f_{Y|X}(C_1 | x)$  和  $f_{Y|X}(C_2 | x)$ 

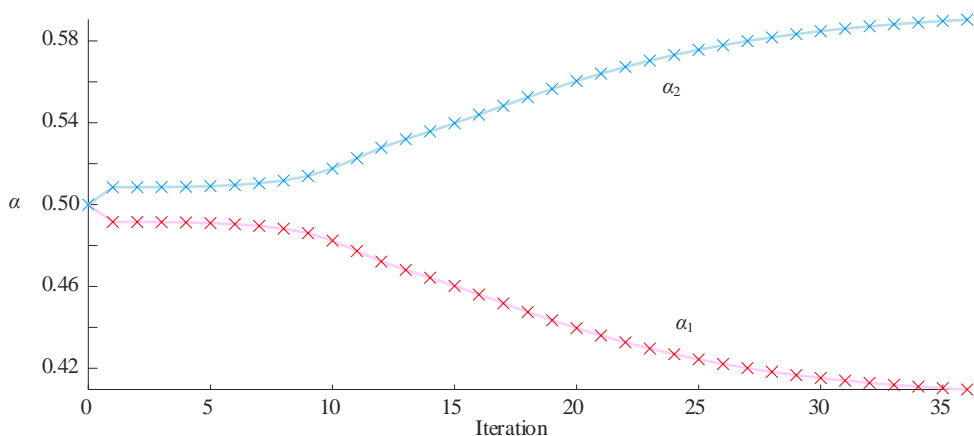
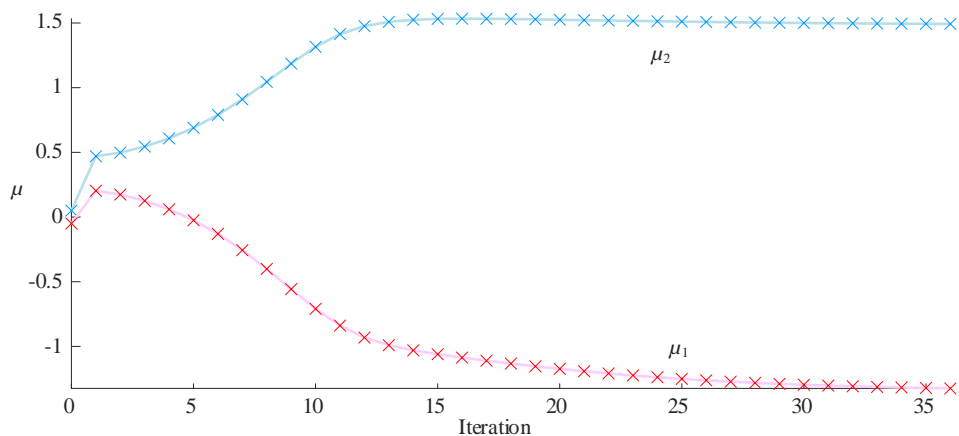
本例设置的迭代截止条件是，要么和上一轮相比对数似然函数  $L(\theta)$  值变化小于 0.00001，要么迭代次数超过 50 次；最先满足两者之一，则迭代停止。

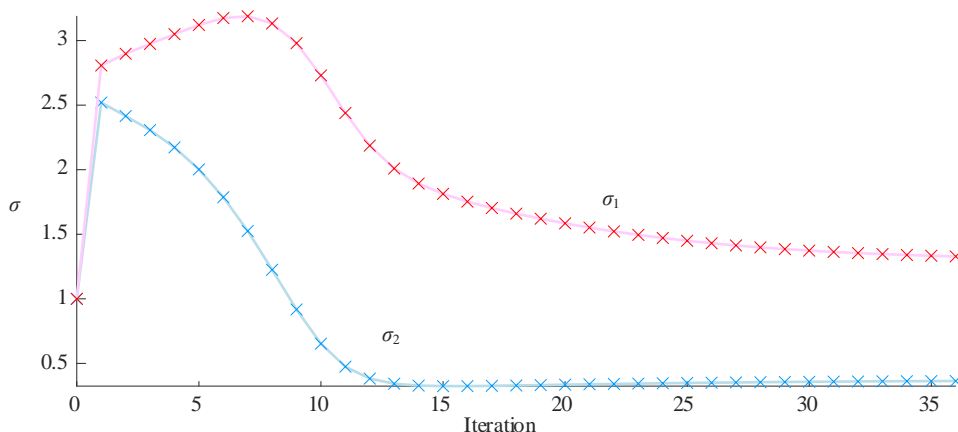
### 迭代收敛过程

图 11 所示为 36 次迭代，对数似然函数  $L(\theta)$  不断收敛过程。第 15 轮迭代之后，对数似然函数  $L(\theta)$  值便趋于稳定。

图 11. 经过 36 次迭代，对数似然函数  $L(\theta)$  不断收敛过程

本例是单特征、二聚类问题，因此  $\theta$  共有 6 个参数；在迭代过程中，这 6 个参数数值也在不断收敛。图 12 展示参数  $\alpha_1$  和  $\alpha_2$  不断收敛过程；图 13 所示为参数  $\mu_1$  和  $\mu_2$  不断收敛过程；图 14 为参数  $\mu_1$  和  $\mu_2$  不断收敛过程。

图 12. 经过 36 次迭代，参数  $\alpha_1$  和  $\alpha_2$  不断收敛过程图 13. 经过 36 次迭代，参数  $\mu_1$  和  $\mu_2$  不断收敛过程

图 14. 经过 36 次迭代，参数  $\sigma_1$  和  $\sigma_2$  不断收敛过程

EM 算法的迭代过程便是随着参数不断迭代更新，对数似然函数  $L(\theta)$  数值不断增大过程，直到满足收敛条件。EM 算法不仅仅是针对  $L(\theta)$  收敛过程，也是对于参数  $\theta$  的收敛过程。

## 22.5 多元 GMM 迭代

多元 EM 算法和本章前文介绍的一元 EM 算法思路完全一致。多元 EM 算法引入大量矩阵运算。本节以二元样本数据聚类为例逐步介绍多元 EM 算法。

图 15 所示为两特征样本数据分布及直方图。

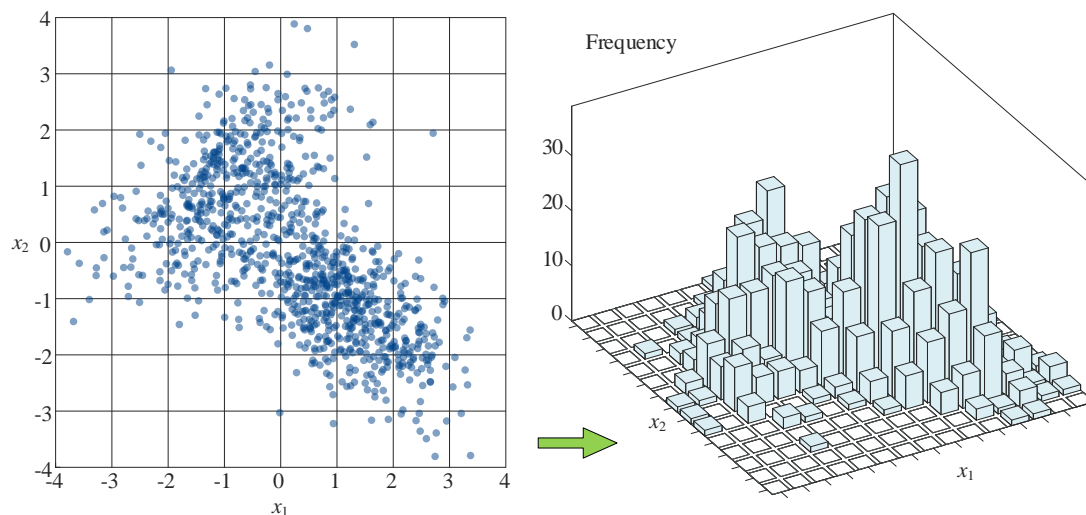


图 15. 两特征样本数据分布

### 初始化

首先初始化参数  $\theta$ :

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

$$\boldsymbol{\theta}^{(0)} = \{\alpha_1^{(0)}, \alpha_2^{(0)}, \boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \boldsymbol{\Sigma}_2^{(0)}\} \quad (19)$$

初始化参数  $\boldsymbol{\theta}$  具体数值如下：

$$\begin{cases} \alpha_1^{(0)} = P(C_1, \boldsymbol{\theta}^{(0)}) = 0.5, & \alpha_2^{(0)} = P(C_2, \boldsymbol{\theta}^{(0)}) = 0.5 \\ \boldsymbol{\mu}_1^{(0)} = [1 \ 0]^T, & \boldsymbol{\mu}_2^{(0)} = [-1 \ 0]^T \\ \boldsymbol{\Sigma}_1^{(0)} = \boldsymbol{\Sigma}_2^{(0)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{cases} \quad (20)$$

## 似然概率

假设  $f_{X|Y}(\mathbf{x} | C_1, \boldsymbol{\theta}^{(0)})$  和  $f_{X|Y}(\mathbf{x} | C_2, \boldsymbol{\theta}^{(0)})$  的概率密度函数 PDF 均为正态分布，具体如下：

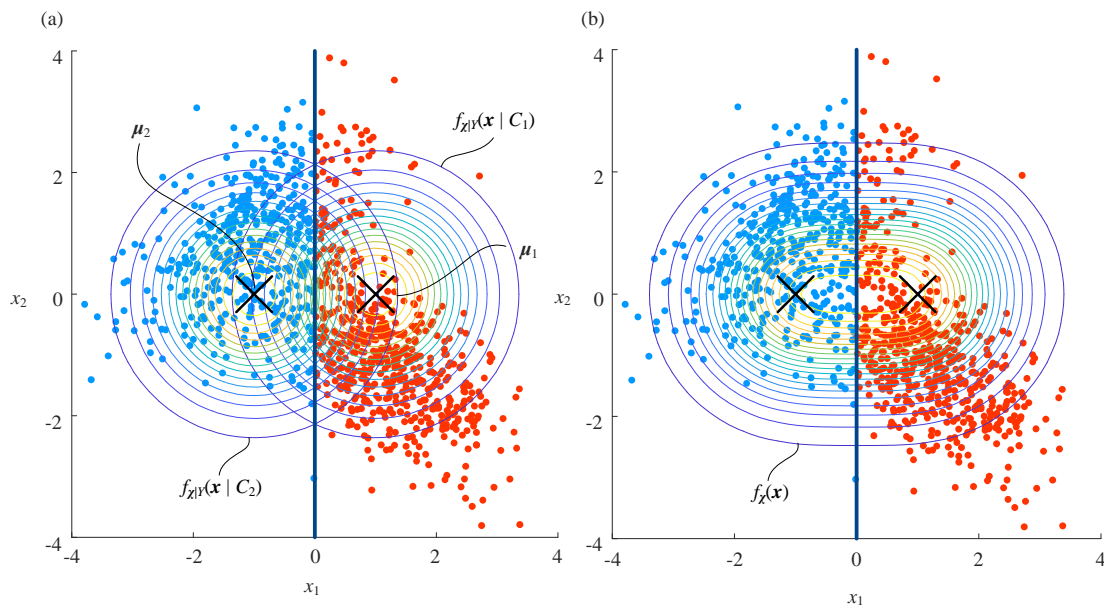
$$\begin{cases} f_{X|Y}(\mathbf{x} | C_1, \boldsymbol{\theta}^{(0)}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1^{(0)})^T (\boldsymbol{\Sigma}_1^{(0)})^{-1} (\mathbf{x} - \boldsymbol{\mu}_1^{(0)})\right)}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_1^{(0)}|}} \\ f_{X|Y}(\mathbf{x} | C_2, \boldsymbol{\theta}^{(0)}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2^{(0)})^T (\boldsymbol{\Sigma}_2^{(0)})^{-1} (\mathbf{x} - \boldsymbol{\mu}_2^{(0)})\right)}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_2^{(0)}|}} \end{cases} \quad (21)$$

## 证据因子

下一步，估算证据因子概率密度函数  $f_X(\mathbf{x}, \boldsymbol{\theta}^{(0)})$ ：

$$\begin{aligned} f_X(\mathbf{x}, \boldsymbol{\theta}^{(0)}) &= f_{X,Y}(\mathbf{x}, C_1, \boldsymbol{\theta}^{(0)}) + f_{X,Y}(\mathbf{x} \cap C_2, \boldsymbol{\theta}^{(0)}) \\ &= p_Y(C_1, \boldsymbol{\theta}^{(0)}) f_{X|Y}(\mathbf{x} | C_1, \boldsymbol{\theta}^{(0)}) + p_Y(C_2, \boldsymbol{\theta}^{(0)}) f_{X|Y}(\mathbf{x} | C_2, \boldsymbol{\theta}^{(0)}) \\ &= \frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1^{(0)})^T (\boldsymbol{\Sigma}_1^{(0)})^{-1} (\mathbf{x} - \boldsymbol{\mu}_1^{(0)})\right)}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_1^{(0)}|}} + \frac{1}{2} \times \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2^{(0)})^T (\boldsymbol{\Sigma}_2^{(0)})^{-1} (\mathbf{x} - \boldsymbol{\mu}_2^{(0)})\right)}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_2^{(0)}|}} \end{aligned} \quad (22)$$

图 16 (a) 展示初始化参数  $\boldsymbol{\theta}^{(0)}$  对应的  $f_{X|Y}(\mathbf{x} | C_1)$  和  $f_{X|Y}(\mathbf{x} | C_2)$  等高线；图 16 (b) 展示  $f_X(\mathbf{x})$  等高线图。

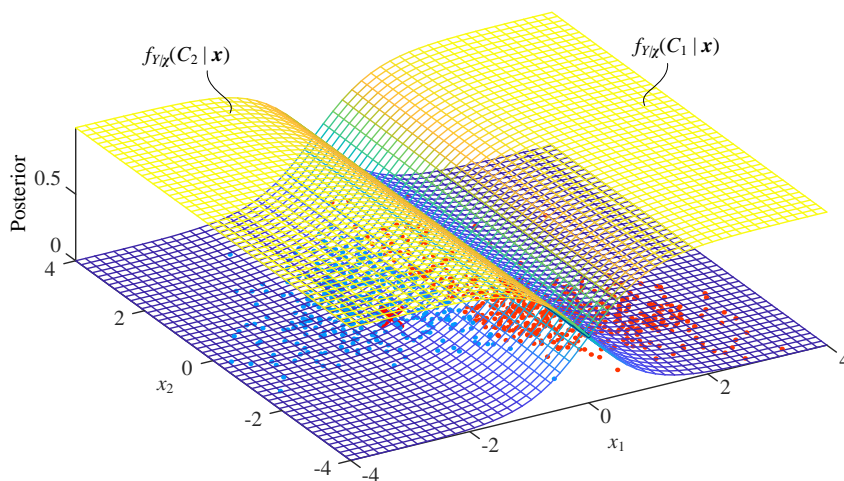
图 16. 初始化参数  $\theta^{(0)}$  对应的  $f_{Y|X}(x | C_1)$  和  $f_{Y|X}(x | C_2)$  等高线, 以及  $f_X(x)$  等高线图

## 后验概率

根据贝叶斯定理, 计算后验概率  $f_{Y|X}(C_1 | \mathbf{x}, \theta^{(0)})$  和  $f_{Y|X}(C_2 | \mathbf{x}, \theta^{(0)})$ :

$$\begin{cases} f_{Y|X}(C_1 | \mathbf{x}, \theta^{(0)}) = \frac{p_Y(C_1, \theta^{(0)}) f_{X|Y}(\mathbf{x} | C_1, \theta^{(0)})}{f_X(\mathbf{x}, \theta^{(0)})} \\ f_{Y|X}(C_2 | \mathbf{x}, \theta^{(0)}) = \frac{p_Y(C_2, \theta^{(0)}) f_{X|Y}(\mathbf{x} | C_2, \theta^{(0)})}{f_X(\mathbf{x}, \theta^{(0)})} \end{cases} \quad (23)$$

图 17 所示为初始化参数  $\theta^{(0)}$  计算得到后验概率曲面。

图 17. 初始化参数  $\theta^{(0)}$  计算得到  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  曲面

## 更新参数

下一步进行 EM 算法中 M 步, 更新参数。

更新参数  $\alpha_1$  和  $\alpha_2$ :

$$\begin{cases} \alpha_1^{(1)} = \frac{\sum_{i=1}^n f_{Y|X}(C_1 | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(0)})}{n} = 0.56019 \\ \alpha_2^{(1)} = \frac{\sum_{i=1}^n f_{Y|X}(C_2 | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(0)})}{n} = 0.43981 \end{cases} \quad (24)$$

更新簇质心  $\boldsymbol{\mu}_1$  和  $\boldsymbol{\mu}_2$ :

$$\begin{cases} \boldsymbol{\mu}_1^{(1)} = \frac{\sum_{i=1}^n \{f_{Y|X}(C_1 | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(0)}) \mathbf{x}^{(i)}\}}{n\alpha_1^{(1)}} = \begin{bmatrix} 1.098 \\ -0.764 \end{bmatrix} \\ \boldsymbol{\mu}_2^{(1)} = \frac{\sum_{i=1}^n \{f_{Y|X}(C_2 | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(0)}) \mathbf{x}^{(i)}\}}{n\alpha_2^{(1)}} = \begin{bmatrix} -0.8924 \\ 0.4627 \end{bmatrix} \end{cases} \quad (25)$$

为了方便运算默认  $\mathbf{x}^{(i)}$  为列向量。

更新簇协方差矩阵  $\boldsymbol{\Sigma}_1$  和  $\boldsymbol{\Sigma}_2$ :

$$\begin{cases} \boldsymbol{\Sigma}_1^{(1)} = \frac{\sum_{i=1}^n \{f_{Y|X}(C_1 | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(0)}) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_1) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_1)^T\}}{n\alpha_1^{(1)}} = \begin{bmatrix} 0.9346 & -0.7809 \\ -0.7809 & 1.787 \end{bmatrix} \\ \boldsymbol{\Sigma}_2^{(1)} = \frac{\sum_{i=1}^n \{f_{Y|X}(C_2 | \mathbf{x}^{(i)}, \boldsymbol{\theta}^{(0)}) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_2) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_2)^T\}}{n\alpha_2^{(1)}} = \begin{bmatrix} 1.034 & -0.1588 \\ -0.1588 & 1.213 \end{bmatrix} \end{cases} \quad (26)$$

这样，我们便得到了一组全新的参数  $\boldsymbol{\theta}^{(1)}$ 。

## 对数似然函数

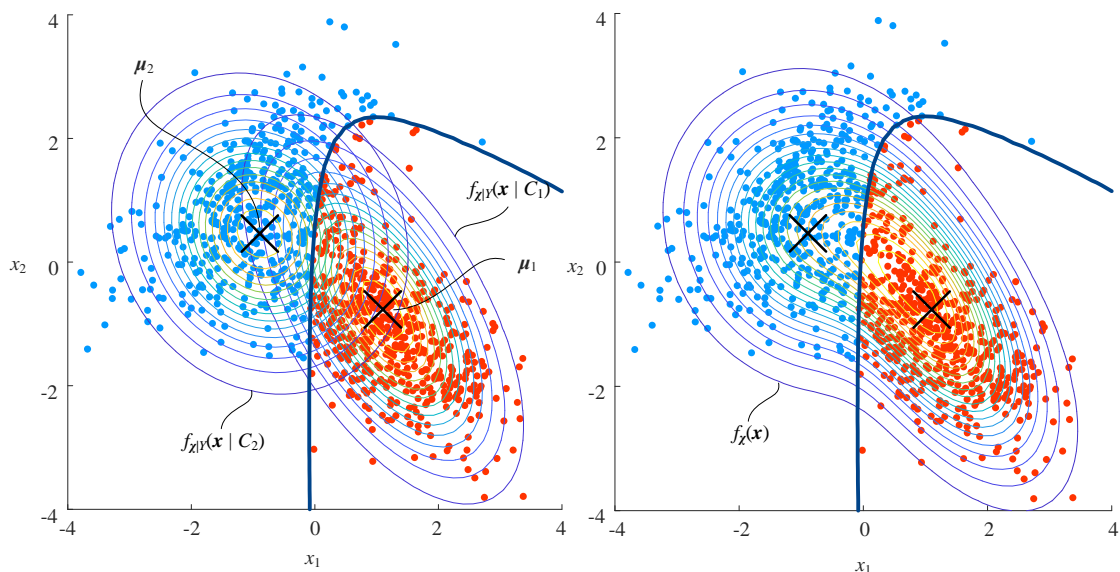
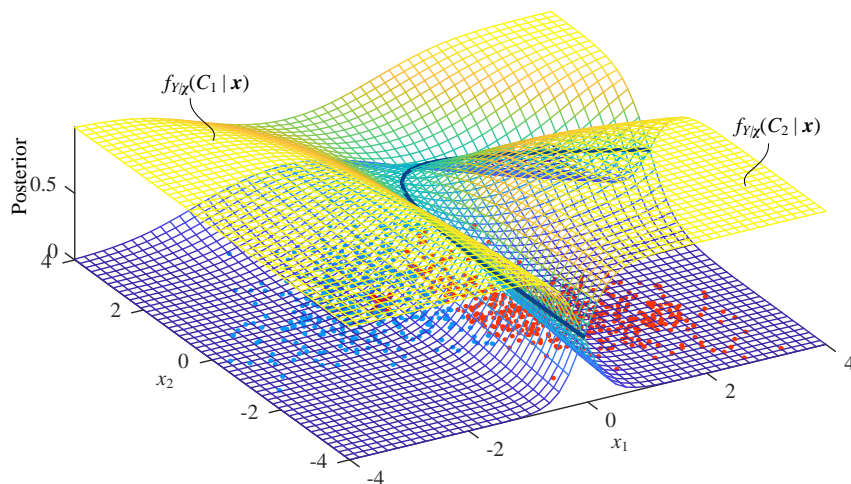
构造对数似然函数  $L(\boldsymbol{\theta})$ ，如下：

$$L(\boldsymbol{\theta}^{(1)}) = \ln \left( \prod_{i=1}^n f_X(\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(1)}) \right) \quad (27)$$

代入 (24)、(25) 和 (26) 中更新得到的参数，计算得到对数似然值  $L(\boldsymbol{\theta}^{(1)}) = -3.213045$ 。

图 18 和图 19 展示  $\boldsymbol{\theta}^{(1)}$  参数对应的概率曲面。分别比较图 16 和图 17，可以发现图 18 和图 19 所示聚类决策边界已经发生显著变化。



图 18. 参数  $\theta^{(1)}$  对应的  $f_{Z|Y}(x | C_1)$  和  $f_{Z|Y}(x | C_2)$  等高线, 以及  $f_Z(x)$  等高线图图 19. 参数  $\theta^{(1)}$  计算得到  $f_{Y|Z}(C_1 | x)$  和  $f_{Y|Z}(C_2 | x)$  曲面

## 第二轮迭代

进入第 2 轮迭代, 更新参数  $\theta^{(2)}$ :

$$\begin{cases} \alpha_1^{(2)} = P(C_1, \theta^{(2)}) = 0.56481, & \alpha_2^{(2)} = P(C_2, \theta^{(2)}) = 0.43519 \\ \mu_1^{(2)} = [1.097 & -0.84]^T, & \mu_2^{(2)} = [-0.9121 & 0.5744]^T \\ \Sigma_1^{(2)} = \begin{bmatrix} 0.9179 & -0.7818 \\ -0.7818 & 1.614 \end{bmatrix}, & \Sigma_2^{(2)} = \begin{bmatrix} 1.02 & 0.07167 \\ 0.07167 & 1.153 \end{bmatrix} \end{cases} \quad (28)$$

## 第 11 轮迭代

经过 11 轮迭代, 满足优化结束条件, 并获得更新参数  $\theta^{(11)}$ :



$$\begin{cases} \alpha_1^{(11)} = P(C_1, \theta^{(11)}) = 0.57516, & \alpha_2^{(11)} = P(C_2, \theta^{(11)}) = 0.42484 \\ \mu_1^{(11)} = [1.096 \quad -1.114]^T, & \mu_2^{(11)} = [-0.9589 \quad 0.9795]^T \\ \Sigma_1^{(11)} = \begin{bmatrix} 0.8938 & -0.4735 \\ -0.4735 & 0.7659 \end{bmatrix}, & \Sigma_2^{(11)} = \begin{bmatrix} 0.9627 & 0.5045 \\ 0.5045 & 0.9269 \end{bmatrix} \end{cases} \quad (29)$$

图 20 和图 21 展示完成迭代后曲面等高线结果。

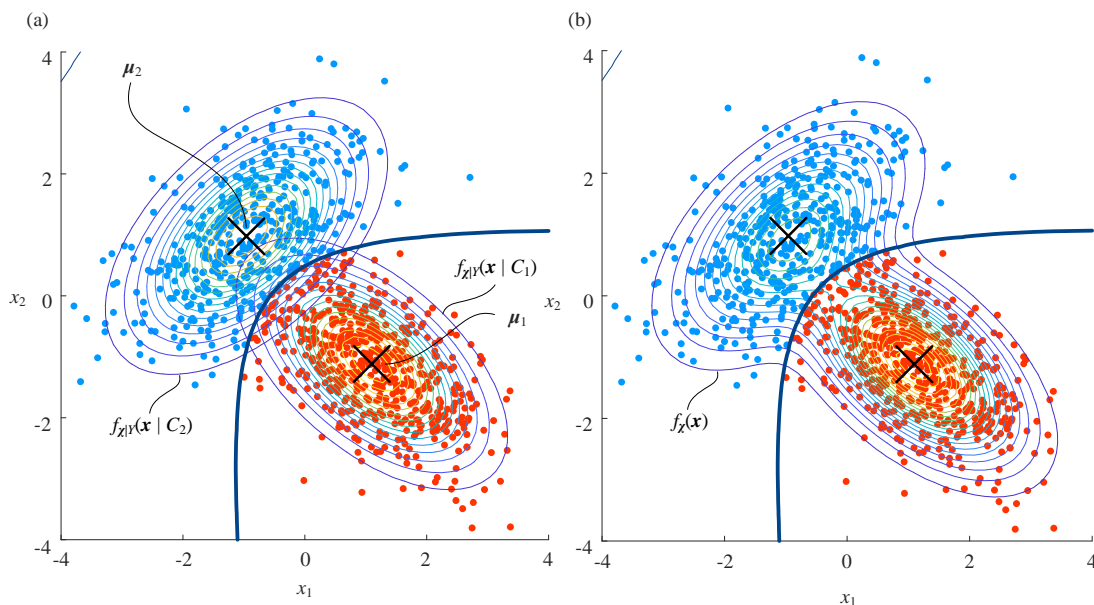


图 20. 参数  $\theta^{(11)}$  对应的  $f_{Y|Z}(x|C_1)$  和  $f_{Y|Z}(x|C_2)$  等高线，以及  $f_Z(x)$  等高线图

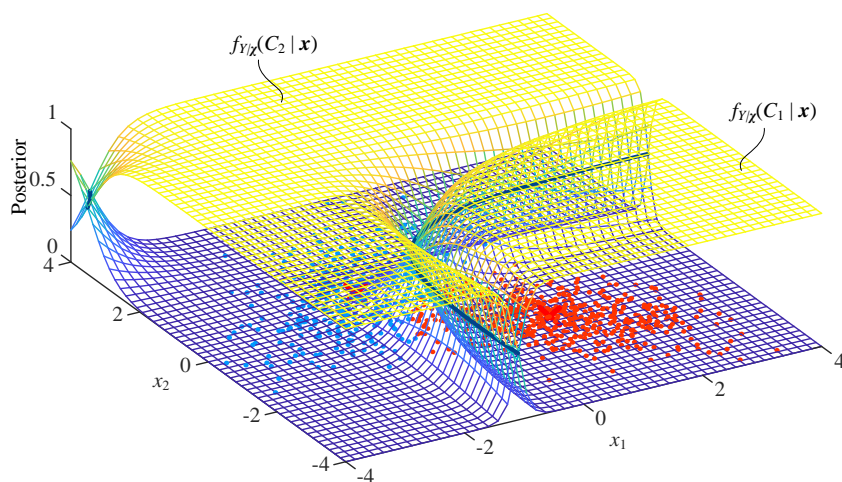


图 21. 参数  $\theta^{(11)}$  计算得到  $f_{Y|Z}(C_1|x)$  和  $f_{Y|Z}(C_2|x)$  曲面

## 迭代收敛过程

图 22 展示的是经过 11 次迭代  $L(\theta)$  递增收敛过程。相信大家看过图 11 和图 22 这两幅图，便明白为什么对数似然函数  $L(\theta)$  是参数  $\theta$  的函数了。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

参数  $\theta$  相当于未知数，由于不存在解析解，只能通过迭代优化求解参数  $\theta$ 。整个过程就是找到描述样本数据集最佳参数  $\theta$ 。

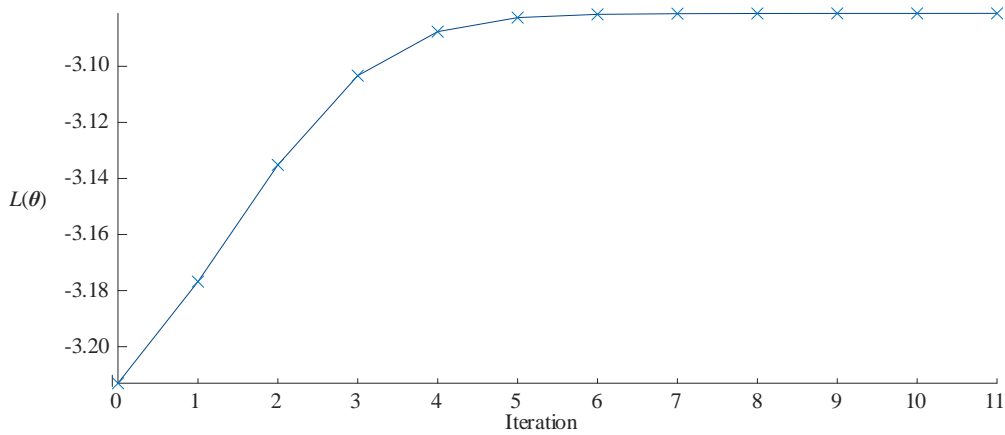


图 22. 经过 11 次迭代，似然函数  $L(\theta)$  不断收敛过程

图 23 所示为经过 11 次迭代，参数  $\alpha_1$  和  $\alpha_2$  不断收敛过程。

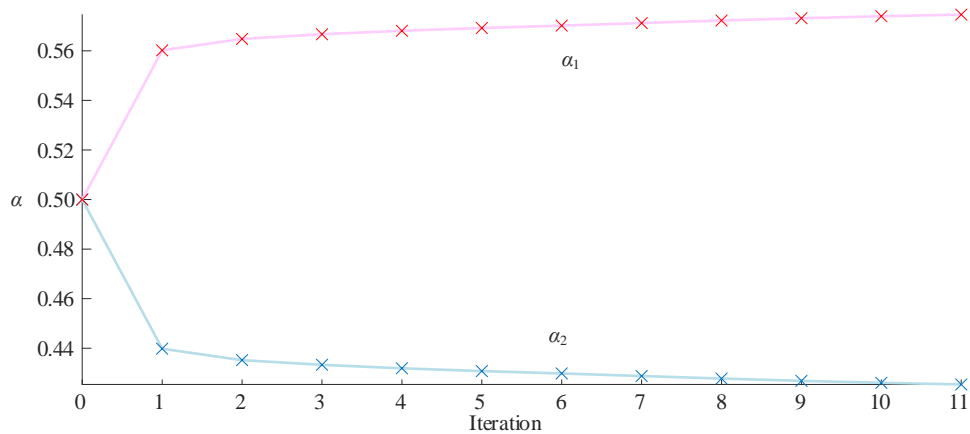


图 23. 11 次迭代，参数  $\alpha_1$  和  $\alpha_2$  不断收敛过程

为了更好地可视化二元高斯分布参数——质心和协方差——变化过程，我们利用椭圆来表达协方差，而椭圆中心所在位置便是簇质心。

图 24 很好地展示 11 次迭代，两个二元高斯分布质心和协方差不断变化过程。图 25 则展示决策边界随着迭代不断变化。

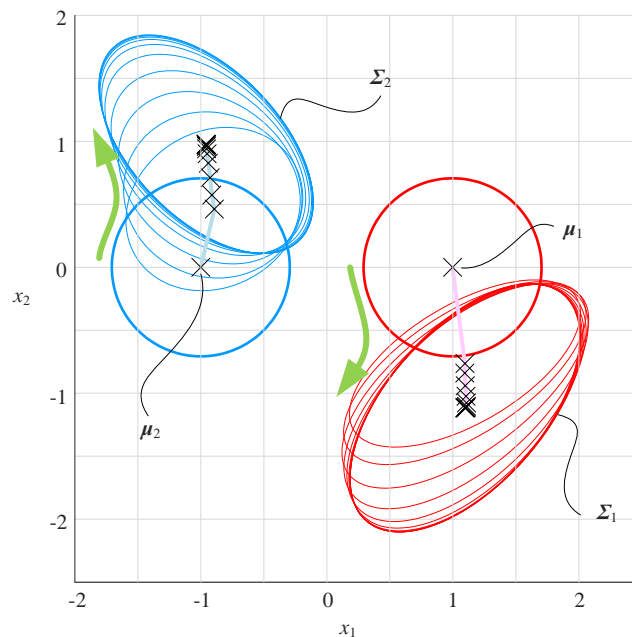


图 24. 11 次迭代，二元高斯分布质心和协方差不断变化过程

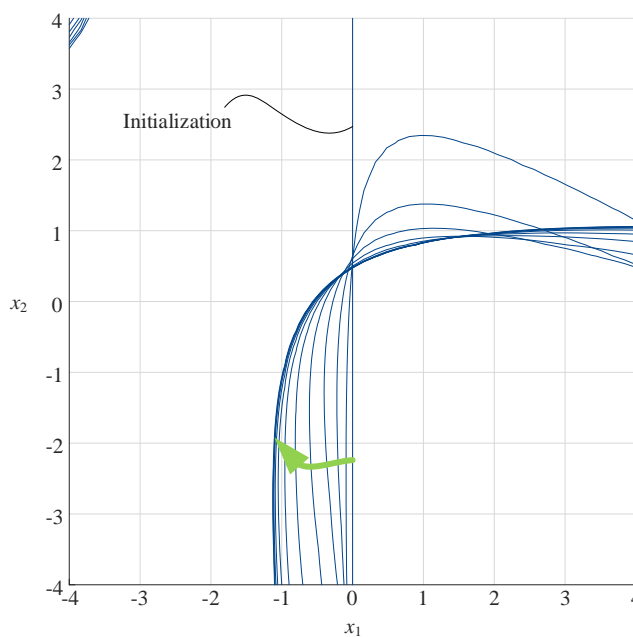
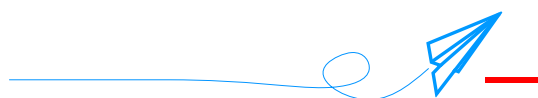


图 25. 11 次迭代，决策边界不断变化过程

EM 算法很有可能迭代收敛在局部极大值处，而非全局最大值；常用的解决办法是，选取不同初始值进行迭代优化；比较对数似然函数  $L(\theta)$  收敛值，从不同优化解中选取理想解。



EM 算法是一种迭代算法，用于在不完全观测的情况下，通过已知的观测数据来估计模型参数。其核心思想是通过不断迭代，利用已知数据计算未知参数的最大似然估计。EM 算法的迭代包括两个步骤：E 步骤和 M 步骤，其中 E 步骤计算隐变量的后验概率，M 步骤利用后验概率重新估计参数。EM 算法通常用于处理混合模型、隐马尔可夫模型等问题，具有广泛的应用，如聚类、密度估计、图像处理等领域。