Statistical Computing
Chap. 2: Expectation–maximization Algorithm

LIU, Ran

March 8, 2025

目录

介绍

统计学中的频率学派方法通常通过极大化观测数据的对数似然函数 $\log L(\boldsymbol{\theta}|\boldsymbol{x})$ 来估计参数 $\boldsymbol{\theta}$。然而，在许多情况下，观测数据并不完整，部分数据缺失，使得直接极大化 $\log L(\boldsymbol{\theta}|\boldsymbol{x})$ 变得困难。

EM 算法（Expectation-Maximization Algorithm）是一种用于处理缺失数据情况下的参数估计方法。其核心思想来源于：

- 通过引入缺失变量 $\boldsymbol{Z}$，将复杂的优化问题转换为较容易求解的形式。
- 迭代地进行期望计算与最大化步骤，以逼近最大似然估计。

EM 算法的理论基础及收敛性分析由 Dempster、Laird 和 Rubin 在 1977 年的论文 *"Maximum Likelihood from Incomplete Data via the EM Algorithm"* 中系统阐述。

基本思想

EM 算法的目标是：

1. 极大化观测数据的对数似然函数：

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \int p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})\mathrm{d}z$$

2. 由于缺失数据 Z 的存在，直接优化 $L(\boldsymbol{\theta}|\boldsymbol{x})$ 变得困难。

为此，EM 算法采用以下策略：

- **E 步（Expectation）**：计算当前参数估计值下，缺失数据的期望，从而构造完整数据的对数似然函数。
- **M 步（Maximization）**：基于 E 步的结果，极大化完整数据的期望对数似然函数，更新参数估计。

EM 算法通过不断迭代 E 步和 M 步，直至参数 θ 收敛，最终获得最优估计值。

目录

数学表达

定义:

- 观测数据（Observed Data）: $\boldsymbol{X}$
- 缺失数据 / 隐变量（Latent Variables）: $\boldsymbol{Z}$
- 完整数据（Complete Data）: $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$
- 观测数据的密度函数: $p_{\boldsymbol{X}}(x \mid \theta)$
- 观测数据的条件密度: $p_{\boldsymbol{X}|\boldsymbol{Z}}(x \mid z, \theta)$
- 完整数据的密度函数: $p_{\boldsymbol{Y}}(y \mid \theta)$
- 缺失数据的条件密度: $p_{\boldsymbol{Z}|\boldsymbol{X}}(z \mid x, \theta)$

EM 算法简介

我们考虑这样一个问题：完整数据 $\boldsymbol{Y}$ 不能被完全观测，而是由观测数据 $\boldsymbol{X}$ 和缺失数据 $\boldsymbol{Z}$ 组成，即：

$$\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$$

完整数据的 **似然函数**可表示为：

$$p(\boldsymbol{Y} \mid \theta) = p(\boldsymbol{X}, \boldsymbol{Z} \mid \theta) = p(\boldsymbol{X} \mid \theta) \cdot p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta)$$

拆分出观测数据的对数似然函数：

$$\log p(\boldsymbol{X} \mid \theta) = \log p(\boldsymbol{Y} \mid \theta) - \log p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta)$$

由于 $\boldsymbol{Z}$ 是未知的，我们无法直接优化 $\log p(\boldsymbol{X} \mid \theta)$。

E 步：计算期望

我们对上式两边取期望：

$$\mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{X} \mid \theta) \mid \boldsymbol{X}, \theta^{(t)}\right]$$

$$= \mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{Y} \mid \theta) \mid \boldsymbol{X}, \theta^{(t)}\right] - \mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta) \mid \boldsymbol{X}, \theta^{(t)}\right]$$

定义：

- **Q 函数（期望对数似然函数）：**

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{Y} \mid \theta) \mid \boldsymbol{X}, \theta^{(t)}\right]$$

- **H 函数（熵相关项）：**

$$H(\theta \mid \theta^{(t)}) = \mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta) \mid \boldsymbol{X}, \theta^{(t)}\right]$$

M 步：最大化 Q 函数

根据定义：

$$\log p(\boldsymbol{X} \mid \theta) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)})$$

其中 Z 已被积分掉。在之后我们会证明：

$$H(\theta^{(t)} \mid \theta^{(t)}) \geqslant H(\theta \mid \theta^{(t)})$$

因此，算法只需保证：

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geqslant Q(\theta^{(t)} \mid \theta^{(t)})$$

即可确保：

$$\ell(\theta^{(t+1)} \mid \boldsymbol{X}) \geqslant \ell(\theta^{(t)} \mid \boldsymbol{X})$$

EM 算法

EM 从 $\boldsymbol{\theta}^0$ 开始，然后在两步之间交替：E 表示期望，M 表示最大化. 该算法概括如下

1. 写出完全数据的似然函数.
2. **E 步**：计算条件期望 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_z\{\log p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\}$.
3. **M 步**：寻求 $\boldsymbol{\theta}$ 来最大化 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. 设 $\boldsymbol{\theta}^{(t+1)}$ 为找到的点.
4. 返回 E 步，直到满足某停止规则为止.

一般需要先设置

- 缺失数据的边际分布 $p(z|\theta)$,
- 观测数据基于缺失数据的条件分布 $p(x|z,\theta)$,

才能得到完整数据的似然函数，也即完整数据的联合概率，然后计算 Q 函数。(当然若能直接设置完整数据的概率密度也行。)

$$L(\theta|y) = p(y|\theta) = p(x,z|\theta) = p(x|z,\theta)p(z|\theta)$$

EM 算法的停止准则

在 EM 算法中，停止条件可以采用以下两种方式：

- 固定迭代次数：设置最大迭代次数 $T_{\max}$，达到后停止。
- 判断参数或目标函数的变化量：当参数或目标函数的变化足够小，则认为算法收敛：
  - 参数变化：

$$\|\theta^{(t+1)} - \theta^{(t)}\|^2 = (\theta^{(t+1)} - \theta^{(t)})^{\mathrm{T}}(\theta^{(t+1)} - \theta^{(t)}) < \epsilon$$

  - 目标函数的变化：

$$\left|Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})\right| < \epsilon$$

为了直观观察 EM 算法的收敛性，我们通常绘制 Q 函数的轨迹图 (Trace Plot)，即 Q 值随迭代次数的变化图，以判断收敛趋势。

例子：混合高斯

假设我们从同一年级的学生总体中采样身高，男生的比例为 p，在未知性别的情况下，身高服从混合高斯模型，它的似然函数是：

$$
\begin{aligned}
& L(\theta \mid X_1, \cdots, X_n) \\
=& L(\mu_1, \mu_2, \sigma_1, \sigma_2, p \mid X_1, \cdots, X_n) \\
=& \prod_{i=1}^{n} \left\{ p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) + (1-p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right\}.
\end{aligned}
$$

如果我们假设男女性别为潜变量 ($Z_i = 0$ 为女性，$Z_i = 1$ 为男性)，那么我们有 $X_i \mid Z_i = 1 \sim N(\mu_1, \sigma_1^2)$ 和 $X_i \mid Z_i = 0 \sim N(\mu_2, \sigma_2^2)$.

所以 full data 的 density 是

$$\begin{aligned}
&\mathbb{P}\left\{X_i, Z_i \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p\right\} \\
=&\mathbb{P}\left\{Z_i \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p\right\} \mathbb{P}\left\{X_i \mid Z_i, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p\right\} \\
=&\left[p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right)\right]^{Z_i} \left[(1-p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}\right)\right]^{1-Z_i}.
\end{aligned}$$

(黑板推导)

E Step

根据定义得到 $\mathcal{Q}$ 函数为

$$
\begin{aligned}
& E\left\{\ell(\theta \mid X, Z) \mid X, \theta^{(t)}\right\} \\
&= \sum_{i=1}^{n} \mathbb{E}\left\{Z_i \mid X_i; \theta^{(t)}\right\}\left[\ln p - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_1^2 - \frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right] \\
&+ \sum_{i=1}^{n} \mathbb{E}\left\{1 - Z_i \mid X_i; \theta^{(t)}\right\}\left[\ln(1-p) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_2^2 - \frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right],
\end{aligned}
$$

E Step

其中推导条件期望的时候，注意我们已有的条件，再进行拆分

$$
\begin{aligned}
&\mathbb{E}\left\{Z_i \mid X_i;\theta^{(t)}\right\} = 0\cdot\mathbb{P}\left\{Z_i=0 \mid X_i;\theta^{(t)}\right\} + 1\cdot\mathbb{P}\left\{Z_i=1 \mid X_i;\theta^{(t)}\right\}\\
&= \mathbb{P}\left\{Z_i=1 \mid X_i;\theta^{(t)}\right\}\\
&= \frac{\mathbb{P}\left\{Z_i=1, X_i;\theta^{(t)}\right\}}{\mathbb{P}\left\{Z_i=1, X_i;\theta^{(t)}\right\} + \mathbb{P}\left\{Z_i=0, X_i;\theta^{(t)}\right\}}\\
&= \frac{\mathbb{P}\left\{X_i \mid Z_i=1;\theta^{(t)}\right\}\mathbb{P}\left\{Z_i=1;\theta^{(t)}\right\}}{\mathbb{P}\left\{X_i \mid Z_i=1;\theta^{(t)}\right\}\mathbb{P}\left\{Z_i=1;\theta^{(t)}\right\} + \mathbb{P}\left\{X_i \mid Z_i=0;\theta^{(t)}\right\}\mathbb{P}\left\{Z_i=0;\theta^{(t)}\right\}}\\
&= \frac{p^{(t)}\frac{1}{\sqrt{2\pi}\sigma_1^{(t)}}\exp\left(-\frac{(x_i-\mu_1^{(t)})^2}{2(\sigma_1^{(t)})^2}\right)}{p^{(t)}\frac{1}{\sqrt{2\pi}\sigma_1^{(t)}}\exp\left(-\frac{(x_i-\mu_1^{(t)})^2}{2(\sigma_1^{(t)})^2}\right) + (1-p^{(t)})\frac{1}{\sqrt{2\pi}\sigma_2^{(t)}}\exp\left(-\frac{(x_i-\mu_2^{(t)})^2}{2(\sigma_2^{(t)})^2}\right)}\\
&:= \xi_i^{(t)}
\end{aligned}
$$

M step

将 $\mathbb{E}\left\{Z_i \mid X_i; \theta^{(t)}\right\}$ 代入原式 $\mathcal{Q}(\theta|\theta^{(t)})$，则有

$$\begin{aligned}\mathcal{Q} = &\sum_{i=1}^{n} \xi_i^{(t)} \left[\ln p - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_1^2 - \frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right] \\ &+ \sum_{i=1}^{n}(1-\xi_i^{(t)}) \left[\ln(1-p) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_2^2 - \frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right],\end{aligned}$$

我们令 $\mathcal{Q}(\theta|\theta^{(t)})$ 的一阶导数等于 0：

$$\frac{\partial Q}{\partial p} = \sum_{i=1}^{n}\left(\frac{\xi_i^{(t)}}{p} - \frac{1-\xi_i^{(t)}}{1-p}\right) = 0 \Rightarrow \widehat{p} = \frac{\sum_{i=1}^{n}\xi_i^{(t)}}{n},$$

$$\frac{\partial Q}{\partial \mu_1} = \sum_{i=1}^{n}\xi_i^{(t)}\frac{-2(x_i-\mu_1)}{2\sigma_1^2} = 0 \Rightarrow \widehat{\mu}_1 = \frac{\sum_{i=1}^{n}\xi_i^{(t)}X_i}{\sum_{i=1}^{n}\xi_i^{(t)}},$$

$$\frac{\partial Q}{\partial \sigma_1^2} = \sum_{i=1}^{n}\xi_i^{(t)}\left[-\frac{1}{2\sigma_1^2} + \frac{(x_i-\mu_1)^2}{2(\sigma_1^2)^2}\right] = 0 \Rightarrow \widehat{\sigma}_1^2 = \frac{\sum_{i=1}^{n}\xi_i^{(t)}(X_i-\widehat{\mu}_1)^2}{\sum_{i=1}^{n}\xi_i^{(t)}}.$$

直观来看，p 的估计值为男生的比例，μ_1 的估计值为男生身高的平均值，可通过这种方法检验我们的公式。

当是多个类别的高斯混合模型，如何求解？

一般混合模型

如果一个数据集是由 K 个不同总体组成，密度函数具有如下混合密度形式

$$p(x;\theta)=\sum_{k=1}^{K}\pi_k p_k(x;\theta_k),$$

其中我们有 $\sum_{k=1}^{K}\pi_k=1$，以及模型参数包括 $(\mathbf{\Pi},\mathbf{\Theta})$.

假定观测到的数据为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，未观测到的数据为 $\boldsymbol{z}=\{z_1,z_2,\cdots,z_n\}$，其中 $z_i\in\{1,\cdots,K\}$ 代表数据的来源，也就是：

$$p(x_i\mid z_i=k;\mathbf{\Theta})=p_k(x;\theta_k),\quad p(z_i=k)=\pi_k.$$

(黑板推导)

则完全数据的似然函数可写成

$$L(\mathbf{\Pi}, \mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k p_k(x_i; \theta_k))^{I(z_i=k)}$$

则对数似然函数为

$$\ell(\mathbf{\Pi}, \mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(z_i = k)[\log \pi_k + \log(p_k(x_i; \theta_k))]$$

接下来再求 Q 函数

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_Z \left\{ \ell(\mathbf{\Pi}, \mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X}, \mathbf{\Pi}^{(t)}, \mathbf{\Theta}^{(t)} \right\} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} E(I(z_i = k) \mid \boldsymbol{X}, \mathbf{\Pi}^{(t)}, \mathbf{\Theta}^{(t)})[\log \pi_k + \log(p_k(x_i; \theta_k))] \end{aligned}$$

其中针对条件期望，我们有

$$\begin{aligned} E(I(z_i = k) \mid \boldsymbol{X}, \boldsymbol{\Pi}^{(t)}, \boldsymbol{\Theta}^{(t)}) &= p(z_i = k \mid \boldsymbol{X}, \boldsymbol{\Pi}^{(t)}, \boldsymbol{\Theta}^{(t)}) \\ &= \frac{p(x_i, z_i = k|\theta_k^{(t)})}{\sum_{l=1}^K p(x_i, z_i = l|\theta_l^{(t)})} \\ &= \frac{\pi_k^{(t)} p_k(x_i|\theta_k^{(t)})}{\sum_{l=1}^K \pi_l^{(t)} p_l(x_i|\theta_l^{(t)})} \end{aligned}$$

代入到 Q 函数里面后，求导得到更新公式。

目录

Jensen's inequality

琴生不等式，它给出积分的凸函数值和凸函数的积分值间的关系，在此不等式最简单形式中，阐明了对一平均做凸函数变换，会小于等于先做凸函数变换再平均。

若将琴生不等式应用在二点上，就回到了凸函数的基本性质：过一个凸函数上任意两点所作割线一定在这两点间的函数图象的上方，即：

$$tf(x_1) + (1-t)f(x_2) \geq f\left(tx_1 + (1-t)x_2\right), 0 \leq t \leq 1.$$

如果 X 是随机变量且 ϕ 是凸函数，则

$$\mathrm{E}[\varphi(X)] \geq \varphi(\mathrm{E}[X]).$$

设 $g(x)$ 是实值可测函数，$f(x)$ 是概率密度函数，更通用的形式为：

$$\int_{-\infty}^{\infty} \varphi(g(x))f(x)dx \geq \varphi\left(\int_{-\infty}^{\infty} g(x)f(x)dx\right).$$

收敛性

为了观察 EM 算法的收敛性质，我们通过说明每个最大化步提高了观测数据的对数似然 $\ell(\theta|x)$ 开始. 首先注意到观测数据密度的对数可重新表达为

$$\ell(\theta|x) = \log p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = \log p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) - \log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})$$

因此，

$$E\{\log p_{\boldsymbol{X}}(x|\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\} = E\{\log p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\} - E\{\log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\},$$

其中期望是关于 $Z|(\boldsymbol{x},\boldsymbol{\theta}^{(t)})$ 求取的. 于是

$$\ell(\theta|x) = \log p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}),$$

其中

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\{\log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\}.$$

我们将证明 $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 是不断减小的.

$$
\begin{aligned}
H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathrm{E}\{\log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta}^{(t)}) - \log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\} \\
&= \int -\log\left[\frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{(t)})}\right] p_{\boldsymbol{Z}|\boldsymbol{X}}(z|\boldsymbol{x},\boldsymbol{\theta}^{(t)})\mathrm{d}z \\
&\geqslant -\log \int p_{\boldsymbol{Z}|X}(z|x,\theta)\mathrm{d}z \\
&= 0.
\end{aligned}
$$

其中的放缩是 Jensen 不等式的应用，因为 $-\log u$ 关于 u 是单调递减，严格凸的. 所以我们的每一步 $\mathcal{Q}$ 增大而 H 减小，

$$
\log p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}^{(t+1)}) - \log p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}^{(t)}) \geqslant 0.
$$

则观测数据的对数似然不断上升.

(黑板推导)

在每次迭代中选择 $\theta^{(t+1)}$ 来最大化 $\mathcal{Q}(\theta|\theta^{(t)})$ 构成了标准的 EM 算法. 如果取而代之的是只简单选取任一个使得 $\mathcal{Q}(\theta^{(t+1)}|\theta^{(t)}) > \mathcal{Q}(\theta^{(t)}|\theta^{(t)})$ 的 $\theta^{(t+1)}$，那么得到的算法称作广义 EM, 或者 GEM.

值得注意的是，若 $-\ell(\hat{\theta}|x)$ 是正定的，则 EM 算法有线性收敛，且收敛速度跟缺失信息比例有关。当缺失信息比例比较大，收敛会非常慢.

为进一步理解 EM 如何工作, 注意到

$$H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geqslant 0$$

$$\Big(Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})\Big) - \Big(Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}|\boldsymbol{x})\Big) \geqslant 0$$

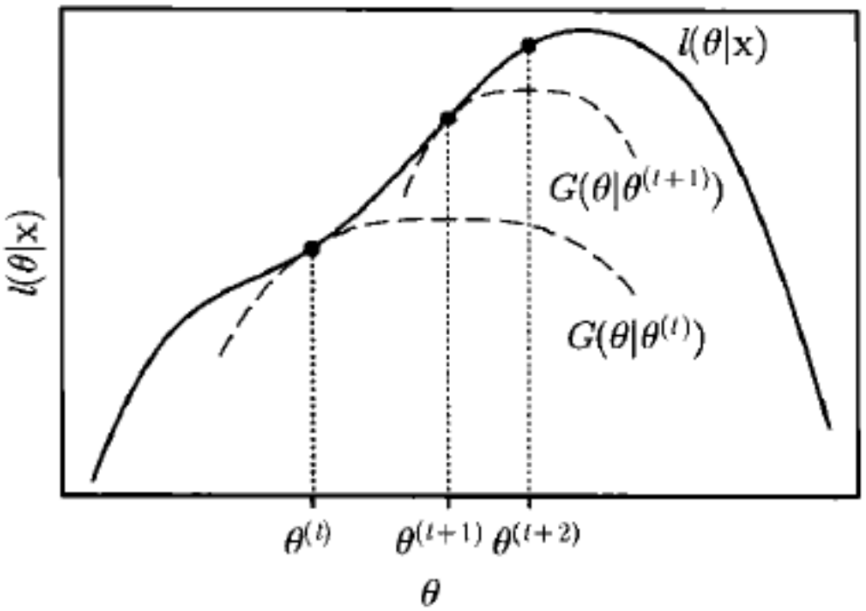
$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) \geqslant Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}|\boldsymbol{x}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) := G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

由于 $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ 和 $\ell(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})$ 独立于 θ, 函数 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 和 $G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 在相同的 θ 达到最大. 此外, G 在 $\theta^{(t)}$ 与 l 相切, 且在任一处低于 l. 我们说 G 是 l 的一个劣化函数 (minorizing function).

EM 策略将优化问题由观测数据的对数似然函数 l 转换到替代函数 G (有效地到 Q，因为后两项与 θ 无关，相当于常数), 这更便于最大化. G 的最大值点保证了在 l 值上的增加.

MM 算法 (Minorize-Maximization)

这个思想在下图中给出了解释. 每个 E 步等同于构造劣化函数 G, 而每个 M 步等同于最大化该函数以给出一个上升的路径.



Minorize-Maximization：每次迭代找到一个目标函数的下界函数（替代函数），求下界函数的最大值。

目录

EM 变型

EM 有时候表达式太过复杂，难以精确地计算期望或者求出最值，在这种情况下，我们选择牺牲一些计算速度或精确性做一些近似计算.

改进 E 步 (MCEM)

Monte Carlo EM：E 步需要找到在观测数据条件下完全数据的期望对数似然. 我们已经用 $Q(\theta|\theta^{(t)})$ 表示该期望，当该期望难以解析计算时，可以用 Monte Carlo 方法来近似.

1. Wei and Tanner 提出第 t 个 E 步可以用下面的两步替代
   1. 从 $p_{\boldsymbol{z}|\boldsymbol{x}}(z|\boldsymbol{x},\boldsymbol{\theta}^{(t)})$ 中抽取独立同分布的缺失数据集 $\boldsymbol{z}_1^{(t)},\cdots,\boldsymbol{z}_{m^{(t)}}^{(t)}$. 每个 $\boldsymbol{z}_j^{(t)}$ 是用来补齐观测数据集的所有缺失值的一个向量，这样 $\boldsymbol{y}_j=(\boldsymbol{x},\boldsymbol{z}_j)$ 表示一个补齐的数据集.
   2. 计算 $\hat{Q}^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})=\frac{1}{m^{(t)}}\sum_{j=1}^{m^{(t)}}\log p_{\boldsymbol{y}}(\boldsymbol{y}_j^{(t)}|\boldsymbol{\theta})$.

   那么 $\hat{Q}^{(t+1)}(\theta|\boldsymbol{\theta}^{(t)})$ 就是 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 的 Monte Carlo 估计.
2. 第 t 个 M 步改为最大化 $\hat{Q}^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

注意 $m^{(t)}$ 跟迭代次数有关，所以我们可在迭代初期选择小的 $m^{(t)}$，末期选择大的 $m^{(t)}$ 来得到精准估计。

MCEM 示例：正态混合模型

考虑一个简单的高斯混合模型 (GMM)：

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2)$$

其中：

- π_k 为类别 k 的混合权重，满足 $\sum_{k=1}^{K} \pi_k = 1$；
- $\mathcal{N}(x|\mu_k, \sigma_k^2)$ 表示均值 μ_k，方差 σ_k^2 (已知) 的正态分布；
- 观测数据 $X = \{x_1, x_2, ..., x_n\}$，但类别标签 $Z = \{z_1, z_2, ..., z_n\}$ 是缺失的.

GMM 的 MCEM 过程

我们希望用 MCEM 估计参数 $\theta = (\pi_k, \mu_k)$. 由于 Z 是缺失变量，我们采用 Monte Carlo 近似 E 步:

E 步 (Monte Carlo 近似):

1. 对于每个观测值 x_i，从条件分布 $p(z_i|x_i, \theta^{(t)})$ 采样 $m^{(t)}$ 个样本，记作 $z_i^{(j)}$.
2. 计算 $\hat{Q}(\theta|\theta^{(t)})$:

$$\hat{Q}(\theta|\theta^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \sum_{i=1}^{n} \log p(x_i, z_i^{(j)}|\theta).$$

M 步:

- **计算新的参数估计值** $\theta^{(t+1)} = \arg\max_\theta \hat{Q}(\theta|\theta^{(t)})$.
- **更新**:

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[I(z_i = k)|x_i, \theta^{(t)}] \leftarrow \frac{1}{n}\frac{1}{m^{(t)}}\sum_{j=1}^{m^{(t)}}\sum_{i=1}^{n} I(z_i^{(j)} = k),$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} x_i\mathbb{E}[I(z_i = k)|x_i, \theta^{(t)}]}{\sum_{i=1}^{n}\mathbb{E}[I(z_i = k)|x_i, \theta^{(t)}]} \leftarrow \frac{\sum_{i=1}^{n}\sum_{j=1}^{m^{(t)}} x_i I(z_i^{(j)} = k)}{\sum_{i=1}^{n}\sum_{j=1}^{m^{(t)}} I(z_i^{(j)} = k)}.$$

改进 M 步

EM 算法的吸引力之一在于 $Q(\theta|\theta^{(t)})$ 的求导和最大化通常比不完全数据极大似然的计算简单，这是因为 $Q(\theta|\theta^{(t)})$ 与完全数据似然有关.

然而，在某些情况下，即使导出 $Q(\theta|\theta^{(t)})$ 的 E 步是直接了当的，M 步也不容易实施. 为此人们提出了多种策略以便于 M 步的实施.

ECM 算法

Meng 和 Rubin 的 ECM 算法是用一系列计算较简单的条件极大化 (conditional maximization) 步骤代替 M 步. 每次条件极大化均被设计为一个简单的优化问题, 该优化问题把 θ 限制在某特殊子空间, 使得可以得到解析解或非常初等的数值解.

我们称第 t 个 E 步后的所有 CM 步所形成的集合为一个 CM 循环. 因此, ECM 的第 t 次迭代包括第 t 个 E 步和第 t 次 CM 循环. 令 S 表示每个 CM 循环里 CM 步的数目. 第 t 次循环里第 s 个 CM 步需要在约束

$$g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$$

下最大化 $Q(\theta|\theta^{(t)})$, 其中 $\theta^{(t+(s-1)/S)}$ 是在当前循环的第 $(s-1)$ 个 CM 步中求得的极大值点. 当 S 个 CM 步的整个循环完成时，我们令 $\theta^{(t+1)} = \theta^{(t+S/S)}$ 并进行第 $(t+1)$ 次迭代的 E 步.

第 t 次循环里第 s 个 CM 步需要在约束

$$g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$$

下最大化 $Q(\theta|\theta^{(t)})$.

这定义看起来很复杂，但其实很好理解，原理就是用之前迭代得到的参数来约束这一步参数的搜索空间，以方便搜索。上式也可写成：

$$\max_{\theta} Q(\theta|\theta^{(t)}), \quad \text{when} \quad g_s(\theta, \theta^{(t+(s-1)/S)}) = 0.$$

构造有效 ECM 算法的技巧在于巧妙地选择约束条件. 通常，可自然地把 θ 分成 S 个子向量 $\theta = (\theta_1, \cdots, \theta_S)$. 然后在第 s 个 CM 步中，我们可以固定 θ 其余的元素而关于 θ_s 寻求最大化 Q.

这等同于用函数 $g_s(\theta) = (\theta_1, \cdots, \theta_{s-1}, \theta_{s+1}, \cdots, \theta_S)$ 导出的约束条件.

类似坐标下降法 (Coordinate Descent).

另外，第 s 个 CM 步也可以在固定 θ_s 下关于 θ 的其他元素最大化 Q. 在这种情况下，$g_s(\theta) = \theta_s$. 也可根据特定的问题背景想象其他的约束体系. 比如线性约束:

$$A\theta = A\theta^{(t)}.$$

ECM 的一种变型是在每两个 CM 步之间插入一个 E 步，由此在 CM 循环的每一个阶段均更新了 Q.

ECM 算法的一个例子：k 个混合正态模型

我们考虑一个 k 组分的混合正态模型 (Gaussian Mixture Model, GMM)，其中观测数据 $X = \{x_1, \ldots, x_n\}$ 被假设来自 k 个正态分布的混合：

$$p(x|\theta) = \sum_{j=1}^{k} \pi_j \mathcal{N}(x|\mu_j, \sigma_j^2),$$

其中 $\theta = (\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2)$ 是参数，且混合权重满足 $\sum_{j=1}^{k} \pi_j = 1$.

我们引入隐变量 Z_i 来指示 x_i 属于哪个正态成分，即：

$$P(Z_i = j) = \pi_j, \quad j = 1, \ldots, k.$$

ECM 算法可以用来求解该模型的极大似然估计.

E 步

在第 t 次迭代的 E 步，我们计算隐变量的后验概率：

$$w_{ij}^{(t)} = P(Z_i = j|x_i, \theta^{(t)}) = \frac{\pi_j^{(t)} \mathcal{N}(x_i|\mu_j^{(t)}, \sigma_j^{2(t)})}{\sum_{l=1}^{k} \pi_l^{(t)} \mathcal{N}(x_i|\mu_l^{(t)}, \sigma_l^{2(t)})}.$$

这表示在当前参数估计下，x_i 来自第 j 组分的概率.

CM 循环

在 ECM 的 M 步，我们不同时优化所有参数，而是将其拆分成多个 CM 步.

CM 步 1: 更新均值 μ_j

在第一个 CM 步，我们固定其他参数，优化对数似然的 Q 函数关于 μ_j:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n w_{ij}^{(t)} x_i}{\sum_{i=1}^n w_{ij}^{(t)}}, \quad j = 1, \ldots, k.$$

CM 步 2: 更新方差 σ_j^2

在第二个 CM 步，我们固定均值，优化 σ_j^2:

$$\sigma_j^{2(t+1)} = \frac{\sum_{i=1}^n w_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n w_{ij}^{(t)}}, \quad j = 1, \ldots, k.$$

CM 步 3: 更新混合权重 π_j

在第三个 CM 步，我们固定均值和方差，优化 π_j，但由于约束 $\sum_{j=1}^{k} \pi_j = 1$，我们使用:

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^{(t)}, \quad j = 1, \ldots, k.$$

完成一次 ECM 迭代

完成所有 CM 步后，我们得到新的参数 $\theta^{(t+1)}$，然后进入下一轮迭代.

* 注意上面更新其实跟普通 EM 算法式子一样，主要原因是 μ_j 和 π_j 使 Q 函数条件最大值的点不包含 σ_j^2.

在标准 EM 算法中，M 步需要同时优化所有参数，这在某些情况下计算复杂. ECM 通过将 M 步拆分为多个条件极大化 (CM) 步，每次只优化部分参数，使得优化问题变得更简单.

在 k 组分混合正态模型的例子中，我们将 M 步拆分为三个 CM 步，分别更新均值、方差和混合权重，每次优化时固定其他参数.

EM 梯度算法

如果最大化不能用解析的方法来实现，那么可以考虑采用迭代优化方法来实施每个 M 步. 这将会产生一个有嵌套迭代循环的算法. ECM 算法在 EM 算法的每次迭代中搬入 S 个条件最大化步骤，这也会产生嵌套迭代.

为避免嵌套循环的计算负担，Lange 提出用单步 Newton 法替代 M 步，从而可近似取得最大值而不用真正地精确求解. M 步是用由

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \boldsymbol{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \boldsymbol{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ &= \theta^{(t)} - Q''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \boldsymbol{l}'(\boldsymbol{\theta}^{(t)}|\boldsymbol{x}),\end{aligned}$$

给出的更新替代的，其中 $l'(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})$ 是当前迭代得分函数的估值. 注意第二个等式是由 $\theta^{(t)}$ 最大化 $Q(\theta|\theta^{(t)}) - \ell(\theta|x)$ 的结论得来的. (一般来说 $l'(\theta|x)$ 的表达式要简单些)

当然我们也可以用最速梯度法代替 Newton 法，这样避免了计算 Q 函数矩阵的逆矩阵.

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \alpha^{(t)} \boldsymbol{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ &= \boldsymbol{\theta}^{(t)} - \alpha^{(t)} \boldsymbol{l}'(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})\end{aligned}$$

一些深度学习的优化方法，如 Adam 也可纳入 EM 梯度算法中。

EM 梯度算法示例

为了更好地理解 EM 梯度算法的应用，我们以高斯混合模型 (GMM) 为例，展示如何使用 EM 梯度算法来优化参数。

假设观测数据 $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ 由两个高斯分布混合生成，即：

$$p(x|\boldsymbol{\theta}) = \pi_1 \mathcal{N}(x|\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x|\mu_2, \sigma_2^2)$$

其中 $\boldsymbol{\theta} = \{\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1, \sigma_2\}$ 是模型参数，且 $\pi_1 + \pi_2 = 1$。

E 步 (期望步)：

定义隐变量 z_i，表示 x_i 由哪个高斯分布生成。计算后验概率：

$$\gamma_{i1} = P(z_i = 1|x_i, \boldsymbol{\theta}^{(t)}) = \frac{\pi_1^{(t)} \mathcal{N}(x_i|\mu_1^{(t)}, \sigma_1^{(t)2})}{\sum_{j=1}^{2} \pi_j^{(t)} \mathcal{N}(x_i|\mu_j^{(t)}, \sigma_j^{(t)2})}$$

EM 梯度优化:

如果使用梯度方法优化 M 步，可以计算 Q 函数的梯度:

$$\frac{\partial Q}{\partial \mu_j} = \sum_{i=1}^{N} \gamma_{ij}(x_i - \mu_j)/\sigma_j^2$$

并使用梯度下降法更新:

$$\mu_j^{(t+1)} = \mu_j^{(t)} + \alpha \sum_{i=1}^{N} \gamma_{ij}(x_i - \mu_j^{(t)})/\sigma_j^{(t)2}$$

类似地，σ_j 和 π_j 也可使用梯度更新，避免求解解析形式的最大值。

目录

针对隐变量的其他处理方法

在许多情况下，缺失数据 Z 可能并非真正缺失，而是作为潜变量 (latent variable) 用于：

- 重新表述问题，使得优化更易处理。
- 通过 EM 算法迭代求解最大似然估计。

另一种常见的处理方式是：

- 通过积分或近似积分（如 Laplace 方法）消去潜变量，直接对剩余变量进行优化。
- MCMC 当中也有相同的操作，比如 collapsed Gibbs.

解析积分消去潜变量

当积分 $\int p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})\mathrm{d}z$ 可以解析计算时，我们可以直接得到观测数据的边际似然函数：

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \log p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = \log \int p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{z}$$

示例：

- 若 $\boldsymbol{Y}$ 服从联合高斯分布：

  $$\boldsymbol{Y} \sim \mathcal{N}(\mu, \Sigma)$$

  其中 $\boldsymbol{Z}$ 也是部分高斯变量，可以通过高斯积分解析消去 $\boldsymbol{Z}$，得到 $\boldsymbol{X}$ 的边际分布。

- 若 $\boldsymbol{Z}$ 服从简单的离散分布 $\{z_k\}$，则边际化可以通过求和：

  $$p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_k p_{\boldsymbol{Y}}(\boldsymbol{x}, z_k|\boldsymbol{\theta})$$

Laplace 方法近似积分

1. 设待积函数 $f(\boldsymbol{z}) = \exp(g(\boldsymbol{z}))$，其中 $g(\boldsymbol{z})$ 是一个光滑函数。
2. 计算 $g(\boldsymbol{z})$ 在其峰值点 $\hat{\boldsymbol{z}}$ 处的二阶泰勒展开：

$$g(\boldsymbol{z}) \approx g(\hat{\boldsymbol{z}}) + \frac{1}{2}(\boldsymbol{z} - \hat{\boldsymbol{z}})^T H (\boldsymbol{z} - \hat{\boldsymbol{z}})$$

   其中 H 是 $g(\boldsymbol{z})$ 的 Hessian 矩阵，$H = \nabla^2 g(\hat{\boldsymbol{z}})$。

3. 近似计算积分：

$$\int \exp(g(\boldsymbol{z}))\mathrm{d}z \approx \exp(g(\hat{\boldsymbol{z}})) \int \exp\left(\frac{1}{2}(\boldsymbol{z} - \hat{\boldsymbol{z}})^T H (\boldsymbol{z} - \hat{\boldsymbol{z}})\right) \mathrm{d}z$$

   右侧积分为高斯积分，可直接计算。

高斯积分公式

考虑标准的高斯积分：

$$\int \exp\left(-\frac{1}{2}\boldsymbol{z}^T A\boldsymbol{z}\right) \mathrm{d}z = \frac{(2\pi)^{d/2}}{|A|^{1/2}}$$

其中：

- A 是 $d \times d$ 的正定矩阵。
- $|A|$ 表示矩阵 A 的行列式。

在 Laplace 方法中，Hessian 矩阵 H 一般是负定的 ($\hat{z}$ 局部极大值)，因此需调整符号：

$$\begin{aligned}
&\int \exp\left(\frac{1}{2}(\boldsymbol{z}-\hat{\boldsymbol{z}})^T H(\boldsymbol{z}-\hat{\boldsymbol{z}})\right) \mathrm{d}z \\
=&\int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\hat{\boldsymbol{z}})^T(-H)(\boldsymbol{z}-\hat{\boldsymbol{z}})\right) \mathrm{d}z
\end{aligned}$$

Laplace 近似的最终表达

结合高斯积分的计算结果，我们得到 Laplace 近似的最终结果：

$$\int \exp(g(\boldsymbol{z}))\mathrm{d}z \approx \exp(g(\hat{\boldsymbol{z}}))(2\pi)^{d/2}|H|^{-1/2}$$

取对数得到：

$$\log(\int \exp(g(\boldsymbol{z}))\mathrm{d}z) \approx g(\hat{\boldsymbol{z}}) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|H|$$

推论：设 h 的值域为正，我们有

$$\int \exp(g(\boldsymbol{z}))h(\boldsymbol{z})\mathrm{d}z \approx \exp(g(\hat{\boldsymbol{z}}))h(\hat{\boldsymbol{z}})(2\pi)^{d/2}|H|^{-1/2}$$

Laplace 近似的总结：

- 通过寻找到 $g(z)$ 的最大值点 $\hat{z}$ 进行二阶近似。
- 计算 Hessian 矩阵 H，并利用高斯积分公式得到近似积分值。
- 适用于高维积分问题，尤其适用于贝叶斯推断中的后验计算：

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}, H^{-1})$$

选择策略：

- 若隐变量 Z 服从高斯分布，且可以解析积分，优先使用 **解析积分方法**。
- 若积分较难但 Z 具有单峰性，优先使用 **Laplace 方法**。
- 若 Z 为混合分布或离散变量，优先使用 **EM 算法**。