混合模型 GMM 及其 EM 算法解

目录

1 期末考纲要求

- 4. **混合模型（变分法** important**）**
  * 估计（似然函数表述，求解困难以及解决方法）（EM，变分，MM）
- 5. GMM **极大似然估计算法问题**
  * 参数更新公式 (E-step, M-step)

下面根据以上考纲，结合讲义中的主要内容，对混合模型、GMM 极大似然估计，以及对应的 EM、变分、MM 等算法思路与推导进行系统整理。

2 混合模型（变分法重要）

2.1 混合模型分类与极大似然估计问题

设有 K 个随机变量 ξ_k，其分布为 N_k。我们定义一个新的随机变量 ξ 表示它来自哪个 ξ_k，且满足：

$$\xi = \begin{cases} \xi_1, & A_1 \text{ 发生且} P(A_1) = \alpha_1, \\ \xi_2, & A_2 \text{ 发生且} P(A_2) = \alpha_2, \\ \quad\cdots \\ \xi_K, & A_K \text{ 发生且} P(A_K) = \alpha_K, \end{cases}$$

其中 $\sum_{k=1}^{K} \alpha_k = 1$ 且 $0 \le \alpha_k \le 1$。事件 $A_1, A_2, \ldots, A_K$ 两两互斥且并集为全集 S。则 ξ 的分布函数为：

$$F(x) = \sum_{k=1}^{K} \alpha_k F_k(x),$$

而其密度函数为：

$$p(x) = \sum_{k=1}^{K} \alpha_k p_k(x),$$

其中 F_k, p_k 分别为 ξ_k 的分布函数与密度函数。

在参数化情形（例如 $\xi_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$），记 θ_k 表示第 k 个分模型的参数（如 μ_k, σ_k^2），并令 $\theta = \{\theta_1, \ldots, \theta_K\}$ 以及 $\alpha_1, \ldots, \alpha_K$ 这 K 个加权系数一起构成混合模型的整体参数：

$$p(x; \theta) = \sum_{k=1}^{K} \alpha_k p_k\big(x; \theta_k\big).$$

给定样本 $\{x_i\}_{i=1}^{I}$，其中 x_i 来自随机变量 ξ 的 I 次独立采样实现。其对数似然函数可写为

$$L(\theta) \;=\; \ln p(\boldsymbol{x}; \theta) \;=\; \sum_{i=1}^{I} \ln\Big(\sum_{k=1}^{K} \alpha_k \, p_k(x_i; \theta_k)\Big).$$

于是，极大似然估计问题变为

$$\theta^* \;=\; \arg\max_{\theta} \sum_{i=1}^{I} \ln\Big(\sum_{k=1}^{K} \alpha_k \, p_k(x_i; \theta_k)\Big).$$

由于求解困难（对数与求和嵌套），我们通常使用 EM、变分法或 MM 算法来求解。

2.2 EM 算法与变分法

2.2.1 EM 算法简介

EM (Expectation-Maximization) 算法通过在似然函数中将 **部分未知量视为“隐变量”**，从而将对数似然进行松弛和迭代优化。记观测数据为 $\mathrm{X} = (X_1, X_2, \ldots, X_I)$，并引入隐变量 $\mathrm{Y} =$

$(Y_1, Y_2, \ldots, Y_I)$，其中 $Y_i \in \{1, \ldots, K\}$ 表示 X_i 来自哪个子分布。

对数似然函数为

$$L(\theta) \;=\; \ln p(\boldsymbol{x}; \theta).$$

借助隐变量 Y 以及条件概率分解，可将 $L(\theta)$ 写成

$$L(\theta) \;=\; \sum_{Y} \; p\big(Y \mid X; \theta^t\big) \; \ln\!\Big(\tfrac{p(X,Y;\theta)}{p(Y|X;\theta^t)}\Big) \;=\; \underbrace{\sum_{Y} p\big(Y \mid X; \theta^t\big) \ln p(X, Y; \theta)}_{Q(\theta;\theta^t)} - \underbrace{\sum_{Y} p\big(Y \mid X; \theta^t\big) \; \ln p(Y \mid X; \theta)}_{H(\theta;\theta^t)}.$$

EM 算法的核心是：只要

$$Q\big(\theta^{t+1}; \theta^t\big) \;\geq\; Q\big(\theta^t; \theta^t\big),$$

就能保证 $L(\theta^{t+1}) \geq L(\theta^t)$。故优化问题可转化为：

$$\theta^{t+1} = \arg\max_{\theta} \; Q\big(\theta; \theta^t\big).$$

EM 算法迭代步骤：

- E-step: 计算后验概率

  $$p\big(k \mid x_i; \theta^t\big) \;=\; \frac{\alpha_k^t \, p_k\big(x_i; \theta_k^t\big)}{\sum_{k=1}^{K} \alpha_k^t \, p_k\big(x_i; \theta_k^t\big)},$$

  这是所谓的 “责任”（responsibility）。

- M-step: 在固定上一步计算的后验概率的条件下，最大化

  $$Q\big(\theta; \theta^t\big) \;=\; \sum_{i=1}^{I} \sum_{k=1}^{K} p\big(k \mid x_i; \theta^t\big) \; \ln\!\Big(\alpha_k \, p_k\big(x_i; \theta_k\big)\Big).$$

  从而得到对 α_k, θ_k 的更新。

2.2.2 变分法视角

变分法中，我们利用凸对偶或鞍点原理，将原始问题中的“$\ln \sum_k$”通过一个变分上界/下界表示出来，从而转化为可分离求解的形式。比如有如下定理：

定理 1 (**变分形式的对数和**). *若* $\mathcal{P}_k(x) > 0$，*则*

$$\ln \sum_{k=1}^{K} \mathcal{P}_k(x) \;=\; \max_{\boldsymbol{u} \in \mathbb{U}} \Bigg\{ \sum_{k=1}^{K} \Big(\ln \mathcal{P}_k(x)\Big) \; u_k(x) \;-\; \sum_{k=1}^{K} u_k(x) \; \ln u_k(x) \Bigg\}, \tag{30}$$

其中 $\mathbb{U} = \Big\{ \boldsymbol{u} = (u_1(x), \ldots, u_K(x)) \;\Big|\; \sum_{k=1}^{K} u_k(x) = 1, \; 0 < u_k(x) < 1 \Big\}$。

定理的证明基于对函数 $-u\ln u$ 的凹性及拉格朗日乘子法，得到

$$u_k^*(x) \ = \ \frac{\mathcal{P}_k(x)}{\sum_{k'}\mathcal{P}_{k'}(x)}.$$

把它带回去正好得到 $\ln\sum_k \mathcal{P}_k(x)$。

对于混合模型的对数似然 $L(\theta)$，我们可写为

$$L(\theta) = \sum_{i=1}^{I}\ln\Big(\sum_{k=1}^{K}\alpha_k p_k(x_i;\theta_k)\Big) = \sum_{i=1}^{I}\max_{\mathrm{u}(x_i)\in\mathbb{U}}\Big\{\sum_{k=1}^{K}u_k(x_i)\ \ln\big(\alpha_k p_k(x_i;\theta_k)\big) - \sum_{k=1}^{K}u_k(x_i)\ \ln u_k(x_i)\Big\}.$$

令

$$\epsilon(\boldsymbol{u},\theta) = \sum_{i=1}^{I}\sum_{k=1}^{K}u_k(x_i)\ \ln\big(\alpha_k p_k(x_i;\theta_k)\big) - \sum_{i=1}^{I}\sum_{k=1}^{K}u_k(x_i)\ \ln u_k(x_i).$$

则最大似然估计可改写成等价的变分形式:

$$\theta^*,\boldsymbol{u}^* = \arg\max_{\theta,\boldsymbol{u}\in\mathbb{U}}\epsilon(\boldsymbol{u},\theta).$$

由此可用**交替极大**算法（类似 EM）:

$$\begin{cases}\boldsymbol{u}^{t+1} = \arg\max_{\boldsymbol{u}\in\mathbb{U}}\ \epsilon(\boldsymbol{u},\theta^t), & \text{(E-step)},\\ \theta^{t+1} = \arg\max_{\theta}\ \epsilon(\boldsymbol{u}^{t+1},\theta), & \text{(M-step)}.\end{cases} \tag{31}$$

可证明，E-step 理论解的 $u_k^{t+1}(x_i)$ 正好是 EM 算法中的后验 $p(k\mid x_i;\theta^t)$。这样就揭示了 EM 算法等价于一个有熵正则项的变分问题求解过程。

2.3 混合模型的 MM 算法思路

设最大化似然函数 $f(x) = \max_\theta L(\theta)$,我们也可考虑 MM (Majorization-Minimization/Maximization) **算法**来实现。若对于某个辅助函数 $g(\theta,\hat{\theta})$，我们有

$$(1)':\quad f(\theta)\ \geq\ g(\theta,\hat{\theta}),\quad \forall\theta,\hat{\theta},$$

$$(2)':\quad f(\hat{\theta})\ =\ g(\hat{\theta},\hat{\theta}),$$

则通过迭代

$$\theta^{t+1}\ =\ \arg\max_{\theta}\ g(\theta,\theta^t), \tag{42}$$

能够保证 $f(\theta^{t+1})\geq f(\theta^t)$。在混合模型极大似然求解中，也能构造出一个满足以上性质的 $g(\theta,\hat{\theta})$，与 EM 中的 $Q(\theta;\hat{\theta})$ 形式类似，从而得到保证单调上升的迭代更新。

3 GMM 极大似然估计与 EM 算法

3.1 高斯混合模型 (Gaussian Mixture Model) 的对数似然

以最常见的 GMM 为例，设

$$p(x;\theta) = \sum_{k=1}^{K} \alpha_k \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\Bigl(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\Bigr),$$

其中 $\theta = \{\alpha_k, \mu_k, \sigma_k^2\}_{k=1}^{K}$ 且 $\sum_{k=1}^{K} \alpha_k = 1$。给定样本集合 $\{x_i\}_{i=1}^{I}$，对数似然函数为

$$L(\theta) = \sum_{i=1}^{I} \ln\Biggl[\sum_{k=1}^{K} \alpha_k \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\bigl(-\tfrac{(x_i-\mu_k)^2}{2\sigma_k^2}\bigr)\Biggr].$$

3.2 EM 算法在 GMM 上的具体推导

E-step:　对每个样本 x_i，计算

$$p\bigl(k \mid x_i;\theta^t\bigr) = \frac{\alpha_k^t \,\frac{1}{\sigma_k^t}\, \exp\bigl(-\frac{(x_i-\mu_k^t)^2}{2(\sigma_k^t)^2}\bigr)}{\sum_{k'=1}^{K} \alpha_{k'}^t \,\frac{1}{\sigma_{k'}^t}\, \exp\bigl(-\frac{(x_i-\mu_{k'}^t)^2}{2(\sigma_{k'}^t)^2}\bigr)}. \tag{26}$$

M-step:　在固定上一步计算的 $p\bigl(k \mid x_i;\theta^t\bigr)$ 的条件下，最大化

$$Q\bigl(\theta;\theta^t\bigr) = \sum_{i=1}^{I}\sum_{k=1}^{K} p\bigl(k \mid x_i;\theta^t\bigr)\, \ln\Bigl[\alpha_k \,\frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\bigl(-\tfrac{(x_i-\mu_k)^2}{2\sigma_k^2}\bigr)\Bigr].$$

结合约束 $\sum_{k=1}^{K} \alpha_k = 1$ 用拉格朗日乘子法，可得更新公式：

- 对 α_k：
$$\alpha_k^{t+1} = \frac{1}{I}\,\sum_{i=1}^{I} p\bigl(k \mid x_i;\theta^t\bigr). \tag{27}$$

- 对 μ_k：
$$\mu_k^{t+1} = \frac{\sum_{i=1}^{I} x_i\, p\bigl(k \mid x_i;\theta^t\bigr)}{\sum_{i=1}^{I} p\bigl(k \mid x_i;\theta^t\bigr)}. \tag{28}$$

- 对 σ_k^2：
$$(\sigma_k^2)^{t+1} = \frac{\sum_{i=1}^{I} (\,x_i - \mu_k^{t+1}\,)^2\, p\bigl(k \mid x_i;\theta^t\bigr)}{\sum_{i=1}^{I} p\bigl(k \mid x_i;\theta^t\bigr)}. \tag{29}$$

将这些更新式反复迭代，便可得到似然函数单调上升直到收敛的 GMM 参数估计结果。

3.3 变分法在混合模型中的应用

如上所述，在更一般的混合模型（如指数族混合分布）里，若 $p_k(x;\theta_k) \propto \exp\big(f_k(x;\theta_k)\big)$，则通过变分法可将

$$\ln \sum_{k=1}^{K} \alpha_k \, p_k(x_i;\theta_k)$$

表示为一个带熵正则项的变分上界，从而与 EM 求解过程相对应。其核心思想与前述定理 (30) 的对数和变分形式相同。这个视角也为很多合并正则项的模型（如图模型、稀疏先验等）提供了可行解法。

3.4 正则化 EM：在目标中加入先验或正则项

有时我们在 $\epsilon(\boldsymbol{u},\theta)$ 中再加一个正则项 $\lambda \mathcal{R}(\boldsymbol{u})$，例如熵以外的空间平滑、图先验等，这就得到所谓的正则 EM 算法。典型形式如

$$\tilde{\epsilon}(\boldsymbol{u},\theta) = \epsilon(\boldsymbol{u},\theta) + \lambda \mathcal{R}(\boldsymbol{u}), \tag{37}$$

然后以同样交替极大方式：

$$\begin{cases} \boldsymbol{u}^{t+1} = \arg\max_{\boldsymbol{u}\in\mathbb{U}} \tilde{\epsilon}\big(\boldsymbol{u},\theta^t\big), \\ \theta^{t+1} = \arg\max_{\theta} \tilde{\epsilon}\big(\boldsymbol{u}^{t+1},\theta\big). \end{cases} \tag{38}$$

具体的 $\mathcal{R}$ 形式各异，往往考虑梯度或散度的正则，或图先验项等。

4 MM (Majorization-Minimization) 算法补充

MM 是另一种通用的迭代方法，可用于最大化或最小化复杂函数。假设我们要解

$$x^* = \arg\min f(x).$$

- 若可构造辅助函数 $g(x,y)$，使得对任意 x,y，满足：

  $$(1)\ \ f(x) \ \le\ g(x,y), \quad (2)\ \ f(y) = g(y,y),$$

  则沿着

  $$x^{t+1} = \arg\min_{x}\ g\big(x,x^t\big) \tag{40}$$

  而迭代，可保证 $f(x^{t+1}) \le f(x^t)$，使目标单调下降（或不增）。

- 如果目标是 $\max f(x)$，且对某个 $g(x,y)$ 有

  $$(1)'\ \ f(x) \ \ge\ g(x,y), \quad (2)'\ \ f(y) = g(y,y),$$

则沿着

$$x^{t+1} = \arg\max_x \; g\big(x, x^t\big) \tag{41}$$

而迭代，可保证 $f(x^{t+1}) \geq f(x^t)$。

在混合模型的极大似然估计里，其实 EM 算法就可视作一种 MM，因为我们找到了一个 $g(\theta, \hat{\theta}) = Q(\theta, \hat{\theta}) + C$ 满足

$$L(\theta) \;\geq\; g(\theta, \hat{\theta}), \quad L(\hat{\theta}) = g(\hat{\theta}, \hat{\theta}),$$

并采用

$$\theta^{t+1} = \arg\max_\theta \; g\big(\theta, \theta^t\big).$$

因此 EM 一种特殊的 MM 算法。

总结:

(1) **混合模型**是通过加权和来描述多分布叠加的情形，其似然函数带有 "$\ln(\sum)$" 的嵌套结构，直接求解常常较困难。(2) EM **算法**将隐变量显式化并分两步迭代：E-step 计算后验；M-step 更新参数，是最常用的混合模型极大似然求解算法。(3) **变分法**提供了对数和的凸对偶表述，能够揭示 EM 算法的本质：它实现了一个带熵正则的变分极大过程，也能方便地纳入更多先验 (正则) 项。(4) MM **算法**是一种更一般的迭代机制，EM 算法可视为其特例。在"最大化"或"最小化"时，通过辅函数 g 使目标函数单调更新。