Gaussian Mixture Model 03 Expectation Maximization

Chen Gong

25 December 2019

上一小节中，我们看到了使用极大似然估计的方法，我们根本就求不出最优参数 θ 的解析解。所以，我们使用迭代的方法来求近似解。

EM 算法的表达式，可以被我们写为：

$$\theta^{(t+1)} = \arg\max_{\theta} \underbrace{\mathbb{E}_{P(Z|X,\theta^{(t)})}\left[\log P(X,Z|\theta)\right]}_{Q(\theta,\theta^{(t)})} \tag{1}$$

经过一系列的迭代，我们可以得到 $\theta^0, \theta^1, \cdots, \theta^{(t)}$，迭代到一定次数以后我们得到的 $\theta^{(N)}$ 就是我们想要得到的结果。EM 算法大体上可以分成两个部分，E-step 和 M-step，

1 E-Step

$$\begin{aligned}
Q(\theta,\theta^{(t)}) &= \int_Z \log P(X,Z|\theta)\cdot P(Z|X,\theta^{(t)})dZ \\
&= \sum_Z \log \prod_{i=1}^N P(x_i,z_i|\theta)\cdot \prod_{i=1}^N P(z_i|x_i,\theta^{(t)})dZ \\
&= \sum_{z_1,\cdots,z_N} \sum_{i=1}^N \log P(x_i,z_i|\theta)\cdot \prod_{i=1}^N P(z_i|x_i,\theta^{(t)})dZ \\
&= \sum_{z_1,\cdots,z_N} \left[\log P(x_1,z_1|\theta) + \log P(x_2,z_2|\theta) + \cdots \log P(x_N,z_N|\theta)\right]\cdot \prod_{i=1}^N P(z_i|x_i,\theta^{(t)})dZ
\end{aligned} \tag{2}$$

为了简化推导，我们首先只取第一项来化简一下，

$$\begin{aligned}
& \sum_{z_1,\cdots,z_N} \log P(x_1,z_1|\theta)\cdot \prod_{i=1}^N P(z_i|x_i,\theta^{(t)})dZ \\
=& \sum_{z_1,\cdots,z_N} \log P(x_1,z_1|\theta)\cdot P(z_1|x_1,\theta^{(t)})\cdot \prod_{i=2}^N P(z_i|x_i,\theta^{(t)})dZ \\
=& \sum_{z_1} \log P(x_1,z_1|\theta)\cdot P(z_1|x_1,\theta^{(t)})\cdot \sum_{z_2,\cdots,z_N} \prod_{i=2}^N P(z_i|x_i,\theta^{(t)})dZ
\end{aligned} \tag{3}$$

而：

$$\begin{aligned}\sum_{z_2,\cdots,z_N}\prod_{i=2}^{N}P(z_i|x_i,\theta^{(t)}) &= \sum_{z_2,\cdots,z_N}P(z_2|x_2,\theta^{(t)})\cdot P(z_3|x_3,\theta^{(t)})\cdots P(z_N|x_N,\theta^{(t)}) \\ &=\sum_{z_2}P(z_2|x_2,\theta^{(t)})\cdot\sum_{z_3}P(z_3|x_3,\theta^{(t)})\cdots\sum_{z_N}P(z_N|x_N,\theta^{(t)}) \\ &=1\cdot 1\cdots 1 \\ &=1\end{aligned} \tag{4}$$

所以，式 (3) 也就等于：

$$\sum_{z_1,\cdots,z_N}\log P(x_1,z_1|\theta)\cdot\prod_{i=1}^{N}P(z_i|x_i,\theta^{(t)})dZ = \sum_{z_1}\log P(x_1,z_1|\theta)\cdot P(z_1|x_1,\theta^{(t)}) \tag{5}$$

将式 (5) 中得到的结果，代入到式 (2) 中，我们就可以得到：

$$\begin{aligned}Q(\theta,\theta^{(t)}) &= \sum_{z_1}\log P(x_1,z_1|\theta)\cdot P(z_1|x_1,\theta^{(t)})+\cdots+\sum_{z_N}\log P(x_N,z_N|\theta)\cdot P(z_N|x_N,\theta^{(t)}) \\ &=\sum_{i=1}^{N}\sum_{Z_i}\log P(x_i,z_i|\theta)\cdot P(z_i|x_i,\theta^{(t)})\end{aligned} \tag{6}$$

那么，下一步我们就是要找到，$P(x_i,z_i|\theta)$ 和 $P(z_i|x_i,\theta^{(t)})$ 的表达方式了。其中：

$$P(X,Z)=P(Z)P(X|Z)=P_Z\cdot\mathcal{N}(X|\mu_Z,\Sigma_Z) \tag{7}$$

$$P(Z|X)=\frac{P(X,Z)}{P(X)}=\frac{P_Z\cdot\mathcal{N}(X|\mu_Z,\Sigma_Z)}{\sum_{i=1}^{K}P_{Zi}\cdot\mathcal{N}(X|\mu_{Zi},\Sigma_{Zi})} \tag{8}$$

所以，我们将式 (8) 代入到式 (6) 中，就可以得到：

$$Q(\theta,\theta^{(t)})=\sum_{i=1}^{N}\sum_{Z_i}\log P_{Z_i}\cdot\mathcal{N}(X|\mu_{Z_i},\Sigma_{Z_i})\cdot\frac{P_{Z_i}^{\theta(t)}\cdot\mathcal{N}(x_i|\mu_{Z_i}^{\theta(t)},\Sigma_{Z_i}^{\theta(t)})}{\sum_{k=1}^{K}P_k^{\theta(t)}\cdot\mathcal{N}(x_i|\mu_k^{\theta(t)},\Sigma_k^{\theta(t)})} \tag{9}$$

2 M-Step

根据我们在 E-Step 中的推导，我们可以得到：

$$\begin{aligned}Q(\theta,\theta^{(t)}) &= \sum_{i=1}^{N}\sum_{Z_i}\log P_{Z_i}\cdot\mathcal{N}(X|\mu_{Z_i},\Sigma_{Z_i})\cdot\underbrace{\frac{P_{Z_i}^{\theta(t)}\cdot\mathcal{N}(x_i|\mu_{Z_i}^{\theta(t)},\Sigma_{Z_i}^{\theta(t)})}{\sum_{k=1}^{K}P_k^{\theta(t)}\cdot\mathcal{N}(x_i|\mu_k^{\theta(t)},\Sigma_k^{\theta(t)})}}_{P(Z_i|X_i,,\theta^{(t)})} \\ &=\sum_{Z_i}\sum_{i=1}^{N}\log\left(P_{Z_i}\cdot\mathcal{N}(X|\mu_{Z_i},\Sigma_{Z_i})\right)\cdot P(Z_i|X_i,,\theta^{(t)}) \\ &=\sum_{k=1}^{K}\sum_{i=1}^{N}\log\left(P_k\cdot\mathcal{N}(X|\mu_k,\Sigma_k)\right)\cdot P(Z_i=C_k|X_i,,\theta^{(t)})\quad(Z_i=C_k) \\ &=\sum_{k=1}^{K}\sum_{i=1}^{N}(\log P_k+\log\mathcal{N}(X_i|\mu_k,\Sigma_k))\cdot P(Z_i=C_k|X_i,\theta^{(t)})\end{aligned} \tag{10}$$

我们的目的也就是进行不断迭代，从而得出最终的解，用公式表达也就是：

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \theta^{(t)}) \tag{11}$$

我们需要求解的参数也就是，$\theta^{(t+1)} = \{P_1^{(t+1)}, \cdots, P_k^{(t+1)}, \mu_1^{(t+1)}, \cdots, \mu_k^{(t+1)}, \Sigma_1^{(t+1)}, \cdots, \Sigma_k^{(t+1)}\}$。

首先，我们来展示一下怎么求解 $P_K^{(t+1)}$：

由于在等式 (10)，$\sum_{k=1}^K \sum_{i=1}^N (\log P_k + \log \mathcal{N}(X|\mu_k, \Sigma_k)) \cdot P(Z_i = C_k|X_i, , \theta^{(t)})$ 中的 $\log \mathcal{N}(X|\mu_k, \Sigma_k)$ 部分和 P_k 并没有什么关系。所以，可以被我们直接忽略掉。所以，求解问题，可以被我们描述为：

$$\begin{cases} \arg\max_{P_k} \sum_{k=1}^K \sum_{i=1}^N \log P_k \cdot P(Z_i = C_k|X_i, \theta^{(t)}) \\ s.t. \quad \sum_{k=1}^K P_k = 1 \end{cases} \tag{12}$$

使用拉格朗日算子法，我们可以写成：

$$\mathcal{L}(P, \lambda) = \sum_{k=1}^K \sum_{i=1}^N \log P_k \cdot P(Z_i = C_k|X_i, \theta^{(t)}) + \lambda(\sum_{k=1}^K P_k - 1) \tag{13}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(P, \lambda)}{\partial P_k} &= \sum_{i=1}^N \frac{1}{P_k} \cdot P(Z_i = C_k|X_i, \theta^{(t)}) + \lambda = 0 \\ &\Rightarrow \sum_{i=1}^N P(Z_i = C_k|X_i, \theta^{(t)}) + P_k \lambda = 0 \\ &\xRightarrow{k=1,\cdots,K} \sum_{i=1}^N \underbrace{\sum_{k=1}^K P(Z_i = C_k|X_i, \theta^{(t)})}_{1} + \underbrace{\sum_{k=1}^K P_k}_{1} \lambda = 0 \\ &\Rightarrow N + \lambda = 0 \end{aligned} \tag{14}$$

所以，我们可以轻易的得到 $\lambda = -N$，所以有

$$P_K^{(t+1)} = \frac{1}{N} \sum_{i=1}^N P(Z_i = C_k|X_i, \theta^{(t)}) \tag{15}$$

那么，我们所有想要求的参数也就是 $P^{(t+1)} = (P_1^{(t+1)}, P_2^{(t+1)}, \cdots, P_k^{(t+1)})$。

求解 $P_k^{(t+1)}$ 是一个有约束的求最大值问题，由于带约束所以我们要使用拉格朗日乘子法。而且这里使用到了一个 track，也就是将从 1 到 k，所有的数据集做一个整合，非常的精彩，这样就直接消掉了 P_k 无法计算的问题。而至于 θ 的其他部分，也就是关于 $\{\mu_1^{(t+1)}, \cdots, \mu_k^{(t+1)}, \Sigma_1^{(t+1)}, \cdots, \Sigma_k^{(t+1)}\}$ 的计算，使用的方法也是一样的，这个问题就留给各位了。

为什么极大似然估计搞不定的问题，放在 EM 算法里面我们就可以搞定了呢？我们来对比一下两个方法中，需要计算极值的公式。

$$\sum_{k=1}^K \sum_{i=1}^N (\log P_k + \log \mathcal{N}(X_i|\mu_k, \Sigma_k)) \cdot P(Z_i = C_k|X_i, \theta^{(t)}) \tag{16}$$

$$\arg\max_{\theta} \sum_{i=1}^N \log \sum_{k=1}^K P_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k) \tag{17}$$

极大似然估计一开始计算的就是 $P(X)$，而 EM 算法中并没有出现有关 $P(X)$ 的计算，而是全程计算都是 $P(X, Z)$。而 $P(X)$ 实际上就是 $P(X, Z)$ 的求和形式。所以，每次单独的考虑 $P(X, Z)$ 就避免了在 log 函数中出现求和操作。