EM算法

许王莉

中国人民大学　统计学院

Contents

- 几个特殊分布参数的EM算法
  - (1). 正态分布参数的EM算法
  - (2). 多项分布参数的EM算法
  - (3). 混合分布参数的EM算法
- 一般混合分布的EM推断
- 高斯混合分布的EM推断

Introduction

- EM is typically used to compute maximum likelihood estimates given incomplete samples.
- The EM algorithm estimates the parameters of a model iteratively. Starting from some initial guess, each iteration consists of an E step (Expectation step) an M step (Maximization step)

对来自正态总体$N(\mu,\sigma^2)$的样本$x=(x_1,x_2,\cdots,x_n)^\tau$, 我们用如下方法给出参数$(\mu,\sigma^2)$的估计。

令

$$
\begin{aligned}
L(\mu,\sigma^2|x) &= f(x|\mu,\sigma^2) = f(x_1|\mu,\sigma^2)\cdots f(x_n|\mu,\sigma^2) \\
&= (\frac{1}{2\pi\sigma^2})^{n/2}\exp\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\}
\end{aligned}
$$

$$
\begin{aligned}
l(\mu,\sigma^2|x) &= \log L(\mu,\sigma^2|x) \\
&= -\frac{n}{2}\log\sigma^2-\frac{n}{2}\log 2\pi-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2+\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i-\frac{n\mu^2}{2\sigma^2}
\end{aligned}
$$

正态分布极大似然估计

令：

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2 | x) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{n\mu}{\sigma^2} = 0$$

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2 | x) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} x_i^2 - \frac{\mu}{\sigma^4} \sum_{i=1}^{n} x_i + \frac{n\mu^2}{2\sigma^4} = 0$$

可得：

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \hat{\mu}^2$$

数据缺失的情况（E步）

假定观测到$x=(x_1,\cdots,x_m,\underbrace{x_{m+1},\cdots,x_n}_{\text{缺失数据}})$，在缺失情况下，

$$\begin{aligned}\hat{\mu} &= \frac{1}{n}\Big(\sum_{i=1}^{m} x_i + \sum_{j=m+1}^{n} x_j\Big) \\ \hat{\sigma}^2 &= \frac{1}{n}\Big(\sum_{i=1}^{m} x_i^2 + \sum_{j=m+1}^{n} x_j^2\Big) - \hat{\mu}^2\end{aligned}$$

由于上式中$\sum_{j=m+1}^{n} x_j$和$\sum_{j=m+1}^{n} x_j^2$未知，我们需要用观测到的数据对这两项做估计，则

$$E\Big(\sum_{j=m+1}^{n} x_j|x\Big) = (n-m)\hat{\mu}^{(t)}, \quad E\Big(\sum_{j=m+1}^{n} x_j^2|x\Big) = (n-m)(\hat{\mu}^{(t)2}+\sigma^{2(t)}),$$

这里：$(\mu^{(t)},\sigma^{2(t)})$是估计参数的第$i$次迭代。

数据缺失的情况（M步）

令：

$$
\begin{aligned}
s_1^{(t)} &= \sum_{i=1}^{m} x_i + (n-m)\hat{\mu}^{(t)} \\
s_2^{(t)} &= \sum_{i=1}^{m} x_i^2 + (n-m)\left(\hat{\mu}^{(t)2} + \sigma^{2(t)}\right),
\end{aligned}
$$

可得：

$$
\begin{aligned}
\mu^{(t+1)} &= \frac{s_1^{(t)}}{n} \\
\sigma^{2(t+1)} &= \frac{s_2^{(t)}}{n} - \hat{\mu}^{(t+1)2}
\end{aligned}
$$

多项分布

某随机实验如果有n个可能结局$A_1, A_2, \cdots, A_n$，它们的概率分布分别是$p_1, p_2, \cdots, p_n$，那么在n次采样$x = (x_1, \cdots, x_n)$的总结果中，A_1出现x_1次、A_2出现x_2 次、…、A_n出现x_n次的这种事件的出现概率P有下面公式：

$$p(x|p_1, \cdots, p_n) = \begin{cases} \frac{N!}{x_1! \cdots x_n!} p_1^{x_1} \cdots, p_n^{x_n} \cdots, & \sum_{i=1}^{n} x_i = n \\ 0, & \text{其他}. \end{cases}$$

在上述多项分布中，假设

$$(p_1, \cdots, p_n) = (\frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4}, \frac{1}{2}),$$

则θ的对数极大似然函数为

$$\begin{aligned} l(\theta|x) &= \log L(\theta|x) \\ &= x_1 \log(\frac{1}{2} - \frac{\theta}{2}) + x_2 \log(\frac{\theta}{4}) + x_3 \log(\frac{\theta}{4}) + const. \end{aligned}$$

令

$$\frac{\partial}{\partial\theta} l(\theta|x) = -\frac{x_1}{1-\theta} + \frac{x_2}{\theta} + \frac{x_3}{\theta} = 0,$$

可得：

$$\hat{\theta} = \frac{x_2 + x_3}{x_1 + x_2 + x_3}$$

Mixed Attributes（E步）

假定$x = (x_1, x_2, \underbrace{x_3 + x_4}_{x_3\text{未知}})^\tau$，此时

$$\hat{\theta} = \frac{x_2 + \color{red}{x_3}}{x_1 + x_2 + \color{red}{x_3}}$$

给定$\theta^{(t)}$, 怎么估计x_3

$$E_{\theta^{(t)}}(x_3|x) = (x_3 + x_4)\frac{\frac{\theta^{(t)}}{4}}{\frac{1}{2} + \frac{\theta^{(t)}}{4}} = \hat{x}_3^{(t)}$$

Mixed Attributes（M步）

M步为：

$$\hat{\theta}^{(t+1)} = \frac{x_2 + \hat{x}_3^{(t)}}{x_1 + x_2 + \hat{x}_3^{(t)}}$$

Exercise

$$x_{obs} = (x_1, x_2, x_3 + x_4)^\tau = (38, 34, 125)^\tau$$

通过EM算法给出θ的估计.

对于如下观测到的数据,

#孩子	0	1	2	3	4	5	6
#妇女	n_0	n_1	n_2	n_3	n_4	n_5	n_6

从这个表中可以看出, n_0个妇女没有生孩子.

用二项分布M表示妇女是否结婚, $M=1$表示结婚, $M=0$表示未婚. 假定$P(M=0)=\xi$, 则$P(M=1)=1-\xi$. 假定给定$M=1$的情况下, 妇女生孩子的个数服从Possion分布$P(\lambda)$, 也就是

$$P(X=x|M=1)=\frac{\lambda^x e^{-\lambda}}{x!}.$$

不难知道$P(X=0|M=0)=1$.

对于上表中n_0个妇女没有生孩子,它包括n_A个结婚的妇女和n_B个没有结婚的妇女. 也就是$n_0 = n_A + n_B$. 完整的数据结构如下:

#孩子	0	0	1	2	3	4	5	6
#妇女	n_A	n_B	n_1	n_2	n_3	n_4	n_5	n_6
对应概率	p_A	p_B	p_1	p_2	p_3	p_4	p_5	p_6

这里:

$$p_A = \xi, \quad p_B = e^{-\lambda}(1-\xi), \quad p_x = \frac{\lambda^x e^{-\lambda}}{x!}(1-\xi), \quad x = 1, 2, \cdots$$

Complete Data Likelihood

假定观测到的数据为$n_{obs} = (n_0, n_1, \cdots, n_6)$, 这里$n_0 = n_A + n_B$. 假定$n = (n_A, n_B, n_1, \cdots, n_6)$, 则似然函数为

$$L(\xi, \lambda | n) = \frac{(n_A + n_B + n_1 + \cdots + n_6)}{n_A! n_B! n_1! \cdots n_6!} p_A^{n_A} p_B^{n_B} p_1^{n_1} \cdots p_6^{n_6}$$

$$= \frac{(n_A + n_B + n_1 + \cdots + n_6)}{n_A! n_B! n_1! \cdots n_6!} \xi^{n_A} [e^{-\lambda}(1-\xi)]^{n_B} \Pi_{x=1}^{6} [\frac{\lambda^x e^{-\lambda}}{x!}(1-\xi)]^{n_x}.$$

Log-Likelihood

对数似然函数为

$$
\begin{aligned}
l(\xi, \lambda | n) \quad = \quad & n_A \log \xi - n_B \lambda + n_B \log(1 - \xi) \\
& + \sum_{x=1}^{6} n_x [-\lambda + x \log \lambda + \log(1 - \xi)] + k.
\end{aligned}
$$

对对数似然函数求导，可得

$$
\frac{l(\xi, \lambda | n)}{\xi} = \frac{n_A}{\xi} - \frac{n_B + n_1 + \cdots + n_6}{1 - \xi} = 0
$$

$$
\Rightarrow \hat{\xi} = \frac{n_A}{N}
$$

Log-Likelihood

对对数似然函数求导，可得

$$\frac{l(\xi,\lambda|n)}{\xi} = \frac{n_A}{\xi} - \frac{n_B + n_1 + \cdots + n_6}{1-\xi} = 0$$

$$\Rightarrow \hat{\xi} = \frac{n_A}{N}$$

$$\frac{l(\xi,\lambda|n)}{\lambda} = -(n_B + n_1 + \cdots + n_6) + \frac{1}{\lambda}\sum_{x=1}^{6} x n_x = 0$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{x=1}^{6} x n_x}{N - n_A}$$

E-Step

给定$\xi^{(t)}$, $\lambda^{(t)}$, 则

$$n_A^{(t)} = E_{\xi^{(t)},\lambda^{(t)}}[n_A | n_{obs}] = \frac{n_0 \xi^{(t)}}{\xi^{(t)} + (1-\xi^{(t)})e^{-\lambda^{(t)}}}$$

M-Step

在M步，可得

$$\xi^{(t+1)} = \frac{n_A^{(t)}}{N}, \quad \lambda^{(t+1)} = \frac{\sum_{x=1}^{6} x n_x}{N - n_A^{(t)}},$$

这里

$$n_A^{(t)} = \frac{n_0 \xi^{(t)}}{\xi^{(t)} + (1 - \xi^{(t)}) e^{-\lambda^{(t)}}}$$

例子

对于如下观测数据，给出ξ和λ的估计

♯孩子	0	1	2	3	4	5	6
♯妇女	3062	587	284	103	33	4	2

迭代结果如下

0	0.750000	0.400000	2502.779	559.221
1	0.614179	1.035478	2503.591	558.409
2	0.614378	1.036013	2504.219	557.781
3	0.614532	1.036427	2504.705	557.295
4	0.614652	1.036748	2505.081	556.919
5	0.614744	1.036996	2505.371	556.629

Maximum Likelihood

假定$X = \{x_1, x_2, \cdots, x_N\}$, 极大似然函数如下:

$$L(\Theta|X) = p(X|\Theta) = \Pi_{i=1}^{N} p(x_i|\Theta).$$

则参数的估计如下

$$\Theta^* = \arg\max_{\Theta} L(\Theta|X).$$

Latent Variables

假定观测到的$X=\{x_1,x_2,\cdots,x_N\}$不是完整数据,记$Y=\{y_1,y_2,\cdots,y_N\}$为没有观测到的数据(潜变量), 完整数据集记为$Z=(X,Y)$

基于完整数据的似然函数如下:

$$L(\Theta|Z)=p(Z|\Theta)=p(X,Y|\Theta)=p(Y|X,\Theta)p(X|\Theta)$$

Complete Data Likelihood

$$L(\Theta|Z) = p(Z|\Theta) = p(X, Y|\Theta) = p(Y|X, \Theta)p(X|\Theta)$$

1. 如果Θ给定，$L(\Theta|Z)$是变量Y的函数；
2. $p(Y|X, \Theta)$是变量Y和参数Θ的函数；
3. $p(X|\Theta)$是参数Θ的函数.

Expectation Step

记$\Theta^{(i-1)}$表示第$(i-1)$步得到的参数的估计，定义

$$
\begin{aligned}
Q(\Theta, \Theta^{(i-1)}) &= E(\log L(\Theta|Z)|X, \Theta^{(i-1)}) \\
&= \begin{cases} \int_{y \in Y} \log p(X, y|\Theta) p(y|X, \Theta^{(i-1)}) dy, & \text{连续;} \\ \sum_{y \in Y} \log p(X, y|\Theta) p(y|X, \Theta^{(i-1)}), & \text{离散.} \end{cases}
\end{aligned}
$$

Maximization Step

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)})$$

Mixture Models

1. If there is a reason to believe that a data set is comprised of several distinct populations, a mixture model can be used.
2. It has the following form:

$$p(x|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(x|\theta_j),$$

这里，$\sum_{j=1}^{M} \alpha_j = 1$, $\Theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M)$.

Mixture Models

假定观测到的数据为$X = \{x_1, x_2, \cdots, x_N\}$, $Y = \{y_1, y_2, \cdots, y_N\}$, 其中$y_i \in \{1, \cdots, M\}$ represents the source that generates the data. 也就是：

$$p(x|y = j, \Theta) = p_j(x|\theta_j), \quad p(y = j|\Theta) = \alpha_j.$$

Mixture Models

记$z_i = (x_i, y_i)$, 则： $p(z_i|\Theta) = p(x_i, y_i|\Theta) = p(y_i|x_i, \Theta)p(x_i|\Theta)$.

$$\begin{aligned}
& p(y_i|x_i, \Theta) \\
= \; & \frac{p(x_i, y_i, \Theta)}{p(x_i, \Theta)} = \frac{p(x_i|y_i, \Theta)p(y_i, \Theta)}{p(x_i, \Theta)} \\
= \; & \frac{p(x_i|y_i, \Theta)p(y_i|\Theta)p(\Theta)}{p(x_i|\Theta)p(\Theta)} = \frac{p(x_i|y_i, \Theta)p(y_i|\Theta)}{p(x_i|\Theta)} \\
= \; & \frac{p_{y_i}(x_i|\theta_{y_i})\alpha_{y_i}}{\sum_{j=1}^{M} \alpha_j p_j(x|\theta_j)}
\end{aligned}$$

Given x and Θ, the conditional density of y can be computed.

Complete-Data Likelihood Function

记$X=\{x_1,\ldots,x_N\}$, $y=\{y_1,\ldots,y_N\}$, $Z=\{z_1,\ldots,z_N\}$, $z_i=(x_i,y_i)$

$$
\begin{aligned}
& L(\Theta|Z) \\
= \quad & p(Z|\Theta)=p(X,y|\Theta)=p(X|y,\Theta)p(y|\Theta) \\
= \quad & \prod_{i=1}^{N} p(x_i|y_i,\Theta)p(y_i|\Theta)=\prod_{i=1}^{N} \alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})
\end{aligned}
$$

这里$\theta_{y_i}=(y_i,\Theta), \alpha_{y_i}=p(y_i|\Theta)$.

Expectation

$$\log L(\Theta|Z) = \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})\right]$$

$$\begin{aligned}
Q(\Theta, \Theta^g) &= E[\log L(\Theta|Z)|X, \Theta^g] \\
&= \sum_{y \in Y} \log L(\Theta|Z) p(y|X, \Theta^g) \\
&= \sum_{y \in Y} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})\right] \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \\
&= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})\right] \prod_{j=1}^{N} p(y_j|x_j, \Theta^g)
\end{aligned}$$

这里, Θ^g guess. $\sum\limits_{y \in Y} = \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M}$.

Expectation

$$
\begin{aligned}
& Q(\Theta, \Theta^g) \\
= & \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})\right] \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \\
= & \sum_{i=1}^{N} \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \log\left[\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i})\right] \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \\
= & \sum_{i=1}^{N} \Big(\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j=1, j\neq i}^{N} p(y_j|x_j, \Theta^g) \Big) \\
& \times \sum_{y_i=1}^{M} \log\left[\alpha_{y_i} p_{y_i}(x_i|\theta_{y_i}) p(y_i|x_i, \Theta^g)\right]
\end{aligned}
$$

$$
Q(\Theta, \Theta^g) = \sum_{y \in Y} \sum_{i=1}^{N} \sum_{l=1}^{M} \delta_{y_i,l} \log\left[\alpha_l p_l(x_i|\theta_l)\right] \prod_{j=1}^{N} p(y_j|x_j, \Theta^g)
$$

Expectation

$$
\begin{aligned}
& Q(\Theta, \Theta^g) \\
= & \sum_{i=1}^{N} \Big(\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j=1, j \neq i}^{N} p(y_j | x_j, \Theta^g) \Big) \\
& \times \sum_{y_i=1}^{M} \log \left[\alpha_{y_i} p_{y_i}(x_i | \theta_{y_i}) p(y_i | x_i, \Theta^g) \right] \\
= & \sum_{y_i=1}^{M} \sum_{i=1}^{N} \log \left[\alpha_{y_i} p_{y_i}(x_i | \theta_{y_i}) \right] \left[\prod_{\substack{j=1 \\ j \neq i}}^{N} \left(\sum_{y_j=1}^{M} p(y_j | x_j, \Theta^g) \right) \right] p(y_i | x_i, \Theta^g) \\
= & \sum_{l=1}^{M} \sum_{i=1}^{N} \log \left[\alpha_l p_l(x_i | \theta_l) \right] \left[\prod_{\substack{j=1 \\ j \neq i}}^{N} \left(\sum_{y_j=1}^{M} p(y_j | x_j, \Theta^g) \right) \right] p(l | x_i, \Theta^g)
\end{aligned}
$$

Expectation

$$
\begin{aligned}
& Q(\Theta, \Theta^g) \\
= \; & \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(x_i|\theta_l)\right] \sum_{y_1=1}^{M} \cdots \sum_{y_i=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{y_i,l} \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \\
= \; & \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(x_i|\theta_l)\right] \left(\sum_{y_1=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{\substack{j=1 \\ j\neq i}}^{N} p(y_j|X, \Theta^g) \right) p(\\
= \; & \sum_{l=1}^{M} \sum_{i=1}^{N} \log\left[\alpha_l p_l(x_i|\theta_l)\right] \left[\prod_{\substack{j=1 \\ j\neq i}}^{N} \left(\sum_{y_j=1}^{M} p(y_j|X, \Theta^g) \right) \right] p(l|x_i, \Theta^g)
\end{aligned}
$$

Expectation

注意到 $\sum_{y_j=1}^{M} p(y_j|x_j, \Theta^g) = 1$, 可得

$$\begin{aligned} Q(\Theta, \Theta^g) &= \sum_{l=1}^{M}\sum_{i=1}^{N} \log\left[\alpha_l p_l(x_i|\theta_l)\right] p(l|x_i, \Theta^g) \\ &= \sum_{l=1}^{M}\sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \sum_{l=1}^{M}\sum_{i=1}^{N} \log[p_l(x_i|\theta_l)] p(l|x_i, \Theta^g) \end{aligned}$$

Maximization

Given the initial guess Θ^g,

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \sum_{l=1}^{M} \sum_{i=1}^{N} \log[p_l(x_i|\theta_l)] p(l|x_i, \Theta^g)$$

We want to find Θ, to maximize the above expectation.

In fact, iteratively.

The GMM (Guassian Mixture Model)

Guassian model of a d-dimensional source, say j:

$$p_j(x|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right]$$

$$\theta_j = (\mu_j, \Sigma_j)$$

GMM with M sources:

$$p_j(x|\mu_1, \Sigma_1, \cdots, \mu_M, \Sigma_M) = \sum_{j=1}^{M} \alpha_j p_j(x|\mu_j, \Sigma_j), \quad \alpha_j \geq 0, \sum \alpha_j = 1.$$

Goal

Mixture Model

$$p(x|\Theta) = \sum_{l=1}^{M} \alpha_l p_l(x|\theta_l)$$

$\Theta = (\alpha_1, \ldots, \alpha_M, \theta_1, \ldots, \theta_M)$ subject to $\sum_{l=1}^{M} \alpha_l = 1$.

To maximize:

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \sum_{l=1}^{M} \sum_{i=1}^{N} \log[p_l(x_i|\theta_l)] p(l|x_i, \Theta^g)$$

Finding α_l

$$Q(\Theta, \Theta^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \sum_{l=1}^{M} \sum_{i=1}^{N} \log[p_l(x_i|\theta_l)] p(l|x_i, \Theta^g)$$

Due to the constraint on α_l’s, we introduce Lagrange Multiplier λ, and solve the following equation.

$$\frac{\partial}{\partial \alpha_l} \left[\sum_{l=1}^{M} \sum_{i=1}^{N} \log(\alpha_l) p(l|x_i, \Theta^g) + \lambda \left(\sum_{l=1}^{M} \alpha_l - 1 \right) \right] = 0, l = 1, \ldots, M$$

可得:

$$\sum_{i=1}^{N} \frac{1}{\alpha_l} p(l|x_i, \Theta^g) + \lambda = 0, \quad l = 1, \ldots, M$$

$$\sum_{i=1}^{N} p(l|x_i, \Theta^g) + \alpha_l \lambda = 0, \quad l = 1, \ldots, M$$

Finding α_l

可得

$$\sum_{l=1}^{M}\sum_{i=1}^{N} p(l|x_i,\Theta^g) + \lambda \sum_{l=1}^{M}\alpha_l = 0$$

由于 $\sum_{l=1}^{M} p(l|x_i,\Theta^g) = 1$, $\sum_{l=1}^{M}\alpha_l = 1$,

$$\lambda = -N$$

$$\alpha_l = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g)$$

$$p(l|x_i,\Theta^g) = \frac{\alpha_l^g p_l(x_i|\theta_l^g)}{\sum_{j=1}^{M}\alpha_j^g p_j(x|\theta_j^g)}$$

Finding θ_l

Consider GMM

$$p_j(x|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right]$$

$$\theta_j = (\mu_j, \Sigma_j)$$

$$\log[p_l(x|\mu_l, \Sigma_l)] = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_l|^{1/2} - \frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l)$$

Finding θ_l

Therefore, we want to maximize:

$$Q'(\Theta, \Theta^g) = \sum_{l=1}^{M} \sum_{i=1}^{N} \left(-\frac{1}{2} \log |\Sigma_l|^{1/2} - \frac{1}{2} (x_i - \mu_l)^T \Sigma_l^{-1} (x_i - \mu_l) \right) p(l|x_i, \Theta^g)$$

How? knowledge on matrix algebra is needed

$$\mu_l = \frac{\sum_{i=1}^{N} x_i p(l|x_i, \Theta^g)}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

$$\Sigma_l = \frac{\sum_{i=1}^{N} p(l|x_i, \Theta^g)(x_i - \mu_l)(x_i - \mu_l)^T}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

Summary

EM algorithm for GMM, Given an initial guess Θ^g, find Θ^{new} as follows

$$\alpha_l^{new} = \frac{1}{N}\sum_{i=1}^{N} p(l|x_i,\Theta^g)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l|x_i,\Theta^g)}{\sum_{i=1}^{N} p(l|x_i,\Theta^g)}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l|x_i,\Theta^g)(x_i-\mu_l^{new})(x_i-\mu_l^{new})^T}{\sum_{i=1}^{N} p(l|x_i,\Theta^g)},$$

$\Theta^g \leftarrow \Theta^{new}$.

Summary

求解μ_l^{new}和Σ_l^{new}的具体过程

我们的任务是最大化如下目标函数:

$$Q(\Theta,\Theta^g)=\sum_{l=1}^{M}\sum_{i=1}^{N}\log(\alpha_l)p(l|x_i,\Theta^g)+\sum_{l=1}^{M}\sum_{i=1}^{N}\log[p_l(x_i|\theta_l)]p(l|x_i,\Theta^g)$$

在GMM模型中，我们是对如下函数求极大值点:

$$\tilde{Q}=\sum_{l=1}^{M}\sum_{i=1}^{N}\left[-\frac{1}{2}\log|\Sigma_l|-\frac{1}{2}(x_i-\mu_l)^T\Sigma_l^{-1}(x_i-\mu_l)\right]p(l|x_i,\Theta^g)$$

Summary

求解μ_l^{new}

$$\frac{\partial \tilde{Q}}{\partial \mu_l} = \frac{\partial}{\partial \mu_l}\left[\sum_{i=1}^{N}\left(x_i^T \Sigma_l^{-1}\mu_l - \frac{1}{2}\mu_l^T \Sigma_l^{-1}\mu_l\right) p(l|x_i, \Theta^g)\right]$$

$$= \sum_{i=1}^{N}\left(\Sigma_l^{-1}x_i - \Sigma_l^{-1}\mu_l\right) p(l|x_i, \Theta^g) = 0$$

则

$$\mu_l^{new} = \hat{\mu}_l = \frac{\sum_{i=1}^{N} x_i p(l|x_i, \Theta^g)}{\sum_{i=1}^{N} p(l|x_i, \Theta^g)}$$

求解Σ_l^{new}

因为

$$-\frac{1}{2}\log|\Sigma_l| = \frac{1}{2}\log|\Sigma_l|^{-1} = \frac{1}{2}\log|\Sigma^{-1}|$$

所以

$$\tilde{Q} = \sum_{l=1}^{M}\sum_{i=1}^{N}\left[\frac{1}{2}\log|\Sigma_l^{-1}| - \frac{1}{2}(x_i-\mu_l)^T\Sigma_l^{-1}(x_i-\mu_l)\right]p(l|x_i,\Theta^g)$$

令$\Sigma_l^{-1} = (a_{st}^{(l)})_{d\times d}$，$\mu_l = (\mu_{l1},\mu_{l2},...,\mu_{ld})^T$，则

$$\tilde{Q} = \sum_{l=1}^{M}\sum_{i=1}^{N}\left[\frac{1}{2}\log|\Sigma_l^{-1}| - \frac{1}{2}\sum_{s=1}^{d}\sum_{t=1}^{d}a_{st}^{(l)}(x_{is}-\mu_{ls})(x_{it}-\mu_{lt})\right]p(l|x_i,\Theta^g)$$

$$= \sum_{l=1}^{M}\sum_{i=1}^{N}\left[\frac{1}{2}\log|\Sigma_l^{-1}| - \frac{1}{2}\sum_{s=1}^{d}\sum_{t=1}^{d}a_{st}^{(l)}(x_{is}-\hat{\mu}_{ls})(x_{it}-\hat{\mu}_{lt})\right]p(l|x_i,\Theta^g)$$