

# EM算法与GMM模型

布衣之莠

December 5, 2018

## 1 GMM模型:

### 1.1 问题引入:

已知N个人身高 $\{y_1, y_2, \dots, y_N\}$ ,要求解人群中身高的分布, 人群中有男有女, 男女符合不同分布, 但是事先我们并不知道某个人服从哪种分布。我们用混合高斯模型, 假设男女的分布各符合一个高斯分布, 用极大似然求解这两个分布的参数。

### 1.2 问题建模:

建立高斯混合模型, 即符合下式的概率分布模型:

$$P(y|\theta) = \sum_k \alpha_k \phi(y|\theta_k) \quad (1.2.1)$$

其中, $\alpha_k$ 是系数,  $\alpha_k \geq 0, \sum_k \alpha_k = 1, \phi(y|\theta_k)$ 是高斯分布密度,成为第k个分模型, 参数是 $\theta_k = (\mu_k, \sigma_k^2)$

$$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y - \mu_k)^2}{2\sigma_k^2}\right) \quad (1.2.2)$$

对于我们上面引入的问题, 显然 $k=1,2$ ,似然函数是:

$$L(\theta) = \sum_j \log p(y_j|\theta) = \sum_j \log \sum_k \alpha_k p(y_j|\theta_k) \quad (1.2.3)$$

以上似然函数是对和求对数, 很难优化, 所以我们采用EM算法去求解。

## 2 EM算法:

考虑下面似然函数的优化问题:

$$L(\theta) = \sum_j \log p(y_j|\theta) = \sum_j \log \sum_i p(y_j, z_i|\theta) \quad (2.1.1)$$

先介绍Jensen不等式。  
对于凸函数 $f(x)$ ,满足:

$$E(f(x)) \geq f(E(x)) \quad (2.1.2)$$

对于凹函数 $f(x)$ ,满足:

$$E(f(x)) \leq f(E(x)) \quad (2.1.3)$$

取等号是 $f(x) = C$ ,即是常数,Jensen不等式意义在于描述了函数值的期望与期望的函数值关系, 如果我们取 $f(x)$ 是 $\log$ , 我们看到上面式子(2.1.3)右边是对和求对数, 而左边是先内部求对数外部再求和, 利用这一点我们可以解决在问题1中遇到的优化难点:对和求对数。我们在(2.1.1)中引入 $Q_j(z_i)$ ,所以:

$$L(\theta) = \sum_j \log \sum_i Q_j(z_i) \frac{p(y_j, z_i|\theta)}{Q_j(z_i)} \geq \sum_j \sum_i Q_j(z_i) \log \frac{p(y_j, z_i|\theta)}{Q_j(z_i)} \quad (2.1.4)$$

上面引入的 $Q(z)$ 是关于 $z$ 的某种分布, 满足:

$$\sum_i Q_j(z_i) = 1 \quad (2.1.5)$$

(2.1.4)中不等号利用了Jensen不等式。如果可以取等号, 原来对和求对数的难题就解决了, 可以转化为先求对数在求和。等号成立, 由Jensen不等式,

$$\frac{p(y_j, z_i|\theta)}{Q_j(z_i)} = C \quad (2.1.6)$$

由(2.1.5),(2.1.6)可以得出:

$$\sum_i p(y_j, z_i|\theta) = C \quad (2.1.7)$$

所以,

$$Q_j(z_i) = \frac{p(y_j, z_i|\theta)}{C} = \frac{p(y_j, z_i|\theta)}{\sum_i p(y_j, z_i|\theta)} \quad (2.1.8)$$

(2.1.8)给出了我们引入的 $Q_j(z_i)$ 的表达式。

EM算法:

选取初始值 $\theta^0$ 初始化 $\theta, t=0$

Repeat {

E step:

$$Q_j^t(z_i) = \frac{p(y_j, z_i|\theta^t)}{\sum_i p(y_j, z_i|\theta^t)} = \frac{p(y_j, z_i|\theta^t)}{p(y_j|\theta^t)} = p(z_i|y_j, \theta^t) \quad (2.1.9)$$

M step:

$$\theta^{t+1} = \arg \max_{\theta} \sum_j \sum_i Q_j^t(z_i) \log \frac{p(y_j, z_i|\theta)}{Q_j^t(z_i)} \quad (2.1.10)$$

$$t = t + 1 \quad (2.1.11)$$

}直到收敛!

### 3 EM算法求解GMM模型:

把上面的EM算法应用于求解我们第一步提出的问题。

$$L(\theta) = \sum_j \log p(y_j|\theta) = \sum_j \log \sum_k \alpha_k p(y_j|\theta_k) = \sum_j \log \sum_k \gamma_{jk} \frac{\alpha_k p(y_j|\theta_k)}{\gamma_{jk}} \quad (3.1.1)$$

由Jensen不等式:

$$L(\theta) \geq \sum_j \sum_k \gamma_{jk} \log \frac{\alpha_k p(y_j|\theta_k)}{\gamma_{jk}} \quad (3.1.2)$$

利用EM算法可以知道当

$$\gamma_{jk} = \frac{p(y_j, z=k|\theta)}{\sum_k p(y_j, z=k|\theta)} = \frac{\alpha_k \phi(y_j|\theta_k)}{\sum_k \alpha_k \phi(y_j|\theta_k)} \quad (3.1.3)$$

(3.1.2)中等号成立。我们将高斯分布密度函数带入公式(3.1.2),有

$$L(\theta) = \sum_j \sum_k \gamma_{jk} \log \frac{p(y_j|\theta_k)}{\gamma_{jk}} = \sum_j \sum_k \gamma_{jk} \log \frac{\frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{(y-\mu_k)^2}{2\sigma_k^2})}{\gamma_{jk}} \quad (3.1.4)$$

化简:

$$L(\theta) = \sum_j \sum_k \gamma_{jk} [\log \alpha_k + \log(\frac{1}{\sqrt{2\pi}}) - \log \sigma_k - \frac{(y - \mu_k)^2}{2\sigma_k^2} - \log \gamma_{jk}] \quad (3.1.5)$$

在约束 $\sum_k \alpha_k = 1$ 下对(3.1.5)进行最小化, 可以利用lagrange乘子法求解此约束优化问题。令:

$$H(\theta) = L(\theta) + \lambda(\sum_k \alpha_k - 1)\Gamma \quad (3.1.6)$$

$$\frac{\partial H(\theta)}{\partial \alpha_k} = 0 \quad (3.1.7)$$

$$\frac{\partial H(\theta)}{\partial \lambda} = 0 \quad (3.1.8)$$

$$\frac{\partial H(\theta)}{\partial \mu_k} = 0 \quad (3.1.9)$$

$$\frac{\partial H(\theta)}{\partial \sigma_k} = 0 \quad (3.1.10)$$

可以推出公式:

$$\alpha_k = \frac{\sum_j \gamma_{jk}}{N} \quad (3.1.11)$$

$$\mu_k = \frac{\sum_j \gamma_{jk} y_j}{\sum_j \gamma_{jk}} \quad (3.1.12)$$

$$\sigma_k^2 = \frac{\sum_j \gamma_{jk} (y_j - \mu_k)^2}{\sum_j \gamma_{jk}} \quad (3.1.13)$$

所以训练GMM模型的算法步骤如下: 选取初始化值初始化 $\theta$

Repeat {

(1)

$$\gamma_{jk} = \frac{\alpha_k \phi(y_j | \theta_k)}{\sum_k \alpha_k \theta(y_j | \theta_k)} \quad (1)$$

(2) 根据 $\gamma_{jk}$ 和式(3.1.11)(3.1.12)(3.1.13)估计每个分模型的参数。

}

## 4 参考资料:

<https://www.cnblogs.com/mindpuzzle/archive/2013/04/05/2998746.html>

<https://www.cnblogs.com/mindpuzzle/archive/2013/04/24/3036447.html>

<https://www.cnblogs.com/tzg198955/p/4097543.html>