# 高斯混合模型 (GMM) 和 EM 算法

2018 年 10 月 7 日

## 0.1 高斯混合模型

高斯混合模型, 是由多个高斯分布混合形成的分布, 可以由下式表示:

$$P(X|\theta) = \sum_{k=1}^{K} w_k f(X|\theta_k)$$

其中 $\theta_k = (\mu_k, \sigma_k)$,$f(X|\theta_k) = N(\mu_k, \sigma_k)$.

一般, 一个简单的概率分布可以根据观测数据 X, 使用极大似然估计估算该分布概率的参数 $\theta$.[1]

$$\hat{\theta} = \arg\max L(\theta)$$

但是由于高斯混合模型含有隐变量 $w_k$, 所以使用极大似然估计没有解析解, 只能通过迭代方法求解.

**隐变量** 关于隐变量, 可以使用三硬币模型说明.[1]

其过程是: 先抛硬币甲,1) 如果硬币甲是正面 (事件 $z_1$), 则抛硬币乙, 并将乙的结果当作最终观测数据;2) 如果硬币甲是反面 (事件 $z_2$), 则抛硬币丙, 并将丙的结果当作最终观测数据.

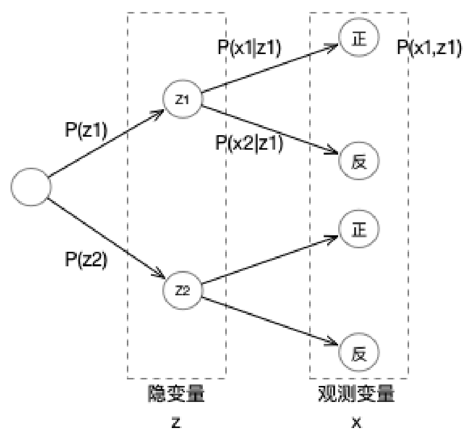


图 1: 三硬币模型

[1]思路是: 若采集到了观测数据 X, 则认定这组观测数据出现的概率在客观上是最大的; 客观上出现概率小的数据不容易被观测到. 这样, 可以假定一个概率分布模型, 以该分布模型的参数 $\theta$ 为变量, 进行最优化, 使得观测数据对应的概率最大, 则该 $\theta$, 为观测数据的极大似然估计 [2].

## 0.2 EM 算法

**算法流程**

1. 选取初值 $\theta^0$.
2. E 步:
$$Q(\theta,\theta^i)=\sum_z P(Z|X,\theta^i)\log P(X,Z|\theta)$$
3. M 步:
$$\theta^{i+1}=\arg\max Q(\theta,\theta^{i+1})$$

**原理及证明** 首先将似然函数 [2]$L(\theta)$ 改写成 $\theta$ 和 $z$ 的函数 $L(\theta,z)$.

似然函数为[2]
$$L(\theta)=\sum_i \log P(x^i|\theta)$$

$$L(\theta)=\sum_i \log P(x^i|\theta)=\sum_i \log\sum_j P(x^i,z^j|\theta) \tag{1}$$

将最右端的隐变量 $z$ 和观测变量 $x$ 的联合概率, 用隐变量 $z$ 的条件概率表示[3]

故
$$L(\theta,z)=\sum_i \log\sum_j P(z^j)\frac{P(x^i,z^j|\theta)}{P(z^j)} \tag{2}$$

**Jensen 不等式** [4]

如果 $f(x)$ 为凸函数, 则有 $E[f(x)]\geq f[E(X)]$.

简单证明:

对于任意点集 $x_i, i=1,..M$, 且 $\lambda_i\geq 0,\sum\limits_{i=1}^{M}\lambda_i=1$.
当 $M=2$ 时, 如图 2 所示, 有 $\lambda_1 f(x_1)+\lambda_2 f(x_2)\geq f(\lambda_1 x_1+\lambda_2 x_2)$
显而易见, 当 $M=\infty$ 时,

$$\sum_i \lambda_i f(x_i)\geq f(\sum_i \lambda_i x_i) \tag{3}$$

---

[2] $L(\theta)=\log\prod\limits_i P(x^i|\theta)=\sum\limits_i logP(x^i|\theta)$

[3] $P(x,z|\theta)=P(z)P(x|z,\theta)$

[4] 虽然表述为期望 $E[]$, 但理解成均值定理的推广更为合适, 如式 3.

即

$$E[f(x)] \geq f[E(X)]$$

当 $x_1 = x_2 = x_3 ... x_M$ 时, 等号成立. 即 $\sum_i \lambda_i f(x_i) = f(\sum_i \lambda_i x_i)$.

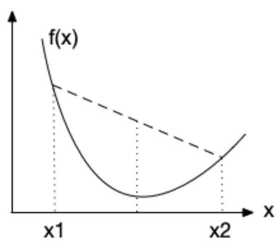


图 2: Jensen 不等式

接式 2,log 函数 (底 $> 1$ 时) 为凹函数. 有 $\sum_i \lambda_i \log(x_i) \leq \log(\sum_i \lambda_i x_i)$. 所以有:

$$L(\theta, z) = \sum_i \log \sum_j P(z^j) \frac{P(x^i, z^j|\theta)}{P(z^j)} \geq \sum_i \sum_j P(z^j) \log \frac{P(x^i, z^j|\theta)}{P(z^j)} \tag{4}$$

上式右端为 $L(\theta, z)$ 的下界. 这里通过极大化式 4 下界的方法求解极大化 $L(\theta, z)$.

**极大化似然函数下界** 当 $\frac{P(x^i, z^j|\theta)}{P(z^j)}$ 等于某常数时, 式 4 等号成立, 即下界最大. 此时有

$$\frac{P(x^i, z^j|\theta)}{P(z^j)} = c$$

$$\sum_j P(z^j) = 1$$

所以有

$$\sum_j P(x^i, z^j|\theta) = c$$

$$P(z^j) = \frac{P(x^i, z^j|\theta)}{c} = \frac{P(x^i, z^j|\theta)}{\sum_j P(x^i, z^j|\theta)} = P(z^j|x^i, \theta) \tag{5}$$

此时下界最大.

至此, 便实现利用 Jensen 不等式求出 $L(\theta, z)$ 的下界, 并求出令下界最大的 $P(z^j)$ 的取值:$P(z^j) = P(z^j|x^i, \theta) = \frac{P(x^i, z^j)}{\sum_j P(x^i, z^j)}$. 如图 3

由图 1所示, 此时已经将隐变量已不再是变量, 可以当作常量, 化简分布模型.

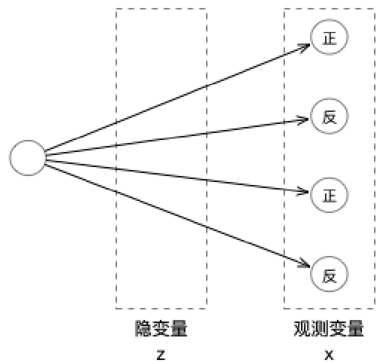


图 3: 简化三硬币模型

将式 5代回式 4, 有

$$L(\theta)=\sum_i\sum_j P(z^j|x^i,\theta)\log P(x^i,z^j|\theta)-\sum_i\sum_j P(z^j|x^i,\theta)\log P(z^j|x^i,\theta) \tag{6}$$

即

$$L(\theta)=\sum_i\sum_j P(z^j)\log P(x^i,z^j|\theta)=E_z[P(X,Z|\theta)]_{|X,\theta}=Q(\theta)$$

上式 6右端第二项, 是 $P(z^j|x^i,\theta)$ 的函数, 为常量, 在极大化 $L(\theta)$ 时可以忽略. 上式便是算法步骤中的 E 步.

至此, 便可以使用常规办法求解 $\hat{\theta}=\arg\max L(\theta)$. 这里便是算法步骤中的 M 步.

**总结** 先假定初值 $(z^0,\theta^0)$, 带入到 E 步公式求解出 $Q$ 函数.$Q$ 函数为调整 $z$ 能获得的最大值. 再对 $Q$ 函数求最大值, 调整 $\theta$. 反复迭代上述过程, 直至收敛, 获得 $(\hat{z},\hat{\theta})$.

# 参考文献

[1] 李航. *统计学习方法*. 清华大学出版社, 2012.

[2] 茆诗松. *概率论与数理统计第二版*. 高等教育出版社, 2011.