EM algorithm

2019年6月19日 14:23

EM算法主要解决问题：带有隐含变量的概率模型的极大似然估计.

有一组观测序列 $Y=(Y_1, Y_2, \cdots Y_n)$. 每个观测值对应一个隐藏值 $Z=(Z_1, Z_2, \cdots Z_n)$

进行极大似然估计 $P_\theta(Y)=\sum_Z P_\theta(Y,Z)=\sum_Z P_\theta(Y|Z)P_\theta(Z)$ 通过 $\theta^*=\max_\theta \log P_\theta(Y)$

例子．三硬币模型． 有A.B.C 3个硬币．先抛A．如果是反面．则抛B．如果是正面．则抛C.

抛n次．得到观测结果 1 1 0 1 0 0 1 0 1 1．求3个硬币正面概率

这里A硬币的结果相当于隐变量 Z．B.C硬币的结果相当于观测变量 Y.

分别设A.B.C抛正面的概率为 p_A, p_B, p_C．这里参数 $\theta=(p_A, p_B, p_C)$

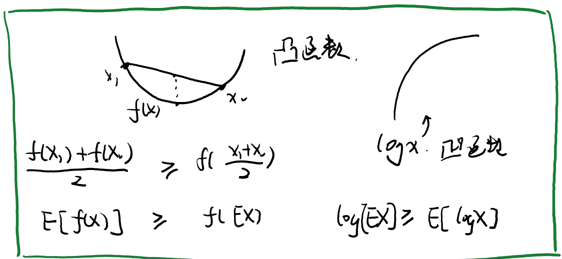
$$P_\theta(Y)=\prod_{i=1}^n P_\theta(Y^{(i)})=\prod_{i=1}^n \sum_{z^{(i)}} P_\theta(Y^{(i)}, z^{(i)})=\prod_{i=1}^n \sum_{z^{(i)}} P_\theta(Y^{(i)}|z^{(i)})P_\theta(z^{(i)})$$

$$=\prod_{i=1}^n P_\theta(z^{(i)}=0)\cdot P_\theta(y^{(i)}|z^{(i)}=0)+P_\theta(z^{(i)}=1)\cdot P_\theta(y^{(i)}|z^{(i)}=1)$$

$$=\prod_{i=1}^n \left[p_A\cdot p_B^{y^{(i)}}\cdot(1-p_B)^{1-y^{(i)}}+(1-p_A)\,p_C^{y^{(i)}}\cdot(1-p_C)^{1-y^{(i)}}\right]$$

然后把 $\theta^{(i)}$ 代入到上式中．通过 $\max_\theta \log P_\theta(Y)$．求得 θ^*．但这里参数都是离散变量．接下来推到一般情况.

$$\log P_\theta(Y)=\log\sum_Z P_\theta(Z)P_\theta(Y|Z)$$ ← log里面加和．用琴生不等式进行缩放.



$$=\log\sum_Z Q(Z)\frac{P_\theta(Z)P_\theta(Y|Z)}{Q(Z)}$$

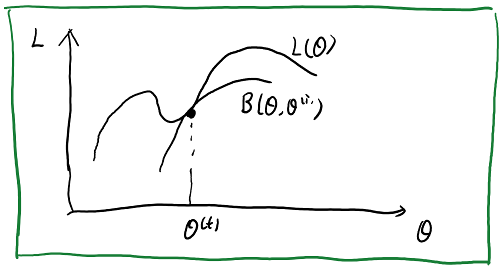
$$\geq \sum_Z Q(Z)\log\frac{P_\theta(Z)P_\theta(Y|Z)}{Q(Z)}$$ → $Q(Z)$ 可以是任意函数.

为了让这个下限尽量的 tight．我们在某个 θ 下．即 $\theta^{(t)}$ 下．让等号成立．则 $\log[\,]$ 为常数.

则 $\frac{P_{\theta^{(t)}}(Z)P_{\theta^{(t)}}(Y|Z)}{Q(Z)}=c$．又 $\sum_Z Q(Z)=1 \Rightarrow \frac{\sum_Z P_{\theta^{(t)}}(Z)P_{\theta^{(t)}}(Y|Z)}{c}=1 \Rightarrow c=P_{\theta^{(t)}}(Y)$

则 $Q(Z)=\frac{P_{\theta^{(t)}}(Z)P_{\theta^{(t)}}(Y|Z)}{P_{\theta^{(t)}}(Y)}=P_{\theta^{(t)}}(Z|Y)$.

即 $\log P_\theta(Y) \geq \sum_Z P_{\theta^{(t)}}(Z|Y)\log\frac{P_\theta(Z)P_\theta(Y|Z)}{P_{\theta^{(t)}}(Z|Y)}$ ← $B(\theta,\theta^{(t)})$



① 为了逐步逼近 $L(\theta)$ 的极值．可以任意取 $\theta^{(1)}$．然后求下限函数 $B(\theta,\theta^{(1)})$

② 然后对下限函数求极值．$\theta^{(2)}=\max_\theta B(\theta,\theta^{(1)})$ ← 这里对B求导比对L求导容易.

① 叫做E步． ② 叫做M步．整个过程称为EM算法

$$L(\theta^{(t)})=B(\theta^{(t)},\theta^{(t)}) \leq B(\theta^{(t+1)},\theta^{(t)}) \leq L(\theta^{(t+1)}) \Rightarrow L(\theta^{(t+1)}) \geq L(\theta^{(t)})$$ 所以可以收敛到最大值.

例子．高斯混合模型.

$Y=(y_1,\cdots y_n)$ 有K个 GMM. 取每个model的概率为 $(\alpha_1,\alpha_2,\cdots\alpha_k)$, $\sum_i\alpha_i=1$. $y_k=N(\mu_k,\sigma_k)$.

因此. $\theta=(\alpha_1,\alpha_2,\cdots\alpha_k,\ \mu_1,\mu_2,\cdots\mu_k,\ \sigma_1,\sigma_2,\cdots\sigma_k)$. 隐变量 $Z=(Z_1,\cdots Z_n)$ 表示抽到了哪个model

用 one-hot 表示则为 $Z_i=(Z_{i1},Z_{i2},\cdots Z_{ik})$. 其中只有1个为1,其余为0.

$$\begin{array}{cc}\begin{bmatrix} Z_1^{(1)}, & Z_2^{(1)}, & \cdots & Z_k^{(1)} \\ Z_1^{(2)} & Z_2^{(2)}, & \cdots & Z_k^{(2)} \\ \vdots & \vdots & & \vdots \\ Z_1^{(n)} & Z_2^{(n)}, & \cdots & Z_k^{(n)} \end{bmatrix} & \begin{matrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{matrix} \\ n_1 \quad n_2 \quad \cdots \quad n_k. & \end{array}$$

$$P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})=\frac{P_{\theta^{(t)}}(y^{(i)}|Z^{(i)}=k)P_{\theta^{(t)}}(Z^{(i)}=k)}{\sum_j P_{\theta^{(t)}}(y^{(i)}|Z^{(i)}=j)\,P_{\theta^{(t)}}(Z^{(i)}=j)}$$

$$\log P_\theta(Y)=\sum_{i=1}^n\log P_\theta(Y^{(i)})=\sum_{i=1}^n\log\sum_k P_\theta(Z^{(i)}=k)\cdot P_\theta(Y^{(i)}|Z^{(i)}=k)$$

$$\geq\sum_{i=1}^n\sum_k P_{\theta^{(t)}}(Z^{(i)}=k|Y^{(i)})\cdot\log\frac{P_\theta(Z^{(i)}=k)\,P_\theta(Y^{(i)}|Z^{(i)}=k)}{P_{\theta^{(t)}}(Z^{(i)}=k|Y^{(i)})}$$

$$=\sum_{i=1}^n\sum_k P_{\theta^{(t)}}(Z^{(i)}=k|Y^{(i)})\cdot\log\frac{\alpha_k\cdot\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y^{(i)}-\mu_k)^2}{2\sigma_k^2}\right)}{P_{\theta^{(t)}}(Z^{(i)}=k|Y^{(i)})}$$

$$B(\theta,\theta^{(t)})=\sum_{i=1}^n\sum_k P_{\theta^{(t)}}(Z^{(i)}=k|Y^{(i)})\left[\log\alpha_k-\log\sqrt{2\pi}-\log\sigma_k-\frac{(y^{(i)}-\mu_k)^2}{2\sigma_k^2}-\log P_{\theta^{(t)}}(Z^{(i)}=k|Y^{(i)})\right]$$

$$=\sum_{i=1}^n\left[\sum_k P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\log\alpha_k-P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\log\sigma_k-P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\frac{(y^{(i)}-\mu_k)^2}{2\sigma_k^2}\right]$$

s.t. $\sum_k\alpha_k=1$

$$\therefore\ \frac{\partial B(\theta,\theta^{(t)})}{\partial\alpha_k}=\frac{1}{\alpha_k}\cdot\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})-\gamma=0.\ \Rightarrow\ \alpha_k\gamma=\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})$$

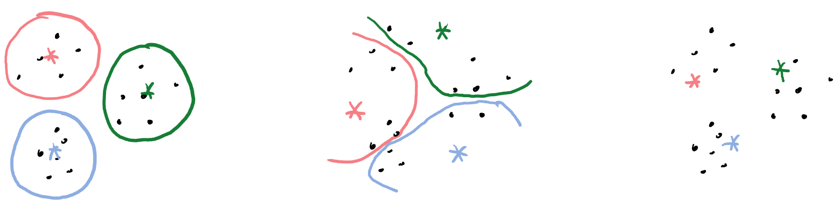
$$\therefore\ \gamma=\sum_{i=1}^N\sum_k P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\ \Rightarrow\ \alpha_k=\frac{\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})}{\sum_k\sum_{i=1}^N P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})}$$

$$\frac{\partial B(\theta,\theta^{(t)})}{\partial\mu_k}=\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\cdot 2(y^{(i)}-\mu_k)=0\ \Rightarrow\ \mu_k=\frac{\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\cdot y^{(i)}}{\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})}.$$

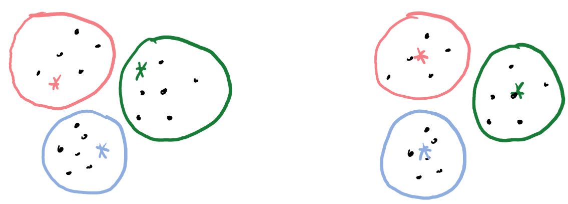
$$\frac{\partial B(\theta,\theta^{(t)})}{\partial\sigma_k}=\frac{1}{\sigma_k}\cdot-\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})+\frac{1}{\sigma_k^3}\cdot\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})\cdot(y^{(i)}-\mu_k)^2=0$$

$$\Rightarrow\ \sigma_k^2=\frac{\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})(y^{(i)}-\mu_k)^2}{\sum_{i=1}^n P_{\theta^{(t)}}(Z^{(i)}=k|y^{(i)})}$$

K-means



$E^{(1)}$ $M^{(1)}$



$E^{(2)}$ $M^{(2)}$

HMM.

$Y=(Y_1,Y_2,\cdots Y_n)$

$Z=(Z_1,Z_2,\cdots Z_n)$

$$\begin{matrix} Z_0\to & Z_1 & \to & Z_2 & \to\cdots\to & Z_n \\ & \downarrow & & \downarrow & & \downarrow \\ & Y_1 & & Y_2 & & Y_n \end{matrix}\qquad\begin{matrix} Z_1 & Z_2 & \cdots & Z_n \\ \downarrow & \downarrow & & \downarrow \\ Y_1 & Y_2 & & Y_n \end{matrix}$$

HMM.

$Y = (Y_1, Y_2, \cdots Y_n)$

$Z = (Z_1, Z_2, \cdots Z_n)$

$$\begin{array}{c|c}
\begin{tikzcd}
Z_0 \arrow[r] & Z_1 \arrow[r] \arrow[d] & Z_2 \arrow[r] \arrow[d] & \cdots \arrow[r] & Z_n \arrow[d] \\
 & Y_1 & Y_2 & & Y_n
\end{tikzcd}
&
\begin{tikzcd}
Z_1 \arrow[d] & Z_2 \arrow[d] & \cdots & Z_n \arrow[d] \\
Y_1 & Y_2 & & Y_n
\end{tikzcd}
\\
\text{HMM} & \text{GMM}
\end{array}$$

$$\log P_\theta(Y) = \log \sum_Z P_\theta P(Y, Z) = \log \sum_Z P_\theta(Y, Z) = \log \sum_{Z_1} P_\theta(Z_1 | Z_0) P_\theta(Y_1 | Z_1) \cdot \prod_{i=2}^{n} \sum_{Z_i} P_\theta(Z_i | Z_{i-1}) P_\theta(Y_i | Z_i).$$

log内求和. 用EM算法进行参数更新.