The EM algorithm

期望最大值算法

jensen's inequality

回忆

如果f(x)的二阶导数>=0，则f是凸函数；
当x是向量时，函数对应的海森矩阵(hessian)H>=0，是半正定矩阵

如果f(x)的二阶导数>0，则f是严格凸函数
当x是向量时，函数对应的海森矩阵(hessian)H>0，是正定矩阵

定理

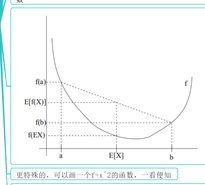
Theorem. Let f be a convex function, and let X be a random variable. Then:

$$\mathrm{E}[f(X)] \geq f(\mathrm{E}X).$$

Moreover, if f is strictly convex, then $\mathrm{E}[f(X)] = f(\mathrm{E}X)$ holds true if and only if $X = \mathrm{E}[X]$ with probability 1 (i.e., if X is a constant).

简而言之就是函数的期望>=期望的函数值

如果f是严格凸函数，则E[f(x)]=f(EX)当且仅当X=E[X]时成立，比如X取常数



更特殊的，可以画一个f=x^2的函数，一看便知

注：当-f是严格凸函数，也就是f是严格凹函数的时候，不等式变号

Mixture of Gaussians

MoG混合高斯模型

之前的高斯混合模型要拟合的参数有 ϕ，μ 和 Σ

E-step

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

这里面的 "$Q_i(z^{(i)}=j)$" 表示的是在分布 Q_i上 $z^{(i)}$ 取值 j 的概率

最大化关于参数 ϕ，μ，Σ 的似然函数的下界函数

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}
\end{aligned}$$

M-step

解 μ l

求偏导

$$\begin{aligned}
&\nabla_{\mu_l} \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}} \\
&= -\nabla_{\mu_l} \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j) \\
&= \frac{1}{2}\sum_{i=1}^{m} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l \\
&= \sum_{i=1}^{m} w_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right)
\end{aligned}$$

令上式为0

$$\mu_l := \frac{\sum_{i=1}^{m} w_l^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_l^{(i)}}$$

解 ϕ j

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j$$

其余式子都可以当成常数看待，一求导就没了

发现只需要最大化

然而还有附加的束 the ϕ_j's sum to 1，因为ϕ_j表示的是 $\phi_j = p(z^{(i)} = j; \phi)$

所以构造拉格朗日函数

$$\mathcal{L}(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta\left(\sum_{j=1}^{k} \phi_j - 1\right)$$

对拉格朗日函数求导

$$\frac{\partial}{\partial \phi_j}\mathcal{L}(\phi) = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + 1$$

令导数为0

$$\phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)}}{-\beta}$$

也就是 $\phi_j \propto \sum_{i=1}^{m} w_j^{(i)}$

对式子两边求和，得到 $-\beta = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} = \sum_{i=1}^{m} 1 = m$

这里使用了条件 $w_j^{(i)} = Q_i(z^{(i)} = j)$

所以最终：

$$\phi_j := \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}$$

The EM algorithm

假如有一个估计问题（estimation problem），训练样本集 {x(1), ..., x(m)} 包含了 m 个独立样本。用模型 p(x, z) 对数据建模，拟合参数（parameters），其中的似然函数

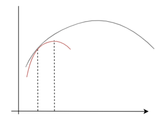
$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(x; \theta) \\
&= \sum_{i=1}^{m} \log \sum_{z} p(x, z; \theta)
\end{aligned}$$

第一步是对极大似然取对数，第二步是对每个样例的每个可能类别 z 求联合分布概率和。

（likelihood）为

因为z(i)是隐含随机变量，是未知的，所以解得 θ 很难，我们的策略是使用一种替代，找一个似然函数的下界函数。在E步找到一个下界函数，在M步对下界函数进行优化

图解



该算法的过程就是：初始化 θ 0，找到一个下界函数，再找到该下界函数的最大值点，设为 θ 1，再根据 θ 1找到新的下界函数，再找该下界函数的最大值点，设为 θ 2，循环反复，最终收敛到(θ)的局部最优值

所以我们设置Qi为p的某个分布，并且满足Qi(z)>=0, $\sum_z Q_i(z) = 1$
(如果z是连续的，则Qi是概率密度函数，需要将求和符号换成积分符号)

则有

$$\begin{aligned}
\sum_i \log p(x^{(i)}; \theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) &(1)\\
&= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} &(2)\\
&\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} &(3)
\end{aligned}$$

公式第三步解释

期望定义：E[g(z)]=∑p(z)g(z)

$\left[p(x^{(i)}, z^{(i)}; \theta)/Q_i(z^{(i)})\right]$ 基于z(i)的期望为 $\sum_{z^{(i)}} Q_i(z^{(i)}) \left[\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right]$

f(x)=logx是凹函数

利用Jensen不等式，就得到了 $f\left(\mathrm{E}_{z^{(i)}\sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right]\right) \geq \mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right)\right]$

其中上面的角标 “z(i) □Qi” 就表明这个期望是对于依据分布 Qi 来确定的 z(i) 的。

也就得到了第三步的式子

Qi的选择(E步骤)

那么Qi该如何选择呢？如果我们对 θ 有某种当前估计，那么将下界函数和 θ 联系起来是很自然的事。为了让下界和 θ 的特定值联系密切，我们有必要让上面的不等式的等号成立，Jensen不等式的时候我们说过，X=EX则等号成立，而X取常数可以使X=EX

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$

所以我们让

$$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta)$$

要实现这一条件，只需

由于我们已经知道Qi是一个概率分布，所以 $\sum_z Q_i(z^{(i)}) = 1$

$$\begin{aligned}
Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\
&= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\
&= p(z^{(i)} | x^{(i)}; \theta)
\end{aligned}$$

所以

因此，在给定 x(i) 和参数 θ 的设置下，我们可以简单地把 Qi 设置为 z(i) 的后验分布（posterior distribution）

至此，Qi已经选择完毕，也就是E步骤完毕，在接下来的M步骤中，需要最大化等式（3），然后得到参数 θ，循环往复

完整EM算法

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

怎么保证该算法收敛呢

设 $\theta^{(t)}$ 和 $\theta^{(t+1)}$ 是上面 EM 迭代过程中的某两个参数

接下来我们只需证明 $\ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)})$，这就表明 EM 迭代过程总是让似然函数对数（log-likelihood）单调递增

参数起点设置为 $\theta^{(t)}$，我们选择 $Q_i^{(t)}(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^{(t)})$

$$\ell(\theta^{(t)}) = \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})}$$

之前已经分析过了，这样选择Qi可以使不等式的等号成立，也就是

对该式最大化，即 $\ell(\theta^{(t+1)})$，

$$\begin{aligned}
\ell(\theta^{(t+1)}) &\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} &(4)\\
&\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} &(5)\\
&= \ell(\theta^{(t)}) &(6)
\end{aligned}$$

第一个不等式来自 $\ell(\theta) \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$

第二个不等式是因为 $\arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$ 来自式子

所以(4)式一定比(5)式大

第三个式子是因为我们前面就保证了Jensen不等式的等号成立了

因此，EM算法能够导致似然函数的单调收敛

$$J(Q, \theta) = \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

备注

如果我们定义

$$\ell(\theta) \geq J(Q, \theta)$$

在前面我们已经知道

EM算法也可以看成是对J(Q, θ)的坐标上升，E 步骤在 Q 上对 J 进行了最大化，M 步骤则在 θ 上对 J 进行最大化。