期望最大算法

Expectation Maximization Algorithm (EM)

① 算法导出

② 收敛性证明

③ 广义EM以及EM的变种

算法导出

模型无隐变量 $\Rightarrow$ MLE, MAP
模型含隐变量 $\Rightarrow$ EM

X: 观测变量, Z: 隐变量 $\Rightarrow$ (X, Z) 完全数据
θ: 参数
同样, 我们的目标是最大似然, (这里是观测变量的似然.)

$$\max \log P(X|\theta)$$

下面从两个角度出发, 导出EM的公式

ELBO + KL Divergence

$$\begin{aligned}\log P(X|\theta) &= \log P(X,Z|\theta) - \log P(Z|X,\theta) \\ &= \log \frac{P(X,Z|\theta)}{q(Z)} - \log \frac{P(Z|X,\theta)}{q(Z)}\end{aligned}$$

两边同时对 Z 求 $q(Z)$ 的期望

$$\text{左边} = \int_Z q(Z) \log(X|\theta)\, dZ = \log(X|\theta) \int_Z q(Z)\, dZ = \log(X|\theta)$$

$$\text{右边} = \underbrace{\int_Z q(Z) \log \frac{P(X,Z|\theta)}{q(Z)}\, dZ}_{\text{ELBO}} - \underbrace{\int_Z q(Z) \log \frac{P(Z|X,\theta)}{q(Z)}\, dZ}_{\text{KL Divergence}}.$$

ELBO (Evidence Lower Bound)

KL Divergence

用于恒量两个分布相似度的指标

定义两个分布：真实分布 $p(x)$，预测分布 $q(x)$.

则 $D_{KL}(p\|q) = \int p(x) \ln \frac{p(x)}{q(x)} dx$.

性质：① 非负性，$D_{KL} \geq 0$

② 仿射变换不变性，$y=ax+b$ $D_{KL}(p(x)\|q(x)) = D_{KL}(p(y)\|q(y))$

③ 非对称性. $D_{KL}(p\|q) \neq D_{KL}(q\|p)$

④ 值域 D_{KL} 在一定条件下才趋近于无穷.

则 $\log p(X|\theta) = ELBO + D_{KL}(q(z)\|p(z|X,\theta))$

$\log p(X|\theta) \geq ELBO$ （当且仅当 $q(z)=p(z|x,\theta)$ 时，等号成立）

对于 $\log p(X|\theta)$ 来说，其内包含有未观测的变量，所以往往直接 $\max \log p(X|\theta)$ 是不行的，得不到解析解. 自然地想到通过迭代的方式求解. 这一部分，我们不直接对 $\max \log p(X|\theta)$ 进行迭代，而是通过 ELBO 进行迭代.

此时，就有了一种想法，我们不去 $\max \log p(X|\theta)$，而是去 $\max ELBO$ 从而使得 $\log p(X|\theta)$ 取得更大的值.

当 $q(z)=p(z|x,\theta)$ 时，两者相等. 故取 $q(z)=p(z|x,\theta)$.

在上述推导中，$q(z)$ 是我们添加进来的分布，其无论为什么都不影响推导过程，故我们将 $q(z)$ 作为迭代的媒介. 令 $q(z)=p(z|x,\theta^{(i)})$，$\theta^{(i)}$ 表示第 i 步迭代得到的 θ.（这里为什么选用 $q(z)$ 作为媒介是我自己分析的）

$$\theta^{(i+1)} = \arg\max_{\theta} \text{ELBO}$$

$$= \arg\max_{\theta} \int_z p(z|x,\theta^{(i)}) \log \frac{p(x,z|\theta)}{p(z|x,\theta^{(i)})} \, dz.$$

$$= \arg\max_{\theta} \int_z p(z|x,\theta^{(i)}) \left(\log p(x,z|\theta) - \log p(z|x,\theta^{(i)}) \right) dz.$$

与θ无关.

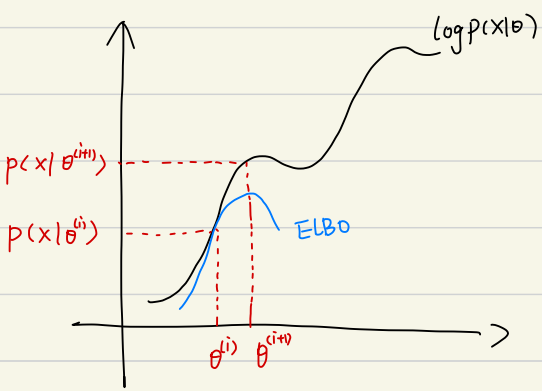
$$= \arg\max_{\theta} \int_z p(z|x,\theta^{(i)}) \log p(x,z|\theta) \, dz$$

$$= \arg\max_{\theta} E_{z\sim p(z|x,\theta^{(i)})} \left[\log p(x,z|\theta) \right]$$

这就是EM算法.

E步：求 $E_{z\sim p(z|x,\theta^{(i)})} \left[\log p(x,z|\theta) \right]$

M步：$\theta^{(i+1)} = \arg\max_{\theta} E_{z\sim p(z|x,\theta^{(i)})} \left[\log p(x,z|\theta) \right]$



左图形象表示了如何通过 max ELBO 去 max $\log p(x|\theta)$.

Jensen Inequality

这一部分用 Jensen 不等式导出 EM 算法.

和之前 ELBO 部分一样，我们需要 $\max \log p(X|\theta)$，这里我们直接 $\max \log p(X|\theta)$，而不通过 ELBO.

Jensen 不等式.

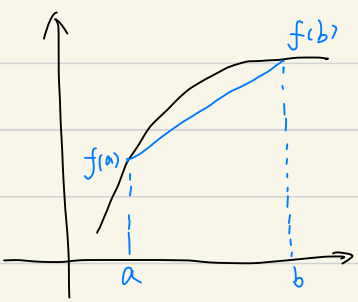
函数 $f(x)$ 是 concave function.

对 $\forall\, t \in [0,1]$,

$$f(ta+(1-t)b) \geq tf(a)+(1-t)f(b)$$

当 $t=\frac{1}{2}$ 时.

$$f\left(\underbrace{\frac{a+b}{2}}_{E}\right) \geq \underbrace{\frac{f(a)+f(b)}{2}}_{E} \Rightarrow f(E) \geq E[f]$$



对 $\log p(X|\theta) = \log \int_z p(X,z|\theta)\, dz$

$$= \log \int_z q(z) \frac{p(X,z|\theta)}{q(z)}\, dz$$

$$= \log E_{z\sim q(z)}\left[\frac{p(X,z|\theta)}{q(z)}\right]$$

$$\geq E_{z\sim q(z)}\left[\log \frac{p(X,z|\theta)}{q(z)}\right]$$ (当且仅当 $\frac{p(X,z|\theta)}{q(z)}=c$ 时，等式成立)

求 $q(z)$ 的表达式：

$$\frac{p(x,z|\theta)}{q(z)} = C$$

$$q(z) = \frac{1}{C} p(x,z|\theta)$$

两边对 z 积分：

$$\int_z q(z)dz = \frac{1}{C}\int_z p(x,z|\theta)\,dz$$

$$1 = \frac{1}{C} p(x|\theta)$$

$$C = p(x|\theta)$$

则 $q(z) = \frac{p(x,z|\theta)}{p(x|\theta)} = p(z|x,\theta)$.

则 $\theta^{(i+1)} = \arg\max_\theta E_{z\sim p(z|x,\theta^{(i)})}\left[\log \frac{p(x,z|\theta)}{p(z|x,\theta^{(i)})}\right]$

$\sim\sim\sim$ 与 θ 无关.

$$= \arg\max_\theta E_{z\sim p(z|x,\theta^{(i)})}\left[\log p(x,z|\theta)\right]$$

李航的《统计学习方法》第九章对 EM 的导出也是用 Jensen 不等式，但切入点和这里有点不同，可以详细地将第九章全部全一遍.

还有 PRML 中，EM 在第九章混合模型中有介绍，是在第4节，一般形式的 EM 算法，也可以看一下.

收敛性证明

EM公式：$\theta^{(i+1)} = \arg\max_{\theta} \int_z p(z|x,\theta^{(i)}) \log p(x,z)\, dz$.

该算法是迭代算法 $\theta^{(i)} \to \theta^{(i+1)}$

目标是 $\max \log p(x|\theta)$.

如果算法可行并收敛，那么就得 $\log p(x|\theta^{(i+1)}) \geq \log p(x|\theta^{(i)})$

且有上界.

显然，$p(x|\theta)$ 是一个概率，最大值为1，故 $\log p(x|\theta)$ 一定有上界.

所以只需证明 $p(x|\theta^{(i+1)}) \geq \log p(x|\theta^{(i)})$

$$\log p(x|\theta) = \log p(x,z|\theta) - \log p(z|x,\theta)$$

两边同时对 $p(z|x,\theta^{(i)})$ 求期望.

左边 $= \int_z p(z|x,\theta^{(i)}) \log p(x|\theta)\, dz = \log p(x|\theta)$

右边 $= \underbrace{\int_z p(z|x,\theta^{(i)}) \log p(x,z|\theta)\, dz}_{Q(\theta,\theta^{(i)})} - \underbrace{\int_z p(z|x,\theta^{(i)}) \log p(z|x,\theta)\, dz}_{H(\theta,\theta^{(i)})}$.

由于 $\theta^{(i+1)} = \arg\max_{\theta} \int_z p(z|x,\theta^{(i)}) \log p(x,z|\theta)\, dz$.

所以，显然，$Q(\theta^{(i+1)},\theta^{(i)}) \geq Q(\theta^{(i)},\theta^{(i)})$

所以，只需要证 $H(\theta^{(i+1)},\theta^{(i)}) \leq H(\theta^{(i)},\theta^{(i)})$ 就可以.

$$H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})$$ 这里也可以用Jensen不等式证（见《统计学习方法》）.

$$= \int_z p(z|x,\theta^{(i)}) \log \frac{p(z|x,\theta^{(i+1)})}{p(z|x,\theta^{(i)})} dz$$

$$= -D_{KL}(p(z|x,\theta^{(i)}) \| p(z|x,\theta^{(i+1)}))$$

$$\leq 0.$$

所以 $H(\theta^{(i+1)}, \theta^{(i)}) \leq H(\theta^{(i)}, \theta^{(i)})$

从而 $\log p(X|\theta^{(i+1)}) \geq \log p(X|\theta^{(i)})$

算法收敛性得到证明.

其实还需要证明最终得到的θ是 $L(\theta) = \log p(X|\theta)$ 的稳定点，但是这里李航的《统计学习方法》中只给出了文献，并没有证明.

广义EM以及EM的变种

EM主要是用来解决概率生成模型的参数估计问题

⇓

X: observed variable

Z: latent variable

(X, Z): complete data

θ: parameter.

生成模型就是假设数据是由隐变量生成的. $Z \longrightarrow X$. 给定Z一个对应的值 $p(z|\theta)$

$p(x, z|\theta) = p(z|\theta)p(x|z,\theta) \Rightarrow p(x|\theta) = \int_z p(x,z|\theta)\,dz$.

或 $p(x|\theta) = \frac{p(x,z|\theta)}{p(z|x,\theta)}$

从而 $\hat{\theta} = \arg\max p(x|\theta)$.

广义EM

狭义EM的局限性:

由前面推导可知: $\log p(x|\theta) = ELBO + D_{KL}(q\|p) \Longrightarrow \log p(x|\theta) \geq ELBO$

(p: 后验概率)

$$ELBO = \int_z q(z)\log\frac{p(x,z|\theta)}{q(z)}\,dz = \underbrace{\int_z q(z)\log p(x,z|\theta)\,dz}_{E_{z\sim q(z)}[\log p(x,z|\theta)]} - \underbrace{\int_z q(z)\log q(z)\,dz}_{\text{熵}: H[q(z)]}$$

$$D_{KL}(q\|p) = -\int_z q(z)\log\frac{p(z|x,\theta)}{q(z)}\,dz$$

对于狭义EM来说, 我们是令 $q(z) = p(z|x,\theta)$.

但是大多数时候, 后验概率 $p(z|x,\theta)$ 是intractable的. 求不出来. 显然我们不能简单地令 $q = p$. 对 q, 我们采用部分最优化, 而不是全局最优化

广义EM的导出.

$\log p(x|\theta) = ELBO + D_{KL}(q\|p)$.

我们通过上式子知：当$D_{KL}(q\|p)$越小时，ELBO越大.

所以我们可以将q的选择也作为迭代的一部分. 令 $ELBO = \mathcal{L}(q,\theta)$

$$\begin{cases} \text{固定}\theta,\ \hat{q} = \arg\min_q D_{KL}(q\|p) = \arg\max_q \mathcal{L}(q,\theta) \\ \text{固定}\hat{q},\ \theta = \arg\max_\theta \mathcal{L}(\hat{q},\theta) \end{cases}$$

对广义的EM为：

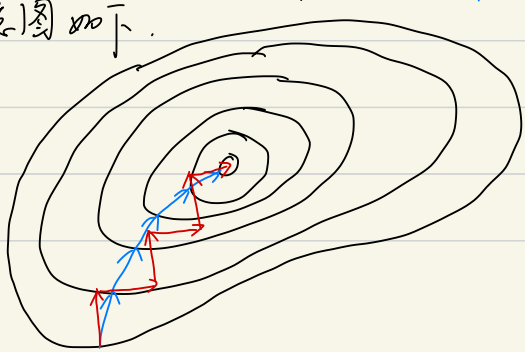
$$\begin{cases} E\text{-step}: & q^{(i+1)} = \arg\max \mathcal{L}(q,\theta^{(i)}) \\ M\text{-step}: & \theta^{(i+1)} = \arg\max \mathcal{L}(q^{(i+1)},\theta) \end{cases}$$

广义的EM有时也被称为F-MM

所以狭义的EM就是直接将 E-step 改为 $q^{(i+1)} = p(z|x,\theta^{(i)})$. 是广义EM的一种特化.

可以看见，广义的EM是对于两个目标q，θ，先固定θ求q. 再固定q求θ. 如此不断迭代. 类似于坐标上升法（SMO算法 -> 在svm中提了一嘴）

算法简单示意图如下.



→：梯度上升法

→：坐标上升法.

EM 的变种

variational Inference / variational Bayes

VBEM / VEM：用 VI/VB（变分推断）来计算 E-step.

MCEM：用 MC 采样方法计算 E-step.