机器学习白板推导系列笔记 (10.1~10.6)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 16 日

1 EM 算法背景

在之前, 对概率模型的参数估计采用极大似然估计法或最大后验概率估计即可, 因为模型只含有观测变量. 当模型既含有观测变量又含有潜在变量时, 就不能简单的使用这些方法. 不能够解析的求出对应的模型参数, 需要使用 EM 算法进行迭代的求解.

2 EM 算法的导出

假设观测变量为 X, 潜在变量为 Z, 要对概率模型进行参数估计, 目标是极大化观测数据对参数的对数似然函数.

$$\log P(X|\theta) \tag{1}$$

下面将从两个角度导出 EM 算法的迭代式.

2.1 KL 散度

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

上式两边分别对 $q(Z)$ 求期望

$$\begin{aligned}\text{左边} &= \int_Z q(Z)\log P(X|\theta)dZ \\ &= \log P(X|\theta)\end{aligned}$$

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

在信息论中, 关于两个分布 P 和 Q 的 KL 散度的定义是

$$KL(P||Q) = \int_x p(x)\log\frac{p(x)}{q(x)}dx \tag{2}$$

KL 散度恒大等于 0, 当且仅当二者同分布时等号成立.

$$\begin{aligned} -KL(P||Q) &= \int_x p(x) \log \frac{q(x)}{p(x)} dx \\ &= \mathbb{E}_{x\sim p(x)}[\log \frac{q(x)}{p(x)}] \end{aligned}$$

又根据杰森不等式, 凸函数有如下性质

$$f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)] \tag{3}$$

又因为 $\log x$ 是凸函数, 因此

$$\begin{aligned} -KL(P||Q) &= \mathbb{E}_{x\sim p(x)}[\log \frac{q(x)}{p(x)}] \\ &\leq \log \mathbb{E}_{x\sim p(x)}[\frac{q(x)}{p(x)}] \\ &= \log \int_x \frac{q(x)}{p(x)} p(x) dx \\ &= 0 \end{aligned}$$

因此

$$KL(P||Q) \geq 0 \tag{4}$$

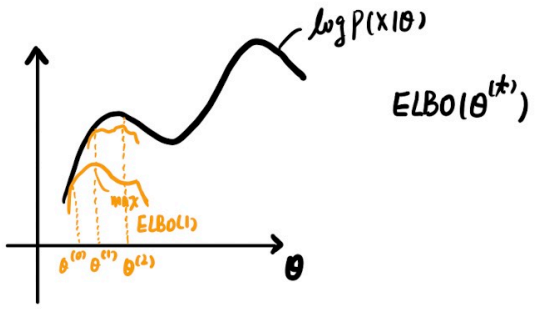
回到原式

$$\log P(X|\theta) = ELBO + KL散度 \tag{5}$$

因此可得

$$\log P(X|\theta) \geq ELBO \tag{6}$$

来想一下这个问题, 我们的最终目的是使得似然函数 $\log P(X|\theta)$ 尽可能大 (但不必担心它无穷, 因为概率是有界的), 下界是 ELBO, 那么就要想办法增大下界, 思路如下所示



如上图所示, 在在某个确定的 θ下的情况就是对数似然函数恒大等于 ELBO, ELBO 也有着对应的曲线 $ELBO(\theta)$, 如何增大下界, 一个很自然的想法就是取最大值, 即

$$\hat{\theta} = \arg\max_{\theta} ELBO(\theta) \tag{7}$$

一直迭代直到对数似然最大. 下面就遇到一个问题, 在确定 θ 下 ELBO 怎么求, 再来观察一下下式

$$ELBO = \int_Z q(Z) \log \frac{P(X, Z|\theta)}{q(Z)} dZ \tag{8}$$

$q(Z)$ 我们是不知道的, 那么可以考虑式 (6) 等号成立的情况, 就可以得出 $q(Z) = P(Z|X, \theta)$. 下面结合上图重新说明一下迭代的过程. 最终目的是使得对数似然函数最大, 以式 (6) 以及概率的约束为依据, 通过增大下界来实现. 对于某一时刻的参数 $\theta^{(t)}$ 对 ELBO 取极大值得到下一时刻的参数 $\theta^{(t+1)}$

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \int_Z q(Z) \log \frac{P(X, Z|\theta)}{q(Z)} dZ \\
&= \arg\max_{\theta} \int_Z P(Z|X, \theta^{(t)}) \log \frac{P(X, Z|\theta)}{P(Z|X, \theta^{(t)})} dZ \\
&= \arg\max_{\theta} \int_Z P(Z|X, \theta^{(t)}) \log P(X, Z|\theta) dZ \\
&= \arg\max_{\theta} \mathbb{E}_{P(Z|X,\theta^{(t)})}[\log P(X, Z|\theta)]
\end{aligned} \tag{9}$$

2.2 杰森不等式

对对数似然函数做适当的变换

$$\begin{aligned}
\log P(X|\theta) &= \log \int_Z P(X, Z|\theta) dZ \\
&= \log \int_Z \frac{P(X, Z|\theta)}{q(Z)} q(Z) dZ \\
&= \log \mathbb{E}_{q(Z)}[\frac{P(X, Z|\theta)}{q(Z)}] \\
&\geq \mathbb{E}_{q(Z)}[\log \frac{P(X, Z|\theta)}{q(Z)}]
\end{aligned} \tag{10}$$

这个其实就是上节的 ELBO, 当 $\frac{P(X,Z|\theta)}{q(Z)} = C$(常数) 时等号成立.

$$\begin{aligned}
q(Z) &= \frac{1}{C} P(X, Z|\theta) \\
\int_Z q(Z) dZ &= \int_Z \frac{1}{C} P(X, Z|\theta) dZ \\
1 &= \frac{1}{C} P(X|\theta) \\
C &= P(X|\theta)
\end{aligned}$$

代入原式

$$q(Z) = \frac{P(X, Z|\theta)}{P(X|\theta)} = P(Z|X, \theta) \tag{11}$$

通过不断优化 $\mathbb{E}_{q(Z)}[\log \frac{P(X,Z|\theta)}{q(Z)}]$ 使得对数似然函数不断上升, 于是

$$\theta^{(t+1)} = \arg\max_{\theta} \mathbb{E}_{P(Z|X,\theta^{(t)})}[\log \frac{P(X, Z|\theta)}{P(Z|X, \theta^{(t)})}] \tag{12}$$

3 EM 算法的收敛性

下面来证明一下 EM 算法的收敛性, 经过前面的描述已经得到参数迭代的式子

$$\theta^{(t+1)} = \arg\max_{\theta} \int_Z P(Z|X, \theta^{(t)}) \log P(X, Z|\theta) dZ \tag{13}$$

我们需要确定的是, 经过每次迭代后对数似然函数是否满足下式, 这样的迭代才是有效的

$$\log P(X|\theta^{(t+1)}) \geq \log P(X|\theta^{(t)}) \tag{14}$$

注. $\log P(X|\theta) \leq 0$

$$\log P(X|\theta) = \log P(X, Z|\theta) - \log P(Z|X, \theta)$$

等式两边分别求关于 $P(Z|X, \theta^{(t)})$ 的期望

$$\begin{aligned} \text{左边} &= \int_Z P(Z|X, \theta^{(t)}) \log P(X|\theta) dZ \\ &= \log P(X|\theta) \end{aligned}$$

$$\text{右边} = \underbrace{\int_Z P(Z|X, \theta^{(t)}) \log P(X, Z|\theta) dZ}_{Q(\theta, \theta^{(t)})} - \underbrace{\int_Z P(Z|X, \theta^{(t)}) \log P(Z|X, \theta) dZ}_{H(\theta, \theta^{(t)})}$$

$Q(\theta, \theta^{(t)})$ 即 ELBO 函数, 根据式 (13) 可知

$$Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta, \theta^{(t)})$$

$$Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)})$$

因此只需要证明$H(\theta^{(t+1)}, \theta^{(t)}) \leq H(\theta^{(t)}, \theta^{(t)})$即可, 二者做差得

$$\begin{aligned} H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) &= \int_Z P(Z|X, \theta^{(t)}) \log P(Z|X, \theta^{(t+1)}) dZ - \int_Z P(Z|X, \theta^{(t)}) \log P(Z|X, \theta^{(t)}) dZ \\ &= \int_Z P(Z|X, \theta^{(t)}) \log \frac{P(Z|X, \theta^{(t+1)})}{P(Z|X, \theta^{(t)})} dZ \\ &= -KL(P(Z|X, \theta^{(t)}) || P(Z|X, \theta^{(t+1)})) \leq 0 \end{aligned}$$

综上可得 $\log P(X|\theta^{(t+1)}) \geq \log P(X|\theta^{(t)})$ 成立, EM 算法是收敛的.

4 EM 算法流程

这里再对 EM 算法做一些回顾, 首先 EM 算法不是一个模型, 而是一种优化的方法, 如 SGD 也属于优化方法. EM 算法主要为了解决概率生成模型中的参数估计问题, 思想还是极大似然估计

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)$$

在知道 X 服从某种分布的情况下, 可以直接进行求解. 但如果 X 的分布未知, 那么需要引入一个归纳偏置(假设服从某个模型), 假设一个隐变量 Z 能够生成 X, 那么

$$P(X) = \int_Z P(X, Z) dZ$$

这里的解释有点不够直观, 到后面应用 EM 算法时会对上面的话有更深的理解.EM 算法的流程如下所示

Algorithm 1 EM 算法

Input: 观测变量数据 X, 隐变量数据 Z, 联合分布 $P(Y,Z|\theta)$, 条件分布 $P(Z|Y,\theta)$

Output: 模型参数 θ

1: 选择模型参数初始值 $\theta^{(0)}$, 开始迭代

2: E 步: 记 $\theta^{(t)}$ 为第 t 次迭代参数 θ 的估计值, 在第 $t+1$ 次迭代的 E 步, 计算

$$Q(\theta,\theta^{(t)}) = \int_Z P(Z|X,\theta^{(t)}) \log P(X,Z|\theta) dZ$$

3: M 步: 求使 $Q(\theta,\theta^{(t)})$ 极大化的 θ, 确定第 $t+1$ 次迭代的参数估计值 $\theta^{(t+1)}$

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta,\theta^{(t)})$$

4: 重复步骤 2, 3 直至收敛.

5 广义 EM

为了解决数据 X 分布未知而无法使用极大似然估计对参数直接进行估计的情况, 假设了一个隐变量, 可以引入一些潜在归纳将问题具体化, 然后使用 EM 算法迭代出参数, 在 EM 算法时, 根据之前的讨论我们注意到, 其中很重要的一部是假设 $q(Z) = P(Z|X,\theta)$, 需要求解出 $P(Z|X,\theta^{(t)})$, 在一些简单的模型中实现这一要求比较容易, 但是在复杂模型中, 它往往是求不出来的, 那该怎么办呢.

再重新观察一下对数似然函数的形式

$$\log P(X|\theta) = ELBO + KL散度 \tag{15}$$

$$ELBO = \int_Z q(Z) \log \frac{P(X,Z|\theta)}{q(Z)} dZ \tag{16}$$

$$KL散度 = \int_Z q(Z) \log \frac{q(Z)}{P(Z|X,\theta)} dZ \tag{17}$$

将 ELBO 记为 $\mathcal{L}(q,\theta)$ 那么

$$\log P(X|\theta) = ELBO + KL散度 \geq \mathcal{L}(q,\theta)$$

因为后验 $P(Z|X,\theta)$ 很难求, 所以考虑这么一个问题, 如果θ 固定, 那么上式左边是定值, ELBO 越大, KL 散度就越小, $q(Z)$ 也就越接近 $P(Z|X,\theta)$, 即然求不出来那就使劲靠近, 于是得到一个最优化问题

$$\hat{q}(Z) = \arg\min_q KL(q||P) = \arg\max_q \mathcal{L}(q,\theta) \tag{18}$$

当得到 $\hat{q}$ 后, 用这个值求 θ

$$\hat{\theta} = \arg\max_\theta \mathcal{L}(\hat{q},\theta) \tag{19}$$

于是就得到了广义的 EM 算法

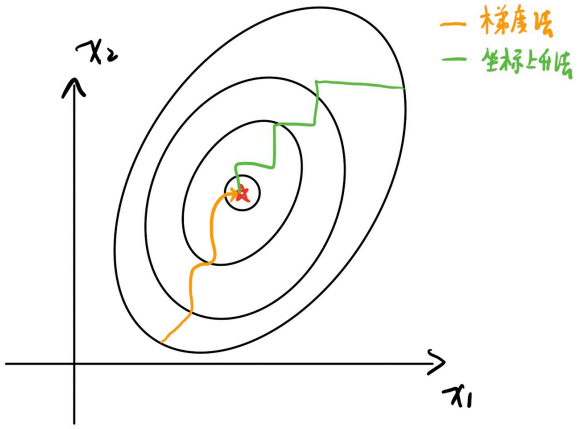
- E-step $q^{(t+1)} = \arg\max_q \mathcal{L}(q,\theta^{(t)})$
- M-step $\theta^{(t+1)} = \arg\max_\theta \mathcal{L}(q^{(t+1)},\theta)$

对 ELBO 可以进一步进行化简

$$\begin{aligned}\mathcal{L}(q,\theta) &= \int_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}dZ \\ &= \int_Z q(Z)\log P(X,Z|\theta)dZ - \int_Z q(Z)\log q(Z)dZ \\ &= \mathbb{E}_{q(Z)}[\log P(X,Z|\theta)] + H[Z]\end{aligned} \tag{20}$$

6 EM 的变种

上节提到的广义 EM 因为两步都是取 max 操作, 因此也称为 MM 算法. 这两步采用的方法是坐标上升法, 下图粗略展示了该方法与梯度法的区别



此外, 广义 EM 的两个优化步骤可以互换, 但其实每一步优化都不容易. 如果 E-Step 采用基于平均场的 VI, 则称为 VBEM/VEM. 也可以采用蒙特卡洛采样法求后验, 称为 MCEM, 这些会在之后介绍. 另外 EM 算法使用梯度法进行优化也会在之后讲解.