EM算法

杨航锋

EM 算法是一种迭代式的算法，用于含有隐变量的概率参数模型的最大似然估计或极大后验概率估计 。 EM 算法解决这个的思路是使用启发式的迭代方法，既然无法直接求出概率模型分布参数，那么是否可以先猜想隐含数据（ EM 算法的 E 步）？ 基于这个思路，对已知数据和猜测的隐含数据一起来极大化对数似然，最后求解概率模型的参数（ EM 算法的 M 步)。由于之前的隐含数据是猜测的，所以此时得到的模型参数一般还不是准确的结果。不过没有关系，基于当前得到的模型参数，继续猜测隐含数据，然后继续极大化对数似然，求解概率模型参数。以此类推，不断的迭代下去，直到概率模型的分布参数基本无变化，算法收敛，即找到了合适的模型参数。 更多文章见GitHub地址

EM算法的推导

极大似然估计的缺陷

假设输入空间是 $X \in \mathbb{R}^n$ 含有m个样本数据$(x^{(1)}, x^{(2)}, \cdots, x^{(m)})$，且这$m$个样本数据服从概率密度函数为 $p(x;\theta)$ 的分布，其中 $x^{(i)} \in X$ 、 θ 为未知参数。通过极大似然估计求解概率模型的未知参数 θ 的过程是

$$\begin{aligned} &\arg\max_{\theta} \ L(\theta) = \prod_{i=1}^{m} p(x^{(i)};\theta) \\ \Leftrightarrow &\arg\max_{\theta} \ \mathcal{L}(\theta) = \sum_{i=1}^{m} \log p(x^{(i)};\theta) \end{aligned}$$

故 $\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{m} \log p(x^{(i)};\theta)$ ，当已知每一个样本数据 $x^{(i)}$ 都对应一个类别变量 $z^{(i)}$ 时，即 $z = (z^{(1)}, z^{(2)}, \ldots, z^{(m)})$ ，此时的极大化模型的对数似然函数可以通过全概率公式展开为

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

因为含有隐变量 z 故极大似然估计并不能够求解上述模型。

詹森不等式

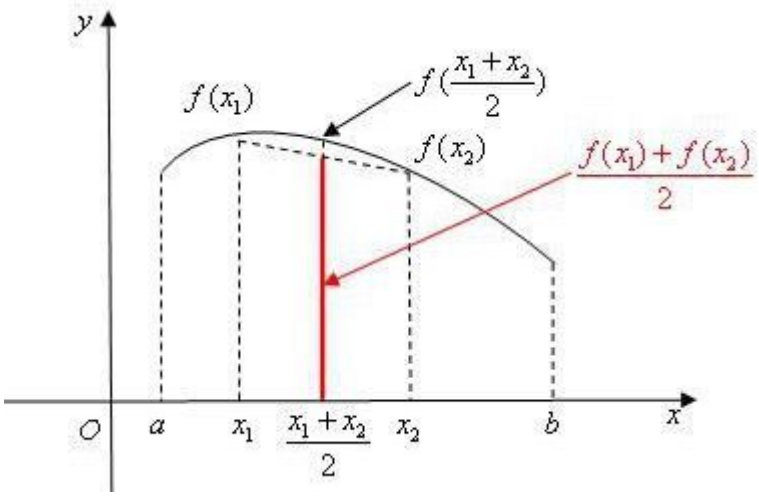
当函数为凸函数时， $f(x)$ 函数的期望大于等于期望的函数，即 $\mathbb{E}[f(x)] \geq f(\mathbb{E}[\mathrm{x}])$ 或者写成凸函数条件表达式的形式， $tf(x_1)+(1-t)f(x_2) \geq f(tx_1+(1-t)x_2)$ 其中 $t \in [0,1]$ ，凹函数相反。

X	x_1	x_2
p	t	$1-t$

根据 X 的分布列，由期望和函数期望的定义可知

$$\mathbb{E}[f(x)] = tf(x_1)+(1-t)f(x_2)$$
$$\mathbb{E}[x] = tx_1+(1-t)x_2$$

从几何上直观的解释为



求解含有隐变量的概率模型

为了求解含有隐变量 z 的概率模型 $\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$ 需要一些特殊的技巧，通过引入隐变量 $z^{(i)}$ 的概率分布为 $Q_i(z^{(i)})$ ，因为 $\log(x)$ 是凹函数故结合凹函数形式下的詹森不等式进行放缩处理

$$\begin{aligned}
\sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) &= \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \\
&= \sum_{i=1}^{m} \log \mathbb{E}\left(\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right) \\
&\geq \sum_{i=1}^{m} \mathbb{E}\left[\log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right] \\
&= \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}
\end{aligned}$$

其中由概率分布的充要条件 $\sum\limits_{z^{(i)}} Q_i(z^{(i)}) = 1$、$Q_i(z^{(i)}) \geq 0$ 可看成下述关于 z 函数分布列的形式

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

这个过程可以看作是对 $\mathcal{L}(\theta)$ 求了下界，假设 θ 已经给定那么 $\mathcal{L}(\theta)$ 的值就取决于 $Q_i(z^{(i)})$ 和 $p(x^{(i)}, z^{(i)})$ 了，因此可以通过调整这两个概率使下界不断上升，以逼近 $\mathcal{L}(\theta)$ 的真实值，当不等式变成等式时说明调整后的概率能够等价于 $\mathcal{L}(\theta)$ ，所以必须找到使得等式成立的条件，即寻找

$$\mathbb{E}[\log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})})] = \log \mathbb{E}[\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}]$$

由期望得性质可知当

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = C, \quad C \in \mathbb{R} \quad (*)$$

等式成立，对上述等式进行变形处理可得

$$\begin{aligned} & p(x^{(i)}, z^{(i)}; \theta) = CQ_i(z^{(i)}) \\ \Leftrightarrow & \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) = C \sum_{z^{(i)}} Q_i(z^{(i)}) = C \\ \Leftrightarrow & \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) = C \qquad (**) \end{aligned}$$

把 $(**)$ 式带入 $(*)$ 化简可知

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

至此，可以推出在固定参数 θ 后， $Q_i(z^{(i)})$ 的计算公式就是后验概率，解决了 $Q_i(z^{(i)})$ 如何选择得问题。这一步称为 E 步，建立 $\mathcal{L}(\theta)$ 得下界；接下来得 M 步，就是在给定 $Q_i(z^{(i)})$ 后，调整 θ 去极大化 $\mathcal{L}(\theta)$ 的下界即

$$
\begin{aligned}
&\arg\max_{\theta} \sum_{i=1}^{m} \log p(x^{(i)};\theta) \\
\Leftrightarrow &\arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} \\
\Leftrightarrow &\arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \left[\log p(x^{(i)},z^{(i)};\theta) - \log Q_i(z^{(i)})\right] \\
\Leftrightarrow &\arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log p(x^{(i)},z^{(i)};\theta)
\end{aligned}
$$

因此EM算法的迭代形式为

Repeats until it converges{

E step：for every $x^{(i)}$ calculate

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

M step：update θ

$$\theta := \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log p(x^{(i)},z^{(i)};\theta)$$

}

EM算法的收敛性

不妨假设 $\theta^{(k)}$ 和 $\theta^{(k+1)}$ 是 EM 算法第 k 次迭代和第 $k+1$ 次迭代的结果，要确保 EM 算法收敛那么等价于证明 $\mathcal{L}(\theta^{(k)}) \leq \mathcal{L}(\theta^{(k+1)})$ 也就是说极大似然估计单调增加，那么算法最终会迭代到极大似然估计的最大值。在选定 $\theta^{(k)}$ 后可以得到E步 $Q_i^{(k)}(z^{(i)}) = p(z^{(i)}|x^{(i)};\theta^{(k)})$ ，这一步保证了在给定 $\theta^{(k)}$ 时，詹森不等式中的等式成立即

$$\mathcal{L}(\theta^{(k)}) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i^{(k)}(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta^{(k)})}{Q_i(z^{(i)})}$$

然后再进行 M 步，固定 $Q_i^{(k)}(z^{(i)})$ 并将 $\theta^{(k)}$ 视作变量，对上式 $\mathcal{L}(\theta^{(k)})$ 求导后得到 $\theta^{(k+1)}$ 因此有如下式子成立

$$
\begin{aligned}
\mathcal{L}(\theta^{(k)}) &= \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i^{(k)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(k)})}{Q_i(z^{(i)})} && (a) \\
&\leq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i^{(k)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(k)})}{Q_i(z^{(i)})} && (b) \\
&\leq \mathcal{L}(\theta^{(k+1)}) && (c)
\end{aligned}
$$

首先(a)式是前面 E 步所保证詹森不等式中的等式成立的条件，(a)到(b)是M步的定义，(b)到(c)对任意参数都成立，而其等式的条件是固定 θ 并调整好 Q 时成立，(b)到(c)只是固定 Q 调整 θ，在得到 $\theta^{(k+1)}$ 时，只是最大化 $\mathcal{L}(\theta^{(k)})$，也就是 $\mathcal{L}(\theta^{(k+1)})$ 的一个下界而没有使等式成立。

总结

EM 算法一句话总结就是： E 步固定 θ 优化 Q， M 步固定 Q 优化 θ。