最大期望算法[1]

简介

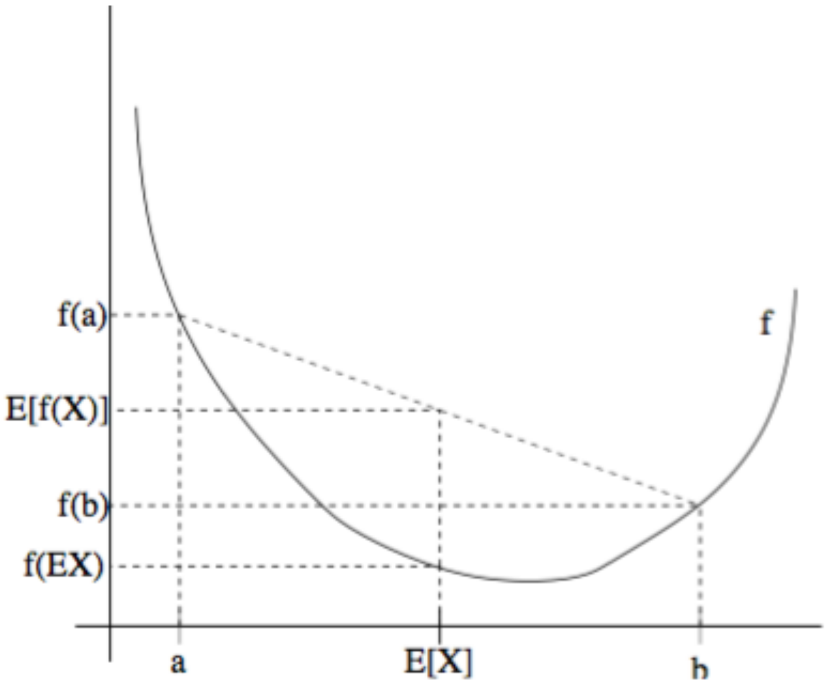
本notes将介绍最大期望算法（Exception-Maximization，EM）的推导过程。

1 Jensen不等式

定理1.1（Jensen不等式） 若f是一个凸函数[2]，且X是一个随机变量，那么有

$$E[f(X)] \geq f(E[X])$$

而且，当且仅当$X = E[X]$概率为1时（即X为一个常数时），等式成立。



话不多说，上图很好地展示了Jensen不等式。而且很容易得到，若f是一个凹函数（即$-f$是一个凸函数），那么有$E[f(X)] \leq f(E[X])$。

2 EM算法

给定训练集$\left\{x^{(1)}, \cdots, x^{(m)}\right\}$，其中包含m个未标记样本。我们希望对$p(x, z; \theta)$建模，但我们只能观察到样本的输入$x$。根据所有已知条件，构造log-likelihood函数

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

直接求出θ的最大似然估计值是很困难的，~~如果可以观察到输入x所对应的输出z，这个过程就会变得简单。~~鉴于此，EM算法提出了一种有效方法：直接最大化$l(\theta)$很困难，我们的策略改为不断地构建l的下界（E-step），然后再优化该下界（M-step）。

对任意$i \in \{1, \cdots, m\}$，令Q_i表示z的某一分布（即$\sum_z Q_i(z) = 1, Q_i(z) \geq 0$），那么有

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

上式的推导过程用到了Jensen不等式，由于$f(x) = \log x$对任意$x \in \mathbb{R}^+$有$f''(x) = -1/x^2 < 0$，所以log函数是严格凹函数，所以得到了式(2)。至于式(1)和式(3)，它们都是求期望的公式，无需多言。

我们得到了l的下界式(3)，那么我们该如何选择Q_i呢？显然，我们希望在θ处得到l的一个紧邻的下界，那么根据Jensen不等式等号成立的条件知

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

其中c为任意常量，上式也可以表示为

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

又由于$\sum_z Q_i\left(z^{(i)}\right) = 1$，所以

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

至此，我们已经选择出了合适的Q_i来构造我们努力最大化的l的下界，这就是E-step。在接下来的M-step中，我们需要对关于θ的式(3)最大化以重新估计θ的值。重复这两步我们就得到了EM算法，即

- 重复以下过程直至收敛
    - （E-step）For each i, set

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

    - （M-step）Set

$$\theta:=\arg\max_{\theta}\sum_i\sum_{z^{(i)}}Q_i\left(z^{(i)}\right)\log\frac{p\left(x^{(i)},z^{(i)};\theta\right)}{Q_i\left(z^{(i)}\right)}\tag{4}$$

我们该如何证明以上算法是收敛的？假设$\theta^{(t)}$和$\theta^{(t+1)}$是EM算法中连续两次迭代过程的参数，只需要证明$l(\theta^{(t)})\le l(\theta^{(t+1)})$就说明算法的每一步迭代都使得log-likelihood单调增大，这样不断逼近最大值最终收敛。那么在第t次迭代中，我们应该选择$Q_i^{(t)}:=p\left(z^{(i)}|x^{(i)};\theta^{(t)}\right)$，带入到式(3)中得

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

上式右边取参数$\theta^{(t+1)}$时值最大，因此有

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

式(5)来自于式(3)，具体地，我们选择$Q_i=Q_i^{(t)},\theta=\theta^{(t+1)}$；式(6)来自于式(3)，具体地，我们选择$\theta=\theta^{(t)}$并更新为$\theta^{(t+1)}$。当$l(\theta)$增长（即$l(\theta^{(t+1)})-l(\theta^{(t)})$）小于某一阀值时我们就可以认为算法收敛。

此外，如果我们定义

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

我们可以将EM算法视作对J的坐标上升算法：在E-step中，我们对Q进行最大化；在M-step中，我们对θ进行最大化。

3 混合高斯模型回顾

得到EM算法的一般形式后，让我们回到混合高斯模型中来观察参数ϕ,μ和Σ的具体拟合过程。

在E-step中我们只是简单地计算

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

接下来，在M-step中我们需要计算ϕ, μ, Σ的最大似然估计，此时

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

令上式对μ_l的偏导等于0可得

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

同理解得Σ（求偏导过程式子太复杂，不给出），对于ϕ，我们只需要最大化

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

由于限制条件$\sum_{j=1}^{k} \phi_j = 1$，我们目标变成最大化拉格朗日函数

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

同样地，令上式对ϕ_j的偏导数等于0可得

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

在限制条件下，我们有$-\beta = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} = \sum_{i=1}^{m} 1 = m$，因此在M-step中

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

---

1. Written by [Jimmy](#) on 2016/03/12.[↩](#)
2. 若f是定义在实数域上的函数，且$f''(x) \geq 0$（for all $x \in \mathbb{R}$），那我们就称f为凸函数；若f是自变量为向量的函数，且它的Hessian矩阵半正定（$H \geq 0$），那么我们称f为凸函数。当这两个命题等号不成立时，f是严格凸函数。[↩](#)