（EM算法）The EM Algorithm

EM是我一直想深入学习的算法之一，第一次听说是在NLP课中的HMM那一节，为了解决HMM的参数估计问题，使用了EM算法。在之后的MT中的词对齐中也用到了。在Mitchell的书中也提到EM可以用于贝叶斯网络中。

下面主要介绍EM的整个推导过程。

1. Jensen不等式

回顾优化理论中的一些概念。设f是定义域为实数的函数，如果对于所有的实数x，$f''(x) \geq 0$，那么f是凸函数。当x是向量时，如果其hessian矩阵H是半正定的（$H \geq 0$），那么f是凸函数。如果$f''(x) > 0$或者$H > 0$，那么称f是严格凸函数。

Jensen不等式表述如下：

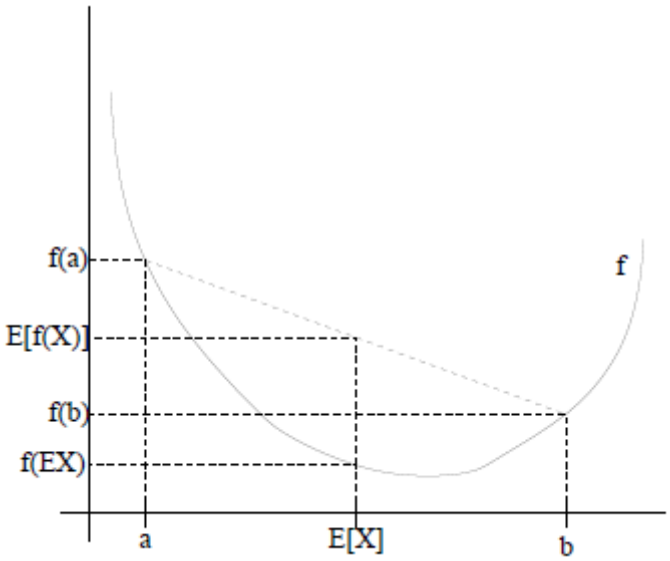
如果f是凸函数，X是随机变量，那么

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

特别地，如果f是严格凸函数，那么$E[f(X)] = f(EX)$当且仅当$p(x = E[X]) = 1$，也就是说X是常量。

这里我们将$f(E[X])$简写为$f(EX)$。

如果用图表示会很清晰：



图中，实线f是凸函数，X是随机变量，有0.5的概率是a，有0.5的概率是b。（就像掷硬币一样）。X的期望值就是a和b的中值了，图中可以看到$E[f(X)] \geq f(EX)$成立。

当f是（严格）凹函数当且仅当-f是（严格）凸函数。

Jensen不等式应用于凹函数时，不等号方向反向，也就是$E[f(X)] \leq f(EX)$。

2. EM算法

给定的训练样本是$\{x^{(1)}, \ldots, x^{(m)}\}$，样例间独立，我们想找到每个样例隐含的类别z，能使得p(x,z)最大。p(x,z)的最大似然估计如下：

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

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。但是直接求θ一般比较困难，因为有隐藏变量z存在，但是一般确定了z后，求解就容易了。

EM是一种解决存在隐含变量优化问题的有效方法。竟然不能直接最大化$\ell(\theta)$，我们可以不断地建立ℓ的下界（E步），然后优化下界（M步）。这句话比较抽象，看下面的。

对于每一个样例i，让Q_i表示该样例隐含变量z的某种分布，Q_i满足的条件是$\sum_z Q_i(z) = 1, Q_i(z) \geq 0$。（如果z是连续性的，那么$Q_i$是概率密度函数，需要将求和符号换做积分符号）。比如要将班上学生聚类，假设隐藏变量z是身高，那么就是连续的高斯分布。如果按照隐藏变量是男女，那么就是伯努利分布了。

可以由前面阐述的内容得到下面的公式：

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

（1）到（2）比较直接，就是分子分母同乘以一个相等的函数。（2）到（3）利用了Jensen不等式，考虑到$\log(x)$是凹函数（二阶导数小于0），而且

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

就是 $\left[p(x^{(i)},z^{(i)};\theta) / Q_i(z^{(i)}) \right]$ 的期望（回想期望公式中的Lazy Statistician规则）

设Y是随机变量X的函数，$Y = g(X)$（g是连续函数），那么

（1）X是离散型随机变量，它的分布律为$P(X = x_k) = p_k$，k=1,2,…。若$\sum_{k=1}^{\infty} g(x_k)p_k$绝对收敛，则有

$$E(Y) = E[g(X)] = \sum_{k=1}^{\infty} g(x_k)p_k$$

（2）X是连续型随机变量，它的概率密度为$f(x)$，若$\int_{-\infty}^{\infty} g(x)f(x)dx$绝对收敛，则有

$$E(Y) = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

对应于上述问题，Y是 $\left[p(x^{(i)}, z^{(i)}; \theta)/Q_i(z^{(i)})\right]$ ，X是$z^{(i)}$，$Q_i(z^{(i)})$是p_k，g是$z^{(i)}$到 $\left[p(x^{(i)}, z^{(i)}; \theta)/Q_i(z^{(i)})\right]$ 的映射。这样解释了式子（2）中的期望，再根据凹函数时的Jensen不等式：

$$f\left(\mathrm{E}_{z^{(i)}\sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]\right) \geq \mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right)\right],$$

可以得到（3）。

这个过程可以看作是对$\ell(\theta)$求了下界。对于Q_i的选择，有多种可能，那种更好的？假设θ已经给定，那么$\ell(\theta)$的值就决定于$Q_i(z^{(i)})$和$p(x^{(i)},z^{(i)})$了。我们可以通过调整这两个概率使下界不断上升，以逼近$\ell(\theta)$的真实值，那么什么时候算是调整好了呢？当不等式变成等式时，说明我们调整后的概率能够等价于$\ell(\theta)$了。按照这个思路，我们要找到等式成立的条件。根据Jensen不等式，要想让等式成立，需要让随机变量变成常数值，这里得到：

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

c为常数，不依赖于$z^{(i)}$。对此式子做进一步推导，我们知道$\sum_z Q_i(z^{(i)}) = 1$，那么也就有$\sum_z p(x^{(i)}, z^{(i)};\theta) = c$，（多个等式分子分母相加不变，这个认为每个样例的两个概率比值都是c），那么有下式：

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

至此，我们推出了在固定其他参数θ后，$Q_i(z^{(i)})$的计算公式就是后验概率，解决了$Q_i(z^{(i)})$如何选择的问题。这一步就是E步，建立$\ell(\theta)$的下界。接下来的M步，就是在给定$Q_i(z^{(i)})$后，调整θ，去极大化$\ell(\theta)$的下界（在固定$Q_i(z^{(i)})$后，下界还可以调整的更大）。那么一般的EM算法的步骤如下：

循环重复直到收敛 {

（E步）对于每一个i，计算

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

（M步）计算

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

那么究竟怎么确保EM收敛？假定$\theta^{(t)}$和$\theta^{(t+1)}$是EM第t次和t+1次迭代后的结果。如果我们证明了$\ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)})$，也就是说极大似然估计单调增加，那么最终我们会到达最大似然估计的最大值。下面来证明，选定$\theta^{(t)}$后，我们得到E步

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

这一步保证了在给定$\theta^{(t)}$时，Jensen不等式中的等式成立，也就是

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

然后进行M步，固定$Q_i^{(t)}(z^{(i)})$，并将$\theta^{(t)}$视作变量，对上面的$\ell(\theta^{(t)})$求导后，得到$\theta^{(t+1)}$，这样经过一些推导会有以下式子成立：

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

解释第（4）步，得到$\theta^{(t+1)}$时，只是最大化$\ell(\theta^{(t)})$，也就是$\ell(\theta^{(t+1)})$的下界，而没有使等式成立，等式成立只有是在固定θ，并按E步得到Q_i时才能成立。

况且根据我们前面得到的下式，对于所有的Q_i和θ都成立

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

第（5）步利用了M步的定义，M步就是将$\theta^{(t)}$调整到$\theta^{(t+1)}$，使得下界最大化。因此（5）成立，（6）是之前的等式结果。

这样就证明了$\ell(\theta)$会单调增加。一种收敛方法是$\ell(\theta)$不再变化，还有一种就是变化幅度很小。

再次解释一下（4）、（5）、（6）。首先（4）对所有的参数都满足，而其等式成立条件只是在固定θ，并调整好Q时成立，而第（4）步只是固定Q，调整θ，不能保证等式一定成立。（4）到（5）就是M步的定义，（5）到（6）是前面E步所保证等式成立条件。也就是说E步会将下界拉到与$\ell(\theta)$一个特定值（这里$\theta^{(t)}$）一样的高度，而此时发现下界仍然可以上升，因此经过M步后，下界又被拉升，但达不到与$\ell(\theta)$另外一个特定值一样的高度，之后E步又将下界拉到与这个特定值一样的高度，重复下去，直到最大值。

如果我们定义

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

从前面的推导中我们知道$\ell(\theta) \geq J(Q,\theta)$，EM可以看作是J的坐标上升法，E步固定$\theta$，优化$Q$，M步固定$Q$优化$\theta$。

3. 重新审视混合高斯模型

我们已经知道了EM的精髓和推导过程，再次审视一下混合高斯模型。之前提到的混合高斯模型的参数$\emptyset$，μ和Σ计算公式都是根据很多假定得出的，有些没有说明来由。为了简单，这里在M步只给出$\emptyset$和μ的推导方法。

E步很简单，按照一般EM公式得到：

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

简单解释就是每个样例i的隐含类别$z^{(i)}$为j的概率可以通过后验概率计算得到。

在M步中，我们需要在固定$Q_i(z^{(i)})$后最大化最大似然估计，也就是

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

这是将$z^{(i)}$的k种情况展开后的样子，未知参数$\emptyset_j$，μ_j和Σ_j。

固定$\emptyset_j$和Σ_j，对μ_j求导得

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

等于0时，得到

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

这就是我们之前模型中的μ的更新公式。

然后推导$\emptyset_j$的更新公式。看之前得到的

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

在$\emptyset$和μ确定后，分子上面的一串都是常数了，实际上需要优化的公式是：

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

需要知道的是，$\emptyset_j$还需要满足一定的约束条件就是$\sum_{j=1}^{k} \emptyset_j = 1$。

这个优化问题我们很熟悉了，直接构造拉格朗日乘子。

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

还有一点就是$\emptyset_j \geq 0$，但这一点会在得到的公式里自动满足。

求导得，

$$\frac{\partial}{\partial \emptyset_j}\mathcal{L}(\emptyset) = \sum_{i=1}^{m}\frac{w_j^{(i)}}{\emptyset_j} + \beta$$

等于0，得到

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

也就是说$\phi_j \propto \sum_{i=1}^{m} w_j^{(i)}$。再次使用$\sum_{j=1}^{k} \emptyset_j = 1$，得到

$$-\beta = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} = \sum_{i=1}^{m} 1 = m.$$

这样就神奇地得到了β。

那么就顺势得到M步中$\emptyset_j$的更新公式：

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

Σ的推导也类似，不过稍微复杂一些，毕竟是矩阵。结果在之前的混合高斯模型中已经给出。

4. 总结

如果将样本看作观察值，潜在类别看作是隐藏变量，那么聚类问题也就是参数估计问题，只不过聚类问题中参数分为隐含类别变量和其他参数，这犹如在x-y坐标系中找一个曲线的极值，然而曲线函数不能直接求导，因此什么梯度下降方法就不适用了。但固定一个变量后，另外一个可以通过求导得到，因此可以使用坐标上升法，一次固定一个变量，对另外的求极值，最后逐步逼近极值。对应到EM上，E步估计隐含变量，M步估计其他参数，交替将极值推向最大。EM中还有“硬”指定和“软”指定的概念，“软”指定看似更为合理，但计算量要大，“硬”指定在某些场合如K-means中更为实用（要是保持一个样本点到其他所有中心的概率，就会很麻烦）。

另外，EM的收敛性证明方法确实很牛，能够利用log的凹函数性质，还能够想到利用创造下界，拉平函数下界，优化下界的方法来逐步逼近极大值。而且每一步迭代都能保证是单调的。最重要的是证明的数学公式非常精妙，硬是分子分母都乘以z的概率变成期望来套上Jensen不等式，前人都是怎么想到的。

在Mitchell的Machine Learning书中也举了一个EM应用的例子，明白地说就是将班上学生的身高都放在一起，要求聚成两个类。这些身高可以看作是男生身高的高斯分布和女生身高的高斯分布组成。因此变成了如何估计每个样例是男生还是女生，然后在确定男女生情况下，如何估计均值和方差，里面也给出了公式，有兴趣可以参考。