一、算法的介绍

1.EM 算法介绍：

一种迭代式的算法，用于含有隐变量的概率参数模型的极大似然估计或极大后验概率估计。

2.什么是隐变量：

隐变量: 比如聚类问题，样本 x 的特征是可观察到的。其还有一个隐藏属性: 所属类别 z。(x,z)整体是一个完整的观测样本，记为 y。

二、算法的推导和步骤

1.算法的推导：

假设现有一批独立同分布的样本数据$\{x_1, x_2, ..., x_m\}$[1]，它们是由某个含有隐变量的模型$p(x, z; \theta)$生成，现尝试用极大似然估计法估计此模型的参数。由对数似然函数的定义可知此时的对数似然函数为（假设 z 为离散型）：

$$LL(\theta) = \sum_{i=1}^{m} lnp(x_i; \theta) = \sum_{i=1}^{m} ln \sum_{z_i} p(x_i, z_i; \theta)$$

显然，此时$LL(\theta)$里含有未知的隐变量 z 以及和（z 为离散型时）或者积分（z 为连续型时）的对数，因此无法按照传统方法直接求出使得$LL(\theta)$达到最大值的模型参数θ，而 EM 算法给出了一种迭代的方法完成对$LL(\theta)$的极大化，下面是具体的推导方式。

设z_i的概率密度函数是$Q_i(z_i)$，对上面函数进行恒等变形：

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^{m} \ln p(x_i; \theta) \\ &= \sum_{i=1}^{m} \ln \sum_{z_i} p(x_i, z_i; \theta) \\ &= \sum_{i=1}^{m} \ln \sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i; \theta)}{Q_i(z_i)} \end{aligned}$$

其中$\sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}$可以看作是对$\frac{p(x_i, z_i; \theta)}{Q_i(z_i)}$关于$z_i$求期望。

那么由 Jensen 不等式可知：

$$\ln\left(E_{z_i}\left[\frac{p(x_i, z_i; \theta)}{Q_i(z_i)}\right]\right) \geq E_{z_i}\left[\ln\left(\frac{p(x_i, z_i; \theta)}{Q_i(z_i)}\right)\right]$$

$$\ln \sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i; \theta)}{Q_i(z_i)} \geq \sum_{z_i} Q_i(z_i) \ln \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}$$

将上述不等式待人到$LL(\theta)$中，有：

$$
\begin{aligned}
LL(\theta) &= \sum_{i=1}^{m} \ln \sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i; \theta)}{Q_i(z_i)} \qquad (A.1) \\
&\geq \sum_{i=1}^{m} \sum_{z_i} Q_i(z_i) \ln \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}
\end{aligned}
$$

若令$B(\theta) = \sum_{i=1}^{m} \sum_{z_i} Q_i(z_i) ln \frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$，此时$B(\theta)$为$LL(\theta)$的下界函数，即$LL(\theta)$为$B(\theta)$的上界函数，所以如果能使得$B(\theta)$=$LL(\theta)$。那么$B(\theta)$取得最大值。由延森不等式可知，如果$\frac{p(x_i,z_i;\theta)}{Q_i(z_i)}$等于一个常量 c，则大于等于号可以取得等号。因此，只要任意选取满足$\frac{p(x_i,z_i;\theta)}{Q_i(z_i)} = c$的$Q_i(z_i)$就可以使之达到最大值。又因为$Q_i(z_i)$为概率密度函数，故$Q_i(z_i)$满足如下约束：

$$
0 < Q_i(z_i) < 1
$$

$$
\sum_{z_i} Q_i(z_i) = 1
$$

于是可以推出：

$$
\frac{p(x_i, z_i; \theta)}{Q_i(z_i)} = c
$$

$$
p(x_i, z_i; \theta) = c \cdot Q_i(z_i)
$$

$$
\sum_{z_i} p(x_i, z_i; \theta) = c \cdot \sum_{z_i} Q_i(z_i)
$$

$$
\sum_{z_i} p(x_i, z_i; \theta) = c
$$

$$
\frac{p(x_i, z_i; \theta)}{Q_i(z_i)} = \sum_{z_i} p(x_i, z_i; \theta)
$$

$$
Q_i(z_i) = \frac{p(x_i, z_i; \theta)}{\sum_{z_i} p(x_i, z_i; \theta)} = \frac{p(x_i, z_i; \theta)}{p(x_i; \theta)} = p(z_i | x_i; \theta)
$$

所以，当且仅当$Q_i(z_i) = p(z_i|x_i;\theta)$时$B(\theta)$取得最大值，将该式代入$B(\theta)$和$LL(\theta)$中可以推得：

$$
\begin{aligned}
LL(\theta) &= \sum_{i=1}^{m} \ln \sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i; \theta)}{Q_i(z_i)} && (A.2.1) \\
&= \sum_{i=1}^{m} \ln \sum_{z_i} p(z_i | x_i; \theta) \frac{p(x_i, z_i; \theta)}{p(z_i | x_i; \theta)} && (A.2.2) \\
&= \sum_{i=1}^{m} \sum_{z_i} p(z_i | x_i; \theta) \ln \frac{p(x_i, z_i; \theta)}{p(z_i | x_i; \theta)} && (A.2.3) \\
&= \max\{B(\theta)\} && (A.2.4)
\end{aligned}
$$

其中，（A.2.3）是（A.1）中不等号取等号时的情形。由以上推导可知，对数似然函数$LL(\theta)$等价于其下界函数的最大值$max\{B(\theta)\}$，所以要想极大化$LL(\theta)$可以通过极大化max{B(θ)}来间接极大化 LL(θ)，因此，下面考虑如何极大化$max\{B(\theta)\}$。假设**已知**第 t 次迭代的参数为$\theta^{(t)}$，而第$t+1$ 次迭代的参数$\theta^{(t+1)}$通过如下方式求得：

$$
\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \max\{B(\theta)\} && (A.3)\\
&= \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z_i} p(z_i|x_i;\theta^{(t)}) \ln \frac{p(x_i,z_i;\theta)}{p(z_i|x_i;\theta^{(t)})}\\
&= \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z_i} p(z_i|x_i;\theta^{(t)}) \ln p(x_i,z_i;\theta)
\end{aligned}
$$

将$\theta^{(t+1)}$代入$LL(\theta^{(t+1)})$则可以进一步推得:(这一步为收敛性的证明)

$$
\begin{aligned}
LL(\theta^{(t+1)}) &= \max\{B(\theta^{(t+1)})\} && (A.4.1)\\
&= \sum_{i=1}^{m} \sum_{z_i} p(z_i|x_i;\theta^{(t+1)}) \ln \frac{p(x_i,z_i;\theta^{(t+1)})}{p(z_i|x_i;\theta^{(t+1)})} && (A.4.2)\\
&\geq \sum_{i=1}^{m} \sum_{z_i} p(z_i|x_i;\theta^{(t)}) \ln \frac{p(x_i,z_i;\theta^{(t+1)})}{p(z_i|x_i;\theta^{(t)})} && (A.4.3)\\
&\geq \sum_{i=1}^{m} \sum_{z_i} p(z_i|x_i;\theta^{(t)}) \ln \frac{p(x_i,z_i;\theta^{(t)})}{p(z_i|x_i;\theta^{(t)})} && (A.4.4)\\
&= \max\{B(\theta^{(t)})\} && (A.4.5)\\
&= LL(\theta^{(t)}) && (A.4.6)
\end{aligned}
$$

其中,(A.4.1)和(A.4.2)由(A.2)推得;(A.4.3)由(A.1)推得;(A.4.4)由(A.3)推得;(A.4.5)和(A.4.6)由(A.2)推得。显然,若令

$$Q(\theta,\theta^{(t)}) = \sum_{i=1}^{m} \sum_{z_i} p(z_i|x_i;\theta^{(t)}) lnp(x_i,z_i;\theta\ )$$

那么由(A.4)可知,凡是能使得$Q(\theta,\theta^{(t)})$达到极大的$\theta^{(t+1)}$一定能使得$LL(\theta^{(t+1)}) \geq LL(\theta^{(t)})$。综上所述即可总结出 EM 算法的"E 步"和"M 步"分别为:

E 步:令 $Q_i(z_i) = \ p(z_i|x_i;\theta)$并写出$Q(\theta,\theta^{(t)})$;

M 步:求使得$Q(\theta,\theta^{(t)})$到达极大的$\theta^{(t+1)}$。

算法流程:

假设现有一批独立同分布的样本数据$\{x_1,x_2,\ldots,x_m\}$,它们是由某个含有隐变量的模型$p(x,z;\theta)$生成,最大迭代数为 J。

2.算法步骤:

(1) 随机初始化模型参数θ的初值$\theta^{(0)}$。

(2) *t*=1,2…*J* 开始进行迭代

- *E* 步计算$Q_i(z_i) = \ p(z_i|x_i;\theta)$ $Q(\theta,\theta^{(t)}) = \sum_{i=1}^{m}\sum_{z_i} p(z_i|x;\theta^{(t)}) lnp(x_i,z_i;\theta\ )$
- *M* 步:求使得$Q(\theta,\theta^{(t)})$到达极大的$\theta^{(t+1)}$。

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

- 如果$\theta^{(t+1)}$已经收敛,则算法结束。否则继续进行迭代。

• 输出：模型参数θ。

3.从信息论角度解释[2]：

首先因为 $p(\boldsymbol{x},\boldsymbol{z};\theta)=p(\boldsymbol{z}|\boldsymbol{x};\theta)p(\boldsymbol{x};\theta)$，有 $\log p(\boldsymbol{x},\boldsymbol{z};\theta)=\log p(\boldsymbol{z}|\boldsymbol{x};\theta)+\log p(\boldsymbol{x};\theta)$，进一步有 $\log p(\boldsymbol{x};\theta)=\log p(\boldsymbol{x},\boldsymbol{z};\theta)-\log p(\boldsymbol{z}|\boldsymbol{x};\theta)$.

这样，对数边际似然 $\log p(\boldsymbol{x};\theta)$ 可以分解为

$$
\begin{aligned}
\log p(\boldsymbol{x};\theta) &= \sum_{\boldsymbol{z}} q(\boldsymbol{z})\log p(\boldsymbol{x};\theta) && (11.46)\\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z})\Big(\log p(\boldsymbol{x},\boldsymbol{z};\theta)-\log p(\boldsymbol{z}|\boldsymbol{x};\theta)\Big) && (11.47)\\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z})\log\frac{p(\boldsymbol{x},\boldsymbol{z};\theta)}{q(\boldsymbol{z})}-\sum_{\boldsymbol{z}} q(\boldsymbol{z})\log\frac{p(\boldsymbol{z}|\boldsymbol{x};\theta)}{q(\boldsymbol{z})} && (11.48)\\
&= ELBO(q,\boldsymbol{x};\theta)+\mathrm{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta)), && (11.49)
\end{aligned}
$$

$\sum_{\boldsymbol{z}} q(\boldsymbol{z})=1.$

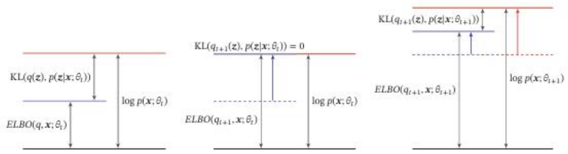
其中 $\mathrm{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta))$ 为分布 $q(\boldsymbol{z})$ 和后验分布 $p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 的 KL 散度.

参见第 E.3.2 节.

由于 $\mathrm{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta))\geq 0$，因此 $ELBO(q,\boldsymbol{x};\theta)$ 为 $\log p(\boldsymbol{x};\theta)$ 的一个下界. 当且仅当 $q(\boldsymbol{z})=p(\boldsymbol{z}|\boldsymbol{x};\theta)$ 时，$\mathrm{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta))=0$，这时 $ELBO(q,\boldsymbol{x};\theta)=\log p(\boldsymbol{x};\theta)$.

我们的思路是：假如 KL 散度这一项为 0，则 $\log p(x;\theta)$这一项就和 ELBO 项相等。而 ELBO 项是关于θ的函数，我们找到使 ELBO 项达到最大的 argmax θ，并不断迭代下去。

图11.11在 EM 算法在第 t 步迭代时的示例. 图11.11a为第 t 步迭代的初始状态，参数为 θ_t，这时通常有 $\mathrm{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta_t))>0$. 图11.11b为 E 步更新：固定参数 θ_t，找到分布 $q_{t+1}(\boldsymbol{z})$ 使得 $\mathrm{KL}(q_{t+1}(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta_t))=0$，这时 $ELBO(q_{t+1},\boldsymbol{x};\theta_t)$ 和 $\log p(\boldsymbol{x};\theta_t)$ 相等. 图11.11c为 M 步更新：固定分布 $q_{t+1}(\boldsymbol{z})$，寻找参数 θ_{t+1} 使得 $ELBO(q_{t+1},\boldsymbol{x};\theta_{t+1})$ 最大. 由于这时通常 $\mathrm{KL}(q_{t+1}(\boldsymbol{z})\|p(\boldsymbol{z}|\boldsymbol{x};\theta_t))>0$，从而 $\log p(\boldsymbol{x};\theta_{t+1})$ 也变大.



(a) 初始状态　(b) E 步更新　(c) M 步更新

图 11-11

4.对于高斯混合模型

假设随机变量X是由K个高斯分布混合而成，各个高斯分布的概率为$\phi_1, \phi_2, \ldots, \phi_K$，第i个高斯分布的均值为$\mu_i$，方差为$\sigma_i$。我们观测到随机变量X的一系列样本值为$x_1, x_2, \ldots, x_n$，计算如下：

第一步：给φ，μ，σ赋初值，开启迭代，高斯混合模型的φ，μ，σ有多个，就分别赋初值；

第二步：E步。如果是首轮迭代，那么φ，μ，σ分别为我们给定的初值；否则φ，μ，σ取决于上一轮迭代的值。有了φ，μ，σ的值，我们按照如下公式计算Q函数：

$$Q_i(z^{(i)}=k)=\frac{\phi_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k}\exp[-\frac{(x^{(i)}-\mu_k)^2}{2\sigma_k^2}]}{\sum_{k=1}^{K}\phi_k \cdot \frac{1}{\sqrt{2\pi}\sigma_k}\exp[-\frac{(x^{(i)}-\mu_k)^2}{2\sigma_k^2}]}$$

其中，φ，σ，μ，x均已知，代入即可，i=1,2,...,N；k=1,2,...,K

第三步：M步。根据计算出来的Q，套进以下公式算出高斯混合模型的各个参数：

$$\mu_k=\frac{\sum_{i=1}^{N}Q_k^{(i)}x^{(i)}}{N_k}$$

$$\sigma_k=\frac{\sum_{i=1}^{N}Q_k^{(i)}(x^{(i)}-\mu_{(k)})(x^{(i)}-\mu_k)^T}{N_k}$$

$$\phi_k=\frac{\sum_{i=1}^{N}Q_k^{(i)}}{N}$$

$$N_k=\sum_{i=1}^{N}Q_k^{(i)}$$

重复2~3步，直至收敛。

5.例子

比如：比如一班二班各有 50 名同学，某次考试这两个班同学没写班级，老师想找出这

100 个成绩里面那些是一班、那些是二班的。

我们想要用两个高斯分布去拟合两个班级的成绩，这样的模型也称为高斯混合模型。
现假设如下分布：
成绩来自一班的同学：

$$p(x|\gamma 1)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$$

成绩来自二班的同学：

$$p(x|\gamma 2)=\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

现在我们假设σ_1、σ_2、μ_1、μ_2为已知量。且假设$p(\gamma 2)=p(\gamma 1)=0.5$,即每个同学的成绩来

自一班或二班的概率相等。

这样的话，根据 *Bayes* 定理：

第x_i个同学来自一班的概率为：*i=1.....100*

$$p(\gamma 1|x_i) = \frac{p(x_i|\gamma 1)p(\gamma 1)}{p(x_i|\gamma 1)p(\gamma 1) + p(x_i|\gamma 2)p(\gamma 2)}$$

第x_i个同学来自二班的概率为：

$$p(\gamma 2|x_i) = \frac{p(x_i|\gamma 2)p(\gamma 2)}{p(x_i|\gamma 1)p(\gamma 1) + p(x_i|\gamma 2)p(\gamma 2)}$$

这就是 E 步。

这样的话我们就可以根据得到的数据来校正之前假设的信息了。

$$\mu_1 = \frac{\gamma 11 * x_1 + \gamma 21 * x_1 + \dots + \gamma 1001 * x_1}{\gamma 11 + \dots + \gamma 1001}$$

$\gamma 11$ 代表第一个同学来自一班的概率，$\gamma 1001$ 代表最后一个同学来自一班的概率。

$$\sigma_1^2 = \frac{\gamma 11 * (x_1 - \mu_1)^2 + \dots + \gamma 1001 * (x_{100} - \mu_1)^2}{\gamma 11 + \dots + \gamma 1001}$$

注意上面是方差。

σ_2和μ_2同理。

并且先验概率：

$$p(\gamma 1) = \frac{\gamma 11 + \dots + \gamma 1001}{100}$$
$$p(\gamma 2) = \frac{\gamma 12 + \dots + \gamma 1002}{100}$$

上面为 M 步。

然后继续迭代。

例：EM 算法详解+通俗例子理解_em 算法实例-CSDN 博客

参考文献：

[1] Dempster, A. P. . (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, 39.

[2]神经网络和深度学习-邱锡鹏