EM 算法

一. EM 算法原理

EM 算法求解问题：

在部分已知相关变量 X 和部分未知潜在变量 Y 的概率模型中，求取参数θ的最大似然估计$l(\theta)$，使潜在 X 变量出现的可能性最大。

解决办法：

因为潜在变量 Y 和参数θ都未知，直接最大化$l(\theta)$很困难，所以先确定$l(\theta)$的下界，并优化下界，直到找到最优结果。

也就是说，首先赋予θ某个初值，得到 Y 的某种估计值，再从 Y 的当前结果出发，重新估计θ的值，交替进行，直至收敛。

1. EM 算法的快速计算

给定样本集合$Z=(X,Y)=\{(x_1,y_1),..,(x_N,y_N)\}$，包括 N 个独立的可观测数据$x_1,...,x_N$和 N 个无法观测的数据$y_1,...,y_N\in\Theta$，即潜在变量，$\Theta$为潜在变量所在的参数空间。$Z=(X,Y)$和 X 分别称为不完全数据集和完全数据集。假设 Z 的联合概率密度函数为$P(X,Y\mid\theta)$，其中θ为需要估计的参数。

对θ的估计可以通过最大似然方法得到，从服从同一分布$P(X,Y\mid\theta)$的概率模型中观测到数据集合$x_1,...,x_N$的似然函数为：

$$L(\theta)=L(x_1,...,x_N\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)$$

这个函数反映了在不同的参数θ下，取得当前这个样本集的可能性。然而$L(\theta)$是连续相乘的，不便于分析，为此定义连续相加的对数似然函数$l(\theta)$：

$$l(\theta)=In\ L(\theta)=In\prod_{i=1}^{N}p(x_i\mid\theta)=\sum_{i=1}^{N}In\ p(x_i\mid\theta)$$

对于不存在的潜在变量，对完全数据集 X 而言，直接求解上述函数的导数或偏导数，

可以得到最大似然估计量：

$$\hat{\theta} = \arg\max l(\theta)$$

然而，完全数据集 X 仅仅是不完全数据集的可观测部分，不完全数据集的对数似然函数为:

$$l(\theta) = In\ L(\theta) = In \prod_{i=1}^{N} p(x_i \mid \theta) = \sum_{i=1}^{N} In \sum_{y_i \in \Theta} p(x_i, y_i \mid \theta)$$

此时，我们的目标变为找到合适的参数θ和$y_1,...,y_N \in \Theta$，使得$l(\theta)$最大，所以很难直接求导得到θ的估计量，但是可以假定指导$y_1,...,y_N \in \Theta$，然后仍然可以通过求导得到θ的估计量。

为使求参数θ进一步简化，假设y_i满足某种分布P_i，运用 Jensen 不等式，则：

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} In \sum_{y_i \in \Theta} p(x_i, y_i \mid \theta) \\ &= \sum_{i=1}^{N} In \sum_{y_i \in \Theta} p_i(y_i) \frac{p(x_i, y_i \mid \theta)}{p_i(y_i)} \\ &\geq \sum_{i=1}^{N} \sum_{y_i \in \Theta} p_i(y_i) In\ \frac{p(x_i, y_i \mid \theta)}{p_i(y_i)} = \tilde{l}(\theta) \end{aligned}$$

因为不等号的存在，对$\tilde{l}(\theta)$求最大似然估计并不意味着对$l(\theta)$的最大似然估计，但$\tilde{l}(\theta)$是$l(\theta)$的一个下界，通过不断最大化这个下界，可以使$l(\theta)$不断逼近最大值。

2. 未知分布$p_i(y_i)$的选择

假设参数θ的情况下，$p_i(y_i)$的最优选择后验概率$p(y_i \mid x_i, \theta)$，因此：

E 步：对每个$i = 1,...,N$，计算：

$$P_i(y_i) = p(y_i \mid x_i, \theta)$$

M 步：计算最大似然估计：

$$\arg\max \sum_{i=1}^{N} \sum_{y_i \in \Theta} p_i(y_i) In\ \frac{p(x_i, y_i \mid \theta)}{p_i(y_i)}$$

更新参数θ。

3. 算法收敛性——终止条件

假设第 k 次迭代的得到参数 θ^k，则第 k+1 次迭代的 E 步，有:

$$P_i^{k+1}(y_i) = p(y_i \mid x_i, \theta^k)$$

则 k+1 次迭代的 M 步中最大似然函数（$P_i^{k+1}(y_i)$ 使等号成立）：

$$l(\theta^k) = \sum_{i=1}^{N} \sum_{y_i \in \Theta} p_i^{k+1}(y_i) In \ \frac{p(x_i, y_i \mid \theta^k)}{p_i^{k+1}(y_i)}$$

通过最大似然估计得到更新参数 θ^{k+1}，所以:

$$\sum_{i=1}^{N} \sum_{y_i \in \Theta} p_i^{k+1}(y_i) In \ \frac{p(x_i, y_i \mid \theta^{k+1})}{p_i^{k+1}(y_i)} \geq \sum_{i=1}^{N} \sum_{y_i \in \Theta} p_i^{k+1}(y_i) In \ \frac{p(x_i, y_i \mid \theta^k)}{p_i^{k+1}(y_i)}$$

同时,

$$l(\theta^{k+1}) \geq \sum_{i=1}^{N} \sum_{y_i \in \Theta} p_i^{k+1}(y_i) In \ \frac{p(x_i, y_i \mid \theta^{k+1})}{p_i^{k+1}(y_i)}$$

因此，收敛的条件为:

$$l(\theta^{k+1}) \geq l(\theta^k)$$

考虑到 $l(\theta)$ 在逼近最大值时增长缓慢，常用终止条件为:

$$\frac{l(\theta^{k+1}) - l(\theta^k)}{l(\theta^k)} \leq T$$

T 是常数，如 $T = 10^{-3}$。

4. 算法特点

EM 算法常用于解决不完全数据集中参数的最大似然估计问题，现也应用于机器学习和数据聚类领域，特点如下：

1） 迭代简单、稳定，用于极大似然估计和计算后验密度函数；
2） 依赖于初始值的选取，容易陷入局部最优解；
3） 复杂度高，收敛速度慢，不适用于大规模数据和高维数据。

5. 仿真实验

EM 算法伪代码：

Input：

$x_1,...,x_N$为可观测数据，Θ为潜在变量$y_1,...,y_N$的参数空间，θ^0为未知参数的初始值，K为最大迭代次数，T为终止条件

For　k=1：K

Step1　E步，计算

$$P_i^k(y_i)=p(y_i\,|\,x_i,\theta^{k-1})$$

Step2　M步，计算

$$\arg\max\sum_{i=1}^{N}\sum_{y_i\in\Theta}p_i^{\,k}(y_i)In\ \frac{p(x_i,y_i\,|\,\theta)}{p_i^{\,k}(y_i)}$$

得到θ^k

if $\dfrac{l(\theta^{k+1})-l(\theta^k)}{l(\theta^k)}\le T$

break;

end

end for

Output:最大似然函数估计值θ^k

假设一组样本数据$x_1,...,x_N$服从高斯混合分布，即数据由c个满足高斯正态分布$N(\mu_j,\Sigma_j)$的不同概率模型混合而成，但不知道每个数据具体来自哪个高斯模型，它可能以概率q_j来自第j个模型；也不知道各高斯模型的具体参数（μ_j,Σ_j），根据这些观测的样本$x_1,...,x_N$求出这些参数$\theta=\{q,\mu,\Sigma\}$。

若用潜在变量$y_1,...,y_N$表示$x_1,...,x_N$所在的模型，则：

$$p(y_i=j\,|\,\theta)=q_j$$

$$x_i\,|\,y_i=j,\theta\sim N(\mu_j,\Sigma_j)$$

对数似然函数有：

$$l(\theta)=\sum_{i=1}^{N}In\ p(x_i\,|\,\theta)=\sum_{i=1}^{N}In\sum_{j=1}^{c}p(x_i,y_i\,|\,\theta)$$

则：

E步：潜在变量的后验概率：

$$
\begin{aligned}
p_j^{k+1}(y_i=j) &= p(y_i=j \mid x_i,\theta^k) \\
&= \frac{p(y_i=j,x_i \mid \theta^k)}{\sum_{j=1}^{c} p(y_i=j,x_i \mid \theta^k)} \\
&= \frac{p(x_i \mid y_i=j,\theta^k)\,p(y_i=j \mid \theta^k)}{\sum_{j=1}^{c} p(x_i \mid y_i=j,\theta^k)\,p(y_i=j \mid \theta^k)} \\
&= \frac{N(x_i \mid \mu_j^k,\Sigma_j^k)q_j^k}{\sum_{j=1}^{c} N(x_i \mid \mu_j^k,\Sigma_j^k)q_j^k}
\end{aligned}
$$

M 步：计算最大似然估计：

$$
\arg\max \sum_{i=1}^{N}\sum_{j=1}^{c} p_i^{k+1}(y_i=j) In\ \frac{p(x_i,y_i=j \mid \theta)}{p_i^{k+1}(y_i=j)}
$$

通过求导得到参数的估计值：

$$
q_j^{k+1} = \frac{\sum_{i=1}^{N} p_i^{k+1}(y_i=j)}{N}
$$

$$
\mu_j^{k+1} = \frac{\sum_{i=1}^{N} p_i^{k+1}(y_i=j)x_i}{\sum_{i=1}^{N} p_i^{k+1}(y_i=j)}
$$

$$
\Sigma_j^{k+1} = \frac{\sum_{i=1}^{N} p_i^{k+1}(y_i=j)(x_i-\mu_j^{k+1})(x_i-\mu_j^{k+1})^T}{\sum_{i=1}^{N} p_i^{k+1}(y_i=j)}
$$