理解 EM 算法

EM（ expectation-maximization，期望最大化）算法是机器学习中与 SVM（支持向量机）、概率图模型并列的难以理解的算法，主要原因在于其原理较为抽象，初学者无法抓住核心的点并理解算法求解的思路。本文对 EM 算法的基本原理进行系统的阐述，并以求解高斯混合模型为例说明其具体的用法。文章是对已经在清华大学出版社出版的《机器学习与应用》一书中 EM 算法的讲解，对部分内容作了扩充。

算法的历史

EM 算法即期望最大化算法，由 Dempster 等人在 1976 年提出[1]。这是一种迭代法，用于求解含有隐变量的最大似然估计、最大后验概率估计问题。至于什么是隐变量，在后面会详细解释。EM 算法在机器学习中有大量成功的应用，典型是求解高斯混合模型，隐马尔可夫模型。如果你要求解的机器学习模型中有隐变量存在，并且要估计模型的参数，EM 算法很多时候是首选算法。

Jensen 不等式

EM 算法的推导、收敛性证明依赖于 Jensen 不等式，我们先对它做一简单介绍。Jensen 不等式的表述是，如果 $f(\mathbf{x})$ 是凸函数， $\mathbf{x}$ 是随机变量，则下面不等式成立

$$\mathrm{E}(f(\mathbf{x})) \geq f(\mathrm{E}(\mathbf{x}))$$

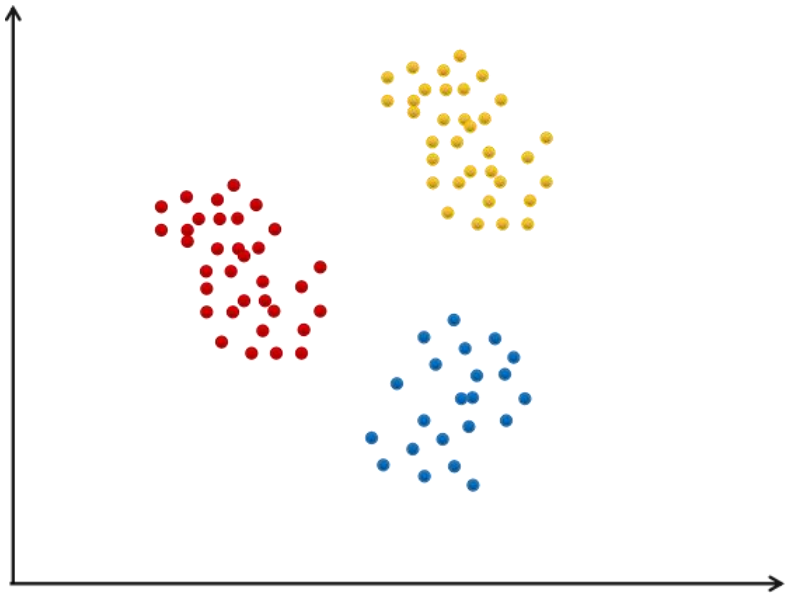
在这里 E 是数学期望，对于离散型随机变量，数学期望是求和，对连续型随机变量则为求定积分。如果 $f(\mathbf{x})$ 是一个严格凸函数，当且仅当 $\mathbf{x}$ 是常数时不等式取等号：

$$\mathrm{E}(f(\mathbf{x})) = f(\mathrm{E}(\mathbf{x}))$$

如果对这一不等式的证明感兴趣，可以阅读相关的数学教材。

高斯混合模型

EM 算法的目标是求解似然函数或后验概率的极值，而样本中具有无法观测的隐含变量。下面以聚类问题和高斯混合模型为例进行说明。有一批样本，分属于 3 个类，假设每个类都服从正态分布，均值和协方差未知，各样本属于哪个类也是未知的，算法需要在此条件下估计每个正态分布的均值和协方差。下图是一个例子，3 类样本都服从正态分布，但每个样本属于哪个类是未知的：



样本所属的类别就是隐变量，我们无法直接观察到它的值，这种隐变量的存在导致了用最大似然估计求解时的困难，后面会解释。上面这个例子可以用高斯混合模型进行描述，它的概率密度函数是多个高斯分布（正态分布）的加权和。

高斯混合模型（Gaussian Mixture Model，简称 GMM）通过多个正态分布的加权和来描述一个随机变量的概率分布，概率密度函数定义为：

$$p(\mathrm{x}) = \sum_{i=1}^{k} w_i N_i(\mathrm{x}; \mu_i, \Sigma_i)$$

其中 x 为随机向量，k 为高斯分布的数量，w_i 为高斯分布的权重，是一个正数，μ 为高斯分布的均值向量，Σ 为协方差矩阵。所有高斯分布的权重之和为 1，即：

$$\sum_{i=1}^{k} w_i = 1$$

高斯混合模型的样本可以看作是这样产生的：

先从 k 个高斯分布中选择出一个，选择第 i 个高斯分布的概率为 w_i，再由第 i 个高斯分布 $N(\mathrm{x}, \mu_i, \Sigma_i)$ 产生出样本数据 x。

高斯混合模型可以逼近任何一个连续的概率分布，因此它可以看做是连续型概率分布的万能逼近器。之所有要保证权重的和为 1，是因为概率密度函数必须满足在 $(-\infty, +\infty)$ 内的积分值为 1。

回忆一下用最大似然估计来确定单个高斯分布的参数的过程，给定一组训练样本，构造它们的对数似然函数，对参数求导并令导数为 0，即可通过最大化对数似然函数而确定高斯分布的参数。

对于高斯混合模型，也可以使用最大似然估计确定模型的参数，但每个样本属于哪个高斯分布是未知的，而计算高斯分布的参数时需要用到这个信息；反过来，样本属于哪个高斯分布又是由高斯分布的参数确定的。因此存在循环依赖，解决此问题的办法是打破此循环依

赖，从高斯分布的一个不准确的初始猜测值开始，计算样本属于每个高斯分布的概率，然后又根据这个概率更新每个高斯分布的参数。这就是 EM 算法求解时的做法。

从另外一个角度看，高斯混合模型的对数似然函数为：

$$\sum_{i=1}^{l}\ln\left(\sum_{j=1}^{k}w_j N_j\left(\mathrm{x};\mu_j,\Sigma_j\right)\right)$$

由于对数函数中有 k 个求和项，以及参数 w_j 的存在，无法像单个高斯模型那样通过最大似然估计求得公式解。如果用梯度下降法近似求解，则要保证 w_j 非负并且和为 1，同样存在困难。

算法的推导

下面考虑一般的情况。有一个概率分布 $p(\mathbf{x};\boldsymbol{\theta})$，从它生成了 l 个样本。每个样本包含观测数据 $\mathbf{x}_i$，以及无法观测到的隐变量 z_i，这个概率分布的参数 θ 是未知的，现在需要根据这些样本估计出参数 θ 的值。因为不知道隐变量的值，所以要消掉它，这通过对其求边缘概率而实现。采用最大似然估计，可以构造出对数似然函数：

$$\begin{aligned}L(\boldsymbol{\theta}) &= \sum_{i=1}^{l}\ln p(\mathbf{x}_i;\boldsymbol{\theta})\\ &= \sum_{i=1}^{l}\ln\sum_{z}p(\mathbf{x}_i,z_i;\boldsymbol{\theta})\end{aligned}$$

这里的 z_i 是一个无法观测到（即不知道它的值）的隐含变量，可以看作离散型随机变量，上式对隐含变量 z 的所有情况下的联合概率 $p(\mathbf{x},z;\boldsymbol{\theta})$ 求和得到 x 的边缘概率。因为隐含变量的存在，无法直接通过最大化似然函数得到参数的公式解。如果使用梯度下降法或牛顿法求解，则要保证隐变量所满足的等式和不等式约束

$$\sum_{z}p(z)=1$$
$$p(z)\geq 0$$

这同样存在困难。

EM 算法所采用的思路是构造出对数似然函数的一个下界函数，这个下界函数更容易优化，然后优化这个下界。不断的改变优化变量的值使得下界函数的值升高，从而使得对数似然函数的值也上升。

对每个样本 i，假设 Q_i 为隐变量 z_i 的一个概率分布，根据对概率分布的要求它必须满足：

$$\sum_z Q_i(z) = 1$$

$$Q_i(z) \geq 0$$

利用这个概率分布，将对数似然函数变形，可以得到：

$$\begin{aligned}\sum_{i=1}^{l} \ln p(\mathbf{x}_i;\boldsymbol{\theta}) &= \sum_{i=1}^{l} \ln \sum_{z_i} p(\mathbf{x}_i, z_i;\boldsymbol{\theta}) \\ &= \sum_{i=1}^{l} \ln \sum_{z_i} Q_i(z_i) \frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)} \\ &\geq \sum_{i=1}^{l} \sum_{z_i} Q_i(z_i) \ln \frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)}\end{aligned}$$

上式的第二步凑出了数学期望，最后一步利用了 Jensen 不等式。如果令函数

$$f(x) = \ln x$$

按照数学期望的定义（注意，在这里 z_i 是随机变量，是对它求数学期望），有：

$$\begin{aligned}&\ln \sum_{z_i} Q_i(z_i) \frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)} = \\ &f\left(\mathrm{E}_{Q_i(z_i)}\left(\frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)}\right)\right) = \ln\left(\mathrm{E}_{Q_i(z_i)}\left(\frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)}\right)\right) \\ &\geq \mathrm{E}_{Q_i(z_i)} f\left(\frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)}\right) = \mathrm{E}_{Q_i(z_i)} \ln\left(\frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)}\right) \\ &= \sum_{z_i} Q_i(z_i) \ln \frac{p(\mathbf{x}_i, z_i;\boldsymbol{\theta})}{Q_i(z_i)}\end{aligned}$$

对数函数是凹函数，Jensen 不等式反号。上式给出了对数似然函数的一个下界，Q_i 可以是任意一个概率分布，因此可以利用参数 $\boldsymbol{\theta}$ 的当前估计值来构造 Q_i。显然，这个下界函数更容易求极值，因为对数函数里面已经没有求和项，对参数求导并令导数为 0 时一般可以得到公式解。

算法的流程

算法在实现时首先初始化参数 $\boldsymbol{\theta}$ 的值，接下来循环迭代直至收敛，每次迭代时分为两步：

E 步，基于当前的参数估计值 $\boldsymbol{\theta}_i$，计算在给定 $\mathbf{x}$ 时对隐变量 z 的条件概率：

$$Q_i(z_i) = p(z_i | \mathrm{x}_i;\boldsymbol{\theta})$$

接下来根据该概率论构造目标函数（下界函数），这个目标函数是对 z 的数学期望，这就是 EM 算法中“E”的含义。

M 步，求解如下极值问题，更新 $\boldsymbol{\theta}$ 的值：

$$\theta = \arg\max_\theta \sum_i \sum_{z_i} Q_i(z_i) \ln \frac{p(\mathrm{x}_i, z_i; \theta)}{Q_i(z_i)}$$

上面的目标函数中对数内部没有求和项，更容易求得$\boldsymbol{\theta}$的公式解。这就是 EM 算法中“M”的含义。由于Q_i可以是任意个概率分布，实现时Q_i可以按照下面的公式计算：

$$Q_i(z_i) = \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta})}{\sum_z p(\mathbf{x}_i, z; \boldsymbol{\theta})}$$

迭代终止的判定规则是相邻两次函数值之差小于指定阈值。

收敛性的证明

假设第t次迭代时的参数值为$\boldsymbol{\theta}_t$，第$t+1$次迭代时的参数值为$\boldsymbol{\theta}_{t+1}$。如果能证明每次迭代时对数似然函数的值单调增，即：

$$L(\boldsymbol{\theta}_t) \le L(\boldsymbol{\theta}_{t+1})$$

则算法能收敛到局部极值点。由于在迭代时选择了：

$$Q_{it}(z_i) = p(z_i | \mathbf{x}_i; \boldsymbol{\theta}_t)$$

因此有：

$$\frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta})}{Q_i(z_i)} = \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta})}{p(z_i | \mathbf{x}_i; \boldsymbol{\theta}_t)} = \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta})}{p(\mathbf{x}_i, z_i; \boldsymbol{\theta}) / p(\mathbf{x}_i; \boldsymbol{\theta})} = p(\mathbf{x}_i; \boldsymbol{\theta})$$

这和z_i无关，因此是一个常数，从而保证 Jensen 不等式可以取等号。因此有下面的等式成立：

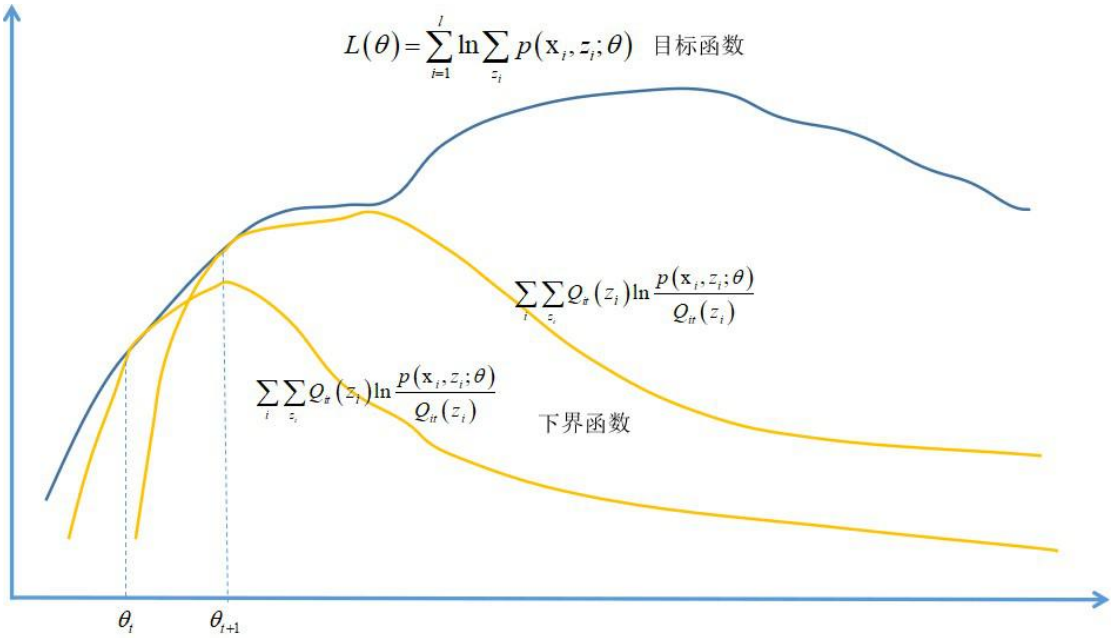
$$L(\boldsymbol{\theta}_t) = \sum_i \ln \sum_{z_i} Q_{it}(z_i) \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta}_t)}{Q_{it}(z_i)} = \sum_i \sum_{z_i} Q_{it}(z_i) \ln \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta}_t)}{Q_{it}(z_i)}$$

从而有：

$$\begin{aligned} L(\boldsymbol{\theta}_{t+1}) &\ge \sum_i \sum_{z_i} Q_{it}(z_i) \ln \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta}_{t+1})}{Q_{it}(z_i)} \\ &\ge \sum_i \sum_{z_i} Q_{it}(z_i) \ln \frac{p(\mathbf{x}_i, z_i; \boldsymbol{\theta}_t)}{Q_{it}(z_i)} \\ &= L(\boldsymbol{\theta}_t) \end{aligned}$$

上式第一步利用了 Jensen 不等式，第二步成立是因为$\boldsymbol{\theta}_{t+1}$是函数的极值，因此会大于等于任意点处的函数值，第三步在上面已经做了说明，是 Jensen 不等式取等式。上面的结论保证了每次迭代时函数值会上升，直到到达局部极大值点处，但只能保证收敛到局部极值。

下图直观的解释了 EM 算法的原理



图中的蓝色曲线为要求解的对数似然函数，黄色曲线为构造出的下界函数。首先用参数的当前估计值$\boldsymbol{\theta}_t$计算出每个训练样本的隐变量的概率分布估计值Q_t，然后用该值构造下界函数，在参数的当前估计值$\boldsymbol{\theta}_t$处，下界函数与对数似然函数的值相等（对应图中左侧第一条虚线）。然后求下界函数的极大值，得到参数新的估计值$\boldsymbol{\theta}_{t+1}$，再以当前的参数值$\boldsymbol{\theta}_{t+1}$计算隐变量的概率分布$Q_{t+1}$，构造出新的下界函数，然后求下界函数的极大值得到$\boldsymbol{\theta}_{t+2}$。如此反复，直到收敛。

算法的精髓在于：

构造下界函数（Jensen 不等式成立），通过巧妙的取Q的值而保证在参数的当前迭代点处下界函数与要求解的目标函数值相等（Jensen 不等式取等号），从而保证优化下界函数后在新的迭代点处目标函数值是上升的。

求解高斯混合模型

下面介绍 EM 算法如何求解高斯混合模型。假设有一批样本$\{\mathbf{x}_1,...,\mathbf{x}_l\}$。为每个样本$\mathbf{x}_i$增加一个隐变量$z_i$，表示样本来自于哪个高斯分布。这是一个离散型的随机变量，取值范围为$\{1,...,k\}$，取每个值的概率为w_i。$\mathbf{x}$和z的联合概率可以写成

$$\begin{aligned}p(\mathbf{x},z=j)&=p(z=j)p(\mathbf{x}|z=j)\\&=w_jN(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\end{aligned}$$

这是样本的隐变量取值为j，并且样本向量值为$\mathbf{x}$的概率。在 E 步构造Q函数

$$Q_i(z_i=j)=q_{ij}=\frac{p(\mathbf{x}_i,z_i=j;\theta)}{\sum_z p(\mathbf{x}_i,z;\theta)}$$
$$=\frac{w_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_{t=1}^{k} w_t N(\mathbf{x};\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}$$

这个值根据 $\boldsymbol{\mu},\boldsymbol{\Sigma},\mathbf{w}$ 的当前迭代值计算，是一个常数。得到 z 的分布即 Q 值之后，要求解的目标函数为

$$L(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\Sigma})=\sum_i\sum_{z_i} Q_i(z_i)\ln\frac{p(\mathbf{x}_i,z_i;\boldsymbol{\theta})}{Q_i(z_i)}$$
$$=\sum_{i=1}^{l}\sum_{j=1}^{k} q_{ij}\ln\frac{w_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{q_{ij}}$$
$$=\sum_{i=1}^{l}\sum_{j=1}^{k} q_{ij}\ln\frac{w_j\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)\right)}{q_{ij}}$$
$$=\sum_{i=1}^{l}\sum_{j=1}^{k} q_{ij}\left(\ln\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}q_{ij}}+\ln w_j-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)\right)$$

在这里 q_{ij} 已经是一个常数而不是 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的函数。对 $\boldsymbol{\mu}_j$ 求梯度并令梯度为 $\mathbf{0}$，可以得到

$$\nabla_{\boldsymbol{\mu}_j}L(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\Sigma})=\nabla_{\boldsymbol{\mu}_j}\sum_{i=1}^{l}\sum_{j=1}^{k} q_{ij}\left(\ln\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}q_{ij}}+\ln w_j-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)\right)$$
$$=-\sum_{i=1}^{l} q_{ij}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)=\mathbf{0}$$

可以解得

$$\boldsymbol{\mu}_j=\frac{\sum_{i=1}^{l} q_{ij}\mathbf{x}_i}{\sum_{i=1}^{l} q_{ij}}$$

对 $\boldsymbol{\Sigma}_j$ 求梯度并令梯度为 $\mathbf{0}$，根据正态分布最大似然估计的结论，可以解到

$$\boldsymbol{\Sigma}_j=\frac{\sum_{i=1}^{l} q_{ij}(\mathbf{x}_i-\boldsymbol{\mu}_j)(\mathbf{x}_i-\boldsymbol{\mu}_j)^{\mathrm{T}}}{\sum_{i=1}^{l} q_{ij}}$$

最后处理 $\mathbf{w}$。上面的目标函数中，只有 $\ln w_j$ 和 $\mathbf{w}$ 有关，因此可以简化。由于 w_i 有等式约束 $\sum_{i=1}^{k} w_i = 0$，因此构造拉格朗日乘子函数

$$L(\mathbf{w},\lambda) = \sum_{i=1}^{l}\sum_{j=1}^{k} q_{ij} \ln w_j + \lambda\left(\sum_{j=1}^{k} w_j - 1\right)$$

对 $\mathbf{w}$ 求梯度并令梯度为 $\mathbf{0}$，可以得到下面的方程组

$$\sum_{i=1}^{l}\sum_{j=1}^{k} \frac{q_{ij}}{w_j} + \lambda = 0$$

$$\sum_{i=1}^{k} w_i = 1$$

最后解得

$$w_j = \frac{1}{l}\sum_{i=1}^{l} q_{ij}$$

由此得到求解高斯混合模型的 EM 算法流程。首先初始化 $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{w}$，接下来循环进行迭代，直至收敛，每次迭代时的操作为：

E 步，根据模型参数的当前估计值，计算第 i 个样本来自第 j 个高斯分布的概率：

$$q_{ij} = p\left(z_i = j \middle| \mathbf{x}_i; \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

M 步，计算模型的参数。权重的计算公式为：

$$w_j = \frac{1}{l}\sum_{i=1}^{l} q_{ij}$$

均值的计算公式为：

$$\mu_j = \frac{\sum_{i=1}^{l} q_{ij}\mathbf{x}_i}{\sum_{i=1}^{l} q_{ij}}$$

协方差的计算公式为：

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{l} q_{ij}\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^{\mathrm{T}}}{\sum_{i=1}^{l} q_{ij}}$$

参考文献

[1] Arthur P Dempster, Nan M Laird, Donald B Rubin. Maximum Likelihood from Incomplete Data

via the EM Algorithm. Journal of the royal statistical society series b-methodological, 1976.