EM 算法(Expectation Maximization Algorithm)

0 前言

EM 算法是一种迭代算法，1977 年由 Dempter 等人总结提出，用于含有隐变量（hidden variable）的概率模型参数的极大似然估计，或极大后验概率估计。EM 算法的每次迭代由两步组成：E 步，求期望(expectation)；M 步，求极大(maximization)。所以这一算法称为期望-极大算法（Expectation Maximization Algorithm），简称 EM 算法。

1 基础知识

1.1 凸函数&凹函数

凸函数[1]**：**凸函数是一个定义在某个向量空间的凸集 C（区间）上的实值函数 f，在其定义域 C 上的任意两点 x、y，以及 t∈[0,1]，有

$$\mathrm{f}(\mathrm{t}*\mathrm{x}+(1-\mathrm{t})*\mathrm{y}) \leq \mathrm{t}*\mathrm{f}(\mathrm{x})+(1-t)*f(y)$$

如果对于任意的 t∈(0,1)有

$$\mathrm{f}(\mathrm{t}*\mathrm{x}+(1-\mathrm{t})*\mathrm{y}) < \mathrm{t}*\mathrm{f}(\mathrm{x})+(1-t)*f(y)$$

则函数 f 是严格凸的。

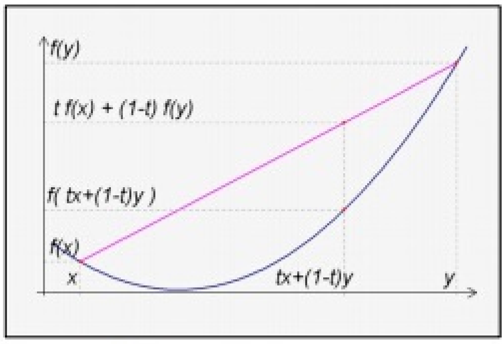


Fig1：凸函数实例

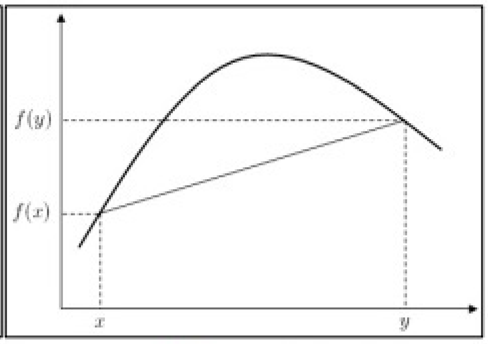


Fig2：凹函数实例

凹函数[2]**：**在数学当中，凹函数是凸函数的相反。凹函数是一个定义在某个向量空间的凹集 C（区间）上的实值函数 f，在其定义域 C 上的任意两点 x、y，以及 t∈[0,1]，有

$$\mathrm{f}(\mathrm{t}*\mathrm{x}+(1-\mathrm{t})*\mathrm{y}) \geq \mathrm{t}*\mathrm{f}(\mathrm{x})+(1-t)*f(y)$$

进一步地，对于凸函数而言，从图像上来看，可以概括为，任意两点的连线在函数曲线的上方。一元可微函数在某个区间上是凸的，当且仅当它的导数在该区间上单调不减，即有$\mathrm{f}''(\mathrm{x}) \geq 0$。当 x 是向量时，如果其 hessian 矩阵 H 是半正定的（H ≥ 0），那么 f 是凸函数。如果$\mathrm{f}''(\mathrm{x}) > 0$或者$\mathrm{H} > 0$，那么称 f 是严格凸函数[5]。

显然，log 函数为凹函数（在 EM 算法将用到）。

1.2 期望(expectation)[3]

在概率论和统计学中，一个离散性随机变量的期望值（或数学期望、或均值，亦简称期望，物理学中称为期待值）是试验中每次可能结果的概率乘以其结果的总和。

采用形式化定义，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 $P(X=x_k)=p_k, k=1,2,\cdots$，若 $\sum_{k=1}^{\infty} g(x_k)p_k$ 绝对收敛，则期望值计算为 $E[Y]=E[g(X)]=\sum_{k=1}^{\infty} g(x_k)p_k$。

（2） X 是连续型随机变量，存在相应的概率密度函数 $f(x)$，若积分 $\int_{-\infty}^{+\infty} g(x)f(x)d$ 绝对收敛，则期望值计算为 $E[Y]=E[g(X)]=\int_{-\infty}^{+\infty} g(x)f(x)dx$。

1.3 Jensen 不等式[4] [5]

Jensen 不等式：如果 f 是凸函数，X 是随机变量，则 $E[f(X)]\geq f(E[X])$，此式等价于 $\sum_{i=1}^{n} p_i f(x_i) \geq f(\sum_{i=1}^{n} p_i x_i)$，其中 $\sum_{i=1}^{n} p_i = 1$。由此可以认为，上文对于凸函数与凹函数定义所用的表达式可以看作 Jensen 的特殊形式。

如果 f 是凹函数，则 $E[f(X)]\leq f(E[X])$。（在 EM 算法中 f(x)为 log 函数，将用到此不等式，特此强调）

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

用图形表示如下：图中实线 f 表示的是凸函数，X 是随机变量。

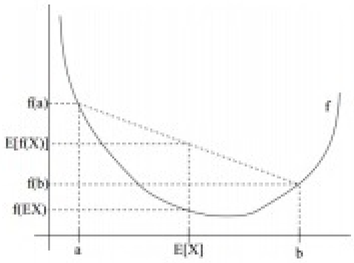


Fig3：Jensen 不等式实例

对于 Jensen 不等式的证明可以参考本文参考资料[4]，在此不做详细证明。

1.4 最大似然估计（MLE）

最大似然估计（MLE）是一种模型参数估计的常用方法。这一方法的使用情境常常是这样的，对于给定的模型以及已经观察到的样本值 $x_1, x_2, \dots x_n$，依据已

2

经得到的样本的观察值$x_1, x_2, \dots x_n$，来估计该模型的参数θ。而其中蕴含的一个直观的想法是：在给定模型下，现在已经得到样本值$x_1, x_2, \dots x_n$，这表示取得这一观察值的概率比较大，而我们所估计出来的参数，正是为了使现有观察情况出现的可能性最大。要特别说明的是，最大似然估计这一方法中，有一个很重要的假设，就是所使用的样本之间是满足独立同分布的。

最大似然估计的一般求解过程可以总结为如下步骤：

（1） 写出似然函数；

（2） 对似然函数取对数，并整理；

（3） 求导数，令导数为 0，得到似然方程；

（4） 解似然方程，得到的参数即为所求。

2 EM 算法

2.1 算法引出

（三硬币模型）[6] 假设有 3 枚硬币，分别记做 A，B，C。这些硬币正面出现的概率分别是π，p 和 q。进行如下掷硬币实验：先掷硬币 A，根据其结果选出硬币 B 或 C，正面选 B，反面选硬币 C；然后投掷选中的硬币，出现正面记作 1，反面记作 0；独立地重复 n 次（这里，n=10），结果为

1,1,0,1,0,0,1,0,1,1

假设只能观测到投掷硬币的结果，不能观测投掷硬币的过程。问如何估计三硬币正面出现的概率，即三硬币模型的参数π，p 和 q。

写出生成一个硬币时的概率：

$$P(y|\theta) = \sum_z P(y, z|\theta) = \sum_z P(z|\theta)P(y|z, \theta)$$

$$= \pi p^y(1-p)^{1-y} + (1-\pi)q^y(1-q)^{1-y}$$

其中，随机变量 y（y=1 或 0）是观测变量，表示一次实验观测的结果是 1 还是 0；随机变量 z 是隐变量，表示未观测到的掷硬币 A 的结果；$\theta = (\pi, p, q)$表模型参数。

将观测数据表示为 Y，未观测数据表示为 Z，则观测数据的似然函数为

$$P(Y|\theta) = \sum_Z P(Y, Z|\theta)$$

为方便计算，取其对数形式

$$L(\theta|Y) = log\, P(Y|\theta) = log \sum_z P(Y, Z|\theta)$$

那么，求模型参数的极大似然估计，即

$$\theta^* = \arg max_\theta\, L(\theta|Y) = \arg max_\theta\, log \sum_z P(Y, Z|\theta)$$

3

然而想用这个 log 套 Σ 的形式直接求解θ往往非常困难。那么这一问题只有通过迭代的方法求解，EM 算法就是此处所采用的迭代算法。到这里已经说明了 EM 算法的使用情景，以及为何需要通过 EM 算法求解（关键在于隐变量 Z 的存在）。

PS1：对于以上问题的参数估计，如果已经知道隐含变量 Z，令 W={Y,Z}，此时相当于 W 为完全已知的变量。问题转化为如下形式

$$L(\theta|Y,Z) = logP(Y,Z|\theta) = logP(W|\theta)$$

避免了 log 套 Σ 的形式，成为只有已知变量的模型参数估计，通过极大似然估计法即可求解。但是实际上 Z 的值并无法直接获得准确值，就难以直接通过以上的方法求解。然而可以在给定θ时求 Z 的后验概率$P(Z|Y,\theta)$，通过概率计算，进而将问题转化为

$$L(\theta|Y,Z) = \log P(Y,Z|\theta) = \sum_{Z} logP(Y,Z|\theta)\,P(Z|Y,\theta)$$

的形式。这相当于用$logP(Y,Z|\theta)$在给定 Y 与θ情况下的期望，来近似表达$\log P(Y,Z|\theta)$，这也是 EM 算法 E 步所做的（个人认为）。这是后话，在 EM 算法中会再说明。

PS2：再次说明 EM 算法的使用情景。EM 算法适用于含有隐变量或潜在变量（如以上问题的 z）的概率模型参数的估计。若不然，即所给问题只有观测变量，则可以通过极大似然估计获得模型参数。

2.2 EM 算法描述

算法前述[6]

一般地，用 Y 表示观测变量的数据，Z 表示隐含变量的数据。Y 和 Z 连在一起就是完全数据，观测数据又称为不完全数据。假设给定观测数据 Y，其概率分布是$P(Y|\theta)$，其中θ是需要估计的模型参数，那么不完全数据的似然函数是$P(Y|\theta)$，其对数似然函数是$L(\theta|Y) = \log P(Y|\theta)$；假设 Y 和 Z 的联合概率分布为$P(Y,Z|\theta)$，那么完全数据的对数似然函数是$\log P(Y,Z|\theta)$。

EM 算法通过迭代求$L(\theta|Y,Z) = \log P(Y,Z|\theta)$的极大似然估计。每次迭代包括两步：E 步，求当前参数下，完全数据的对数似然函数的期望；M 步，极大化 E 步的结果，获得新一轮的模型参数值。下面将给出 EM 算法的具体描述。

算法流程[6]

输入：观测变量数据 Y，隐变量数据 Z，联合分布$P(Y,Z|\theta)$，条件分布$P(Z|Y,\theta)$；

输出：模型参数θ。

（1）选择参数的初值$\theta^{(0)}$，开始迭代；

（2）E 步：记$\theta^{(i)}$为第 i 次迭代参数θ的估计值，在第 i+1 次迭代的 E 步，计算

$$E_{Z|Y,\theta^{(i)}}[L(\theta|Y,Z)] = E_{Z|Y,\theta^{(i)}}[\log P(Y,Z|\theta)]$$

4

$$= \sum_{Z} logP(Y,Z|\theta)\, P(Z|Y,\theta^{(i)})$$

这里，$P(Z|Y,\theta^{(i)})$ 是在给定观测数据 Y 和当前参数估计$\theta^{(i)}$下隐变量 Z 的条件概率分布。这一步所列的函数是完全数据的对数似然函数$\log P(Y,Z|\theta)$，在给定观测数据 Y 和当前参数$\theta^{(i)}$下对隐变量数据 Z 的条件概率分布$P(Z|Y,\theta^{(i)})$的期望，这一期望称为 Q 函数，下文会做进一步说明。

（3）M 步：求使上式极大化的θ，确定第 i+1 次迭代的参数的估计值$\theta^{(i+1)}$

$$\theta^{(i+1)} = \mathrm{arg}max_{\theta} E_{Z|Y,\theta^{(i)}}[\log P(Y,Z|\theta)]$$

（4）重复第（2）步和第（3）步，直到收敛。

算法说明[7]

EM 算法每次迭代都建立在上一轮迭代 M 步所得到的参数θ的最优值的估计$\theta^{(i)}$上，通过它可以得到 Z 的后验概率$P(Z|Y,\theta^{(i)})$，进而在 E 步求出$L(\theta|Y,Z) = \log P(Y,Z|\theta)$在分布 Z～$P(Z|Y,\theta^{(i)})$的期望$E_{Z|Y,\theta^{(i)}}[L(\theta|Y,Z)]$；在此基础上，M 步通过最大化这一期望得到新一轮的参数θ。如此往复，直至收敛。

前面已经说到，$\mathrm{arg}max_{\theta}\, L(\theta|Y,Z)$在 Z 不确定的情况下难以直接计算。Z 的值虽然不确定，但是在给定 Y 和θ的条件下 Z 的条件概率分布是可以获得的，利用这一概率分布，我们可以得到$L(\theta|Y,Z)$的期望$E_{Z|Y,\theta^{(i)}}[L(\theta|Y,Z)]$。于是 EM 算法通过最大化它的期望$E_{Z|Y,\theta^{(i)}}[L(\theta|Y,Z)]$来逼近θ的最优值，得到$\theta^{(i+1)}$。注意由于$L(\theta|Y,Z)$的这个期望是在 Z 的一个分布上求的，这样得到的表达式就只剩下θ一个未知量，因而绕过了 Z 未知的问题。而$\theta^{(i+1)}$又可以作为下一轮迭代的基础，继续向最优逼近。算法中 E 步就是在利用$\theta^{(i)}$求期望$E_{Z|Y,\theta^{(i)}}[L(\theta|Y,Z)]$，这就是所谓“Expectation”；M 步就是通过寻找$\theta^{(i+1)}$最大化这个期望来逼近θ的最优值，这就叫“Maximization”。EM 算法因此得名。

至此 EM 算法过程介绍完毕，下面进一步介绍 E 步中所提到的 Q 函数以及 EM 算法每一步的几点说明。

Q 函数[6]

完全数据的对数似然函数$\log P(Y,Z|\theta)$关于在给定观测数据 Y 和当前参数$\theta^{(i)}$下对隐变量数据 Z 的条件概率分布$P(Z|Y,\theta^{(i)})$的期望称为 Q 函数，即

5

$$Q(\theta,\theta^{(i)}) = E_{Z|Y,\theta^{(i)}}[\log P(Y,Z|\theta)]$$

其中$Q(\theta,\theta^{(i)})$的第一个变元表示要极大化的参数，第二个变元表示参数的当前估计值。Q 函数是 EM 算法的核心，每次迭代实际在求 Q 函数及其最大。

补充说明[6]

（1）参数的初值可以任意选择，但需要注意的是，EM 算法对初值敏感。

（2）算法的每次迭代使似然函数增大或达到局部极值。EM 算法不能保证找到全局最优值。

（3）给出算法迭代停止的条件，一般是对较小的正数ε_1，ε_2，若满足

$$\left|\theta^{(i+1)}-\theta^{(i)}\right| < \varepsilon_1 \text{ 或 } \left|Q(\theta^{(i+1)},\theta^{(i)}) - Q(\theta^{(i)},\theta^{(i)})\right| < \varepsilon_2$$

则停止迭代。

2.3 EM 算法的原理[6] [7]

对于给定的问题，已知的是观测数据 Y，那么按照极大似然的思想，我们要做的实际上是极大化观测数据 Y 关于参数θ的对数似然函数，即极大化

$$L(\theta|Y) = log\, P(Y|\theta) = log \sum_z P(Y,Z|\theta)$$

观察 EM 算法的执行过程，可以看到，我们是通过不断地迭代计算

$$\arg max_\theta E_{Z|Y,\theta^{(i)}}[\log P(Y,Z|\theta)]$$

来逼近θ的最优值的，然而为什么这样做是有效的？换言之，为什么迭代计算$\arg max_\theta E_{Z|Y,\theta^{(i)}}[\log P(Y,Z|\theta)]$的过程，可以做到对对数似然函数$L(\theta|Y)$的优化？

事实上，每次迭代得到的$\theta^{(i+1)}$一定比$\theta^{(i)}$更优，算法的迭代过程是对θ最优值的单调逼近。下面进行检验的推导

$$L(\theta|Y) = log\, P(Y|\theta) = log \sum_z P(Y,Z|\theta)$$

$$= \log \sum_Z P(Z|Y,\theta^{(i)}) \frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(i)})}$$

$$\geq \sum_Z P(Z|Y,\theta^{(i)}) \log \frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(i)})}$$

以上推导中，引入$P(Z|Y,\theta^{(i)})$方便推导。并且在最后一步利用了 Jensen 不等式，注意$\sum_Z P(Z|Y,\theta^{(i)}) = 1$是满足的。另外 Jensen 不等式的引入解除了 log 内套

6

∑的形式，也是计算容易一些。

通过以上也可以看出，$\sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(i)})}$是$L(\theta|Y)$的一个下界，那么任何可以使$\sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(i)})}$增大的$\theta$，也可以使$L(\theta|Y)$增大。为了使$L(\theta|Y)$尽可能的增长，选择$\theta^{(i+1)}$使$\sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(i)})}$达到极大，则

$$\theta^{(i+1)} = \arg max_\theta \sum_Z P(Z|Y,\theta^{(i)})\log\frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(i)})}$$

$$= \arg max_\theta \sum_Z [P(Z|Y,\theta^{(i)}) logP(Y,Z|\theta) - P(Z|Y,\theta^{(i)})\log P(Z|Y,\theta^{(i)})]$$

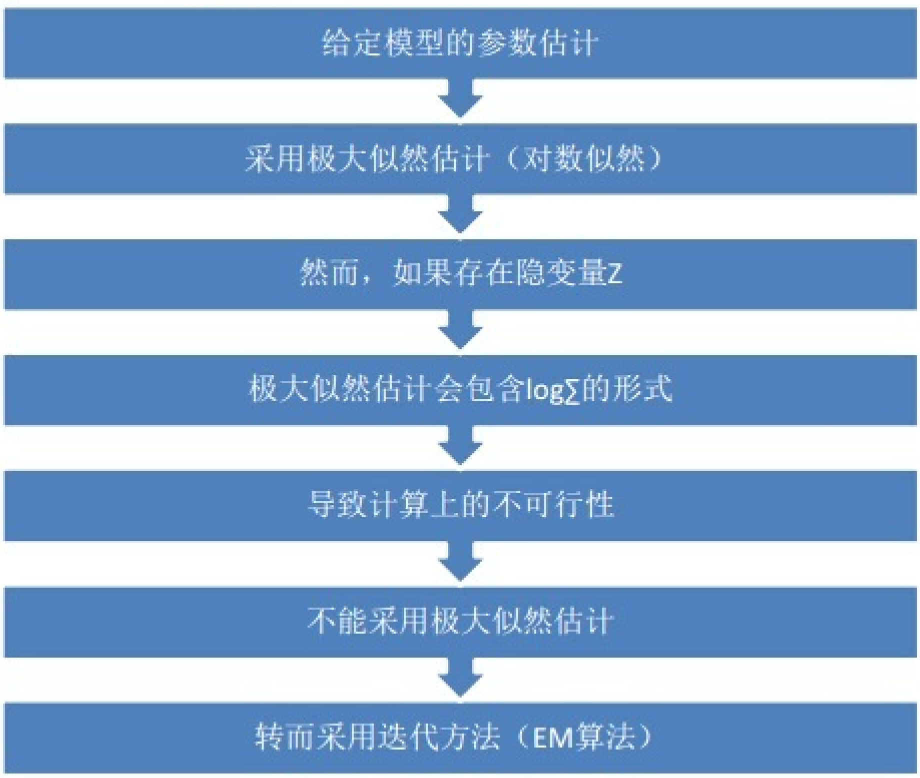
$$= \arg max_\theta \sum_Z P(Z|Y,\theta^{(i)}) logP(Y,Z|\theta)$$

$$= \arg max_\theta E_{Z|Y,\theta^{(i)}}[\log P(Y,Z|\theta)]$$

其中倒数第二步因为$-P(Z|Y,\theta^{(i)})\log P(Z|Y,\theta^{(i)})$这一项与$\theta$无关，可以直接去掉，这样也就得到 EM 算法中的形式，从而证明了 EM 算法所用的途径是有效的。

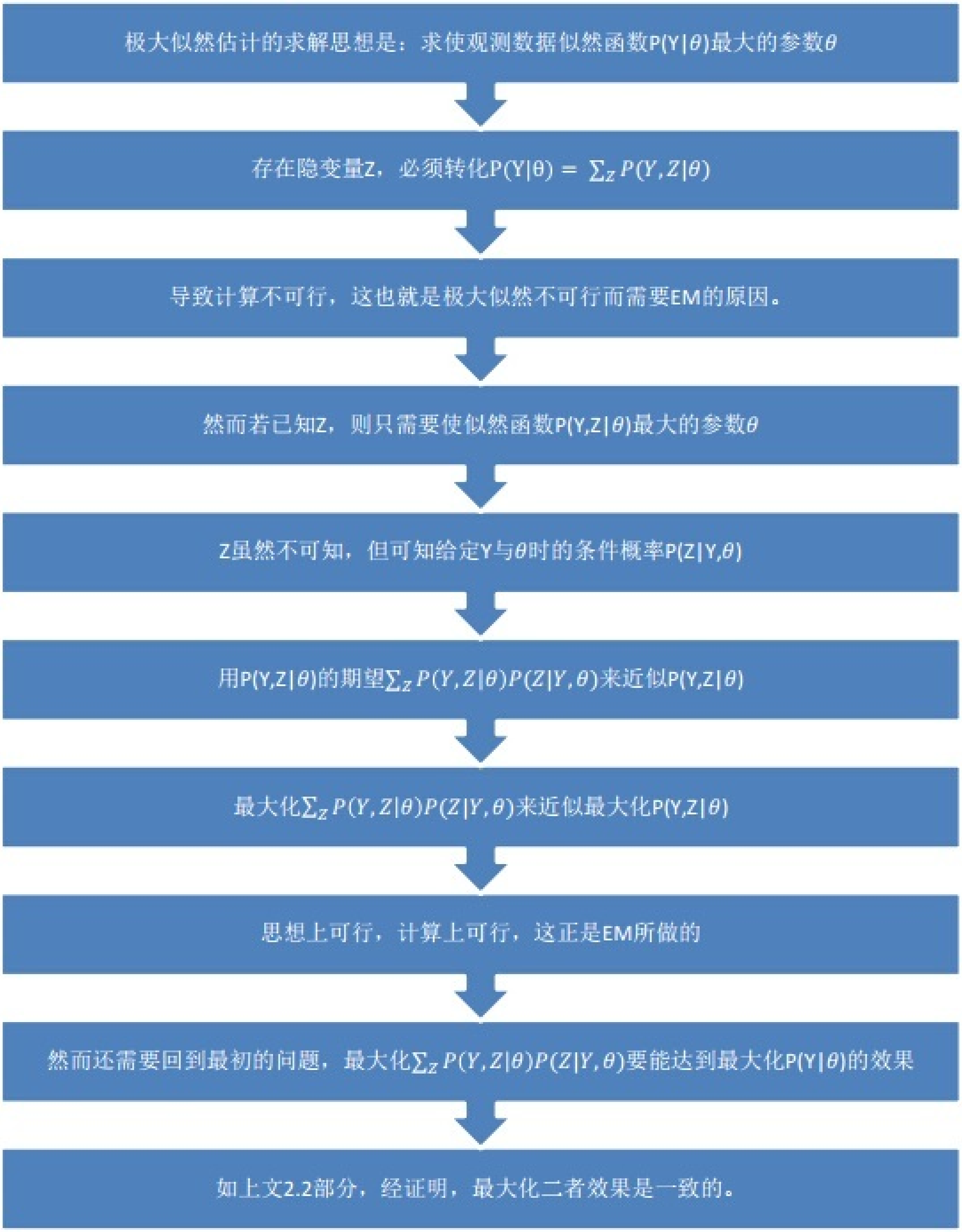
2.3 概括总结

（1）为什么需要用 EM 算法？



7

（2）为什么 EM 算法 Q 函数定义如此，且迭代计算 Q 函数及其最大？



2.4 EM 算法的收敛性

下面给出 EM 算法收敛性的两个定理，证明从略，参考《统计学习方法》相应章节即可。

定理一：设P(Y|θ)为观测数据的似然函数，$\theta^{(i)}$(i=1,2,...)为 EM 算法得到的参数估计序列，$P(Y|\theta^{(i)})$ (i=1,2,...)为对应的似然函数序列，则$P(Y|\theta^{(i)})$是单调递

8

增的，即

$$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)})$$

定理二： 设$L(\theta) = log\,P(Y|\theta)$为观测数据的对数似然函数，$\theta^{(i)}$(i=1,2,...)为EM 算法得到的参数估计序列，$L(\theta^{(i)})$(i=1,2,...) 为对应的对数似然函数序列.

（1）如果$P(Y|\theta)$有上界，则$L(\theta^{(i)}) = log\,P(Y|\theta^{(i)})$收敛到某一值 L^*；

（2）在函数$Q(\theta,\theta^{'})$与$L(\theta)$满足一定条件下，由 EM 算法得到的参数估计序列$\theta^{(i)}$的收敛值θ^*是$L(\theta)$的稳定点。

2.5 EM 算法的应用

EM 算法在机器学习、计算机视觉和自然语言处理应用非常广泛，典型的像是聚类算法 K-means 和高斯混合模型（GMM）以及隐马尔可夫模型（HMM）。例如 HMM 的非监督学习算法 Baum-Welch 算法正是 EM 算法的应用。

对于 EM 算法的进一步学习可以参考资料[8]。

参考资料

[1] https://zh.wikipedia.org/wiki/%E5%87%B8%E5%87%BD%E6%95%B0
[2] https://zh.wikipedia.org/wiki/%E5%87%B9%E5%87%BD%E6%95%B0
[3] https://zh.wikipedia.org/wiki/%E6%9C%9F%E6%9C%9B%E5%80%BC
[4] http://blog.csdn.net/wang_yi_wen/article/details/8917396
[5] http://www.cnblogs.com/jerrylead/archive/2011/04/06/2006936.html
[6] 李航. 统计学习方法[J]. 2012.
[7] http://blog.tomtung.com/2011/10/em-algorithm/
[8] McLachlan G, Krishnan T. The EM algorithm and extensions[M]. John Wiley & Sons, 2007.