# 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。

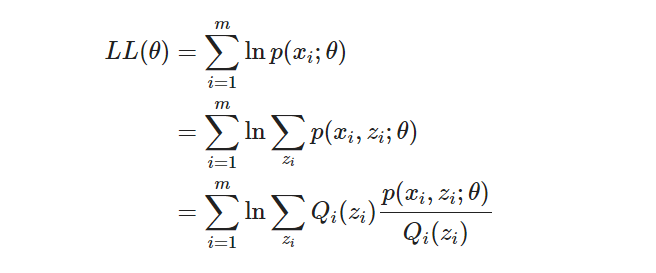
# 二、算法的推导和步骤

# 1.算法的推导：

假设现有一批独立同分布的样本数据[1]，它们是由某个含有隐变量的模型生成，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）：

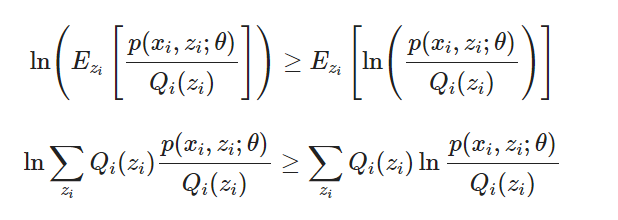
显然，此时里含有未知的隐变量z以及和（z为离散型时）或者积分（z为连续型时）的对数，因此无法按照传统方法直接求出使得达到最大值的模型参数θ，而EM算法给出了一种迭代的方法完成对的极大化，下面是具体的推导方式。

设的概率密度函数是，对上面函数进行恒等变形：

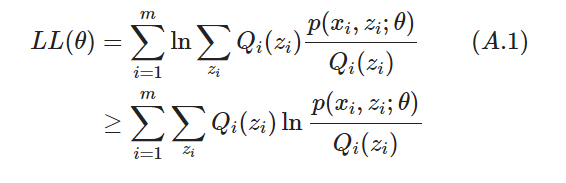


其中可以看作是对关于求期望。

那么由Jensen不等式可知：

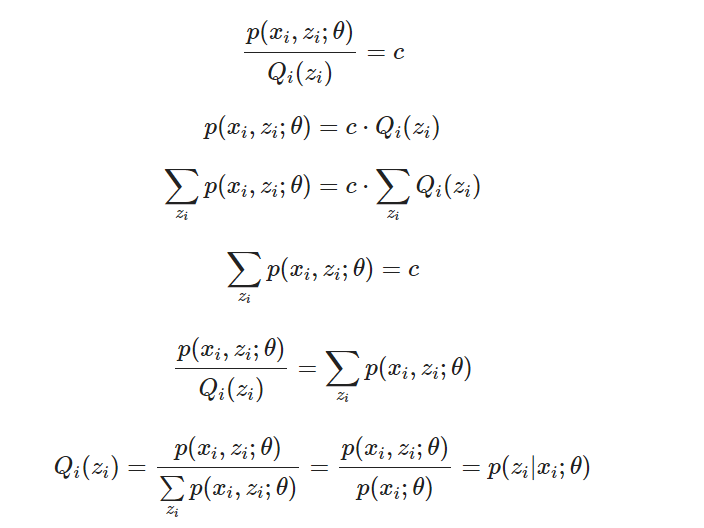


将上述不等式待人到中，有：

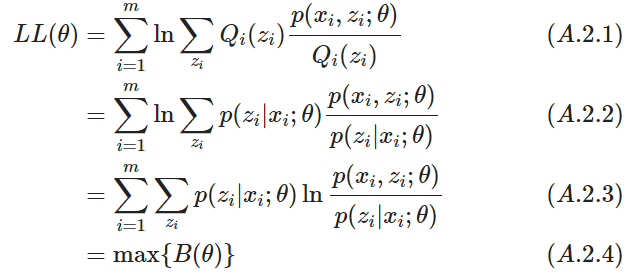


若令，此时为的下界函数，即为的上界函数，所以如果能使得=。那么取得最大值。由延森不等式可知，如果等于一个常量c，则大于等于号可以取得等号。因此，只要任意选取满足的就可以使之达到最大值。又因为为概率密度函数，故满足如下约束：

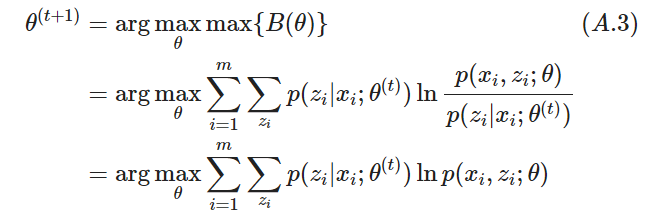
于是可以推出：



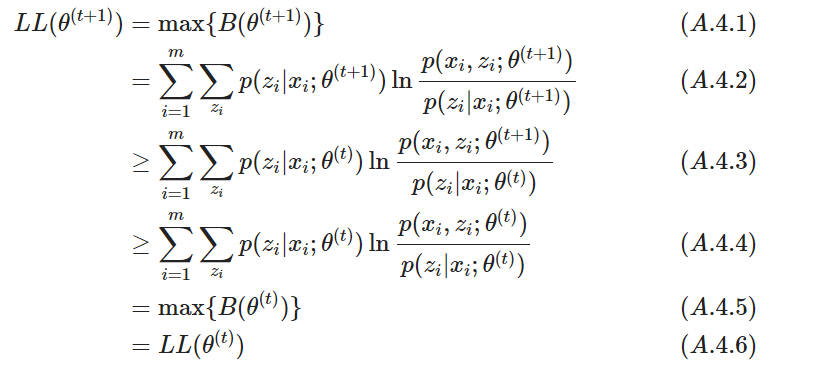
所以，当且仅当时取得最大值，将该式代入和中可以推得：



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，对数似然函数等价于其下界函数的最大值，所以要想极大化可以通过极大化max{B(θ)}来间接极大化LL(θ)，因此，下面考虑如何极大化。假设**已知**第t次迭代的参数为，而第次迭代的参数通过如下方式求得：



将代入则可以进一步推得：（这一步为收敛性的证明）



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

那么由（A.4）可知，凡是能使得达到极大的一定能使得。综上所述即可总结出EM算法的“E步”和“M步”分别为：

**E步：**令并写出；

**M步：**求使得到达极大的。

算法流程：

假设现有一批独立同分布的样本数据，它们是由某个含有隐变量的模型生成，最大迭代数为J。

# 2.算法步骤：

1. 随机初始化模型参数的初值。
2. t=1,2…J开始进行迭代

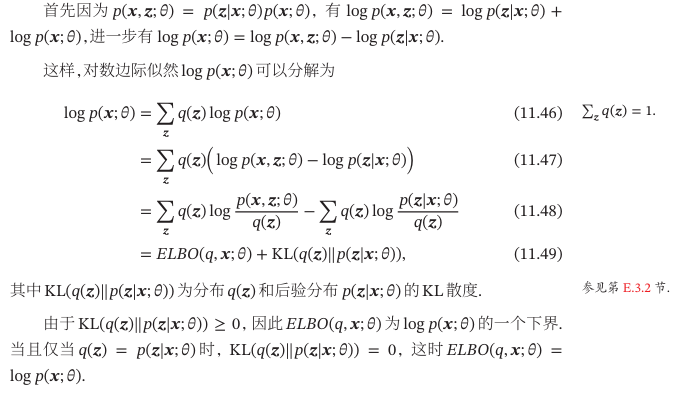
·E步计算

·M步：求使得到达极大的。

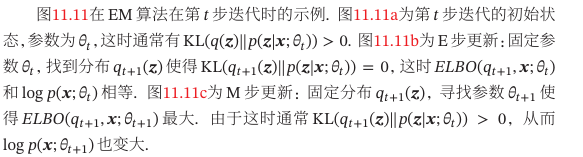
·如果已经收敛，则算法结束。否则继续进行迭代。

·输出：模型参数。

# 3.从信息论角度解释[2]：



我们的思路是：假如KL散度这一项为0，则这一项就和ELBO项相等。而ELBO项是关于的函数，我们找到使ELBO项达到最大的argmax *，*并不断迭代下去。



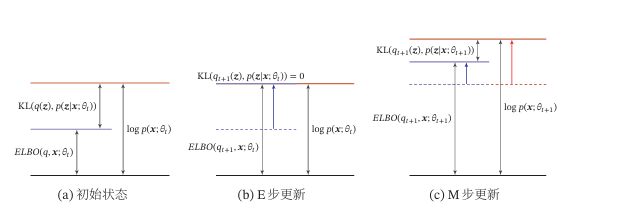
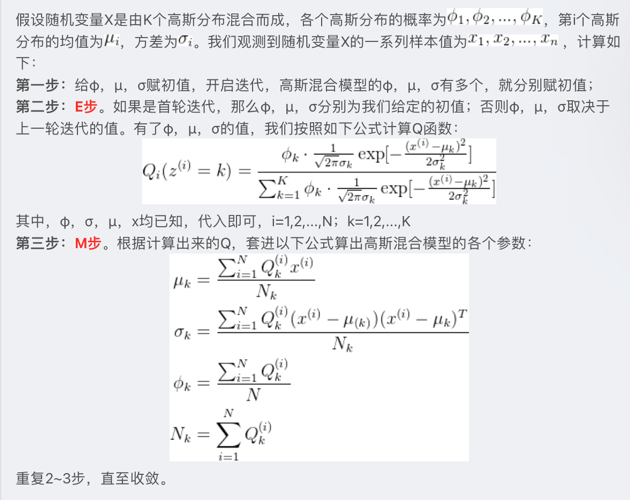


图11-11

# 4.对于高斯混合模型



# 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：

成绩来自二班的同学：

现在我们假设、、、为已知量。且假设,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：

第个同学来自一班的概率为：i=1…..100

第个同学来自二班的概率为：

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。

代表第一个同学来自一班的概率，代表最后一个同学来自一班的概率。

注意上面是方差。

和同理。

并且先验概率：

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](https://blog.csdn.net/qq_41554005/article/details/100591525)

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