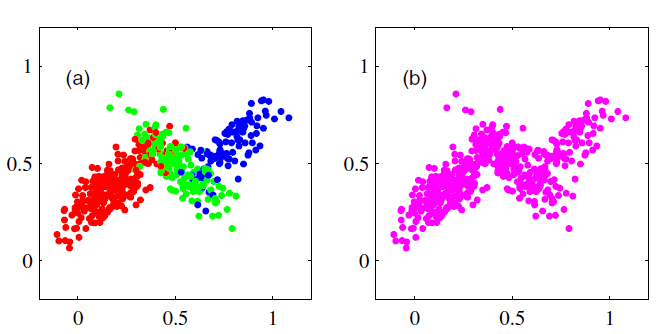
我所理解的EM

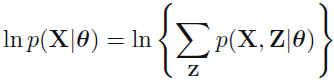
By erikhu(胡繁星)

# 1 为什么要EM？什么情况下EM？

很多情况下，概率模型是含有隐变量的。隐变量是指我们观察不到的变量。比如混合高斯模型下，我们能观察到右边的事件，也叫不完全数据；左边是完全的事件，也叫完全数据。



一般情况下，我们用X表示观察变量（显示变量），Z表示隐藏变量，表示模型的参数。我们的目标是求观测变量的最大似然，

 (1)

如果我们能拿到完全数据，问题就变得简单了，最大似然直接求，

(2)

我们要区分观测变量的最大似然和完全数据的最大似然，式(1)在ln里面含有累加运算，如果直接对其求导，分母很复杂，没法直接求。我们知道一般概率分布都是指数族分布，指数族分布，形式如下，



指数分布的ln形式非常适合求导，式(2)一般服从指数族分布，所以直接求解式(2)比求解式(1)容易得多。

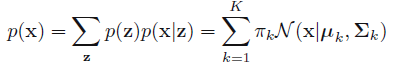
**EM算法适用的场景：含有隐藏变量**。比如混合高斯、混合多项分布、PLSA、LDA、HMM。

**为什么需要EM：因为含有隐变量的最大似然很难求解**。

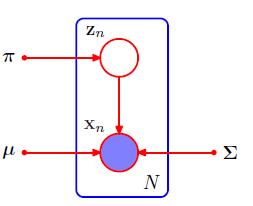
# 2 EM的直观解释（混合高斯为例）

**2.1 高斯分布的EM算法**

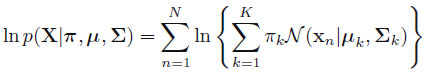
对于混合高斯分布，分布为，

 (3)

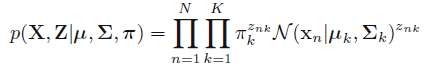
混合高斯分布的图模型表示如下，



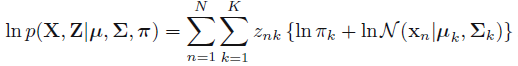
不完全数据的ln似然函数为，

 (4)

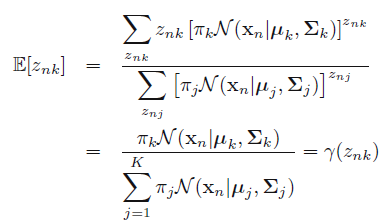
可以看到ln里面有累加操作，没法直接求解。我们再看完全数据的似然函数，

 (5)

取ln变为，

 (6)

对比(4)和(6)，我们发现，完全数据的ln似然函数非常好求解。表示第n个数据点由第k个高斯分布产生，如果是为1，否则为0。那么问题来了，我们观测不到隐变量z，怎么办？一个直观的想法，如果我们能估计出就好了。我们基于当前的知识，对最好的估计就是的后验概率。先写出的后验概率的期望，

 (7)

对于式(7)，我们把目前得到的模型参数和观察变量作为已知值，带入可求得当前模型、当前观察求得的隐变量的后验概率的期望。

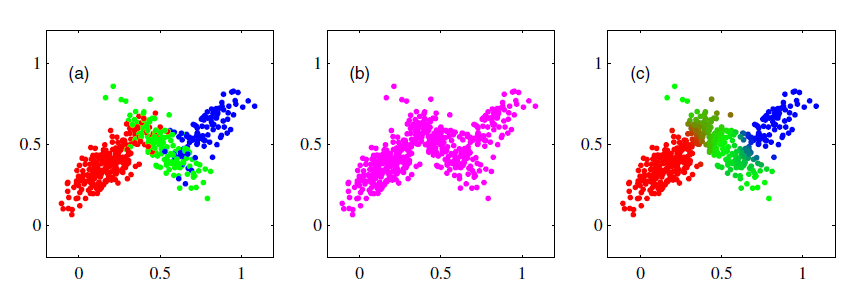
这样，我们把带入到式(6)里面，用的期望代替隐变量，得到了只含有模型参数的式子。我们把模型参数作为(6)的变量，最大化(6)，可以解出。然后在利用解出的带入式(7)重新计算的后验期望。按照这个过程一直迭代下去，直到稳定下来。

这里重新整理了高斯分布的EM算法

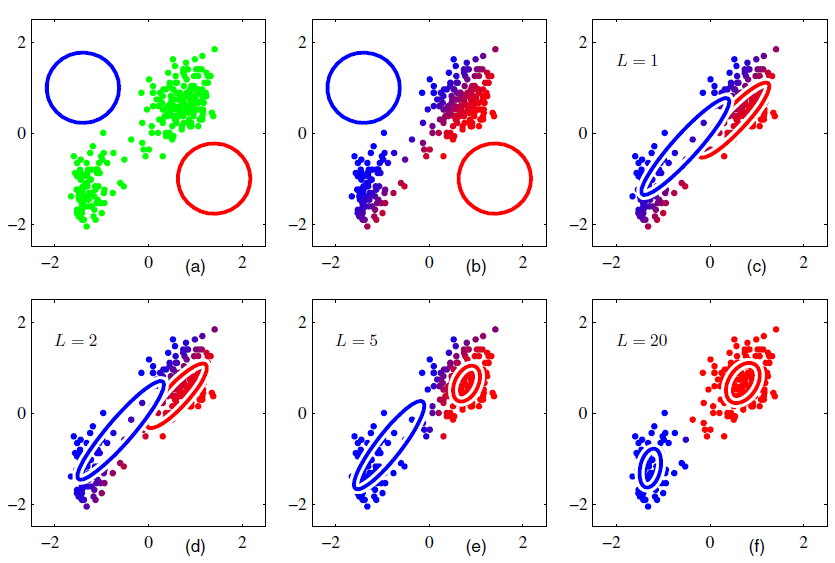
高斯分布的EM算法

|  |
| --- |
| 1. 随机给赋初值 2. E-step:把带入式(7)，求得 3. M-step:把带入下面的式子更新参数      1. 把参数带入不完全数据的ln似然函数，     看这个函数是否还在递增，同时看所有的参数是否还有变化，如果是继续转到2步骤。 |

对于混合高斯分布，我们可以在第三幅图中看到的值。



可以看到混合高斯的迭代过程，



混合高斯和k-means：

1）k-means是混合高斯的硬分类情况，无法给出分类的可信度。

2）k-means是方差矩阵为单位矩阵的混合高斯。

3）k-means可以用来做混合高斯的初值

**2.1 通用的EM算法**

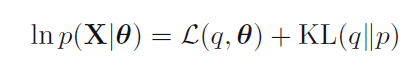
|  |
| --- |
|  |

这里分别解释一下E步和M步。E步，根据前一轮估计的模型参数，求隐变量相对于当前观测结果的后验概率。一般用贝叶斯方法，直接求解，

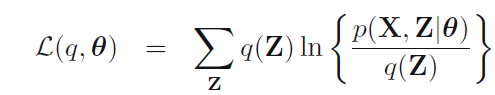
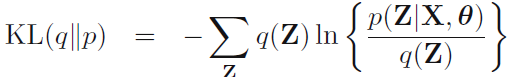
M步。重点看Q函数，Q表示的是完全数据ln似然函数相对于隐变量的期望。这样做最大的好处是完全数据的ln似然函数大部分是简单的累加形式，容易求解。

# 3 为什么EM有效？

我们把不完全数据的ln似然做分解，

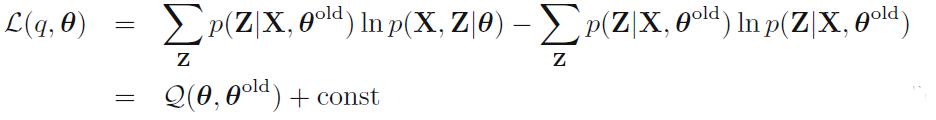
 (8)

其中，

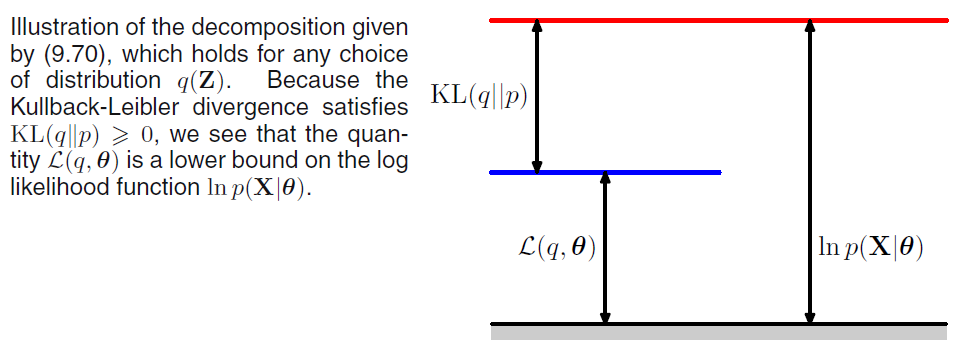
 (9) (10)

式(10)表示q(Z)和的KL散度（Kullback-Leibler divergence），也叫做这两分布的相对熵。任意两个分布的相对熵KL(q||p) >= 0，当两个分布相同的时候等号成立。

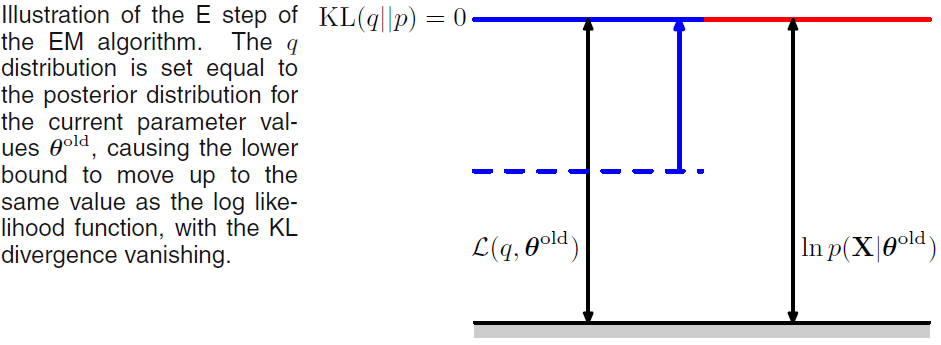
我们再看看式(9)，

(11)

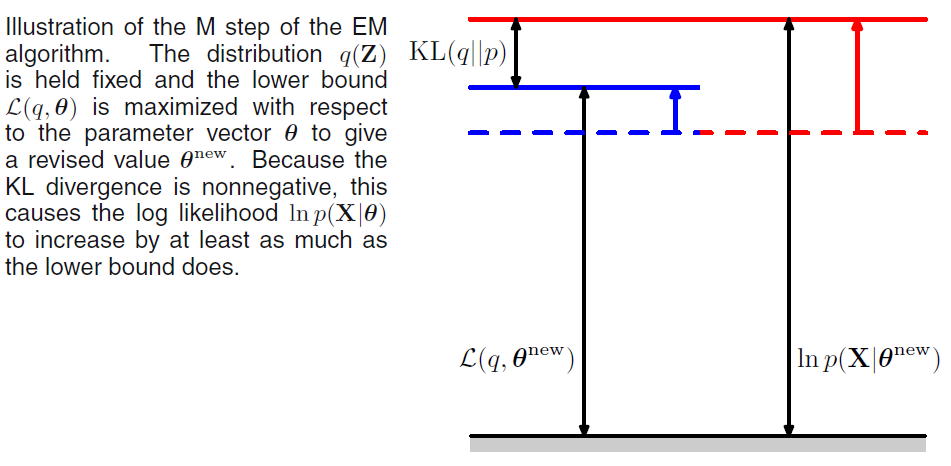
式(9)可以拆解成，Q函数和一个与无关的项（q分布的熵）。所以Q函数是的下限。我们再看看EM算法是如何提升。



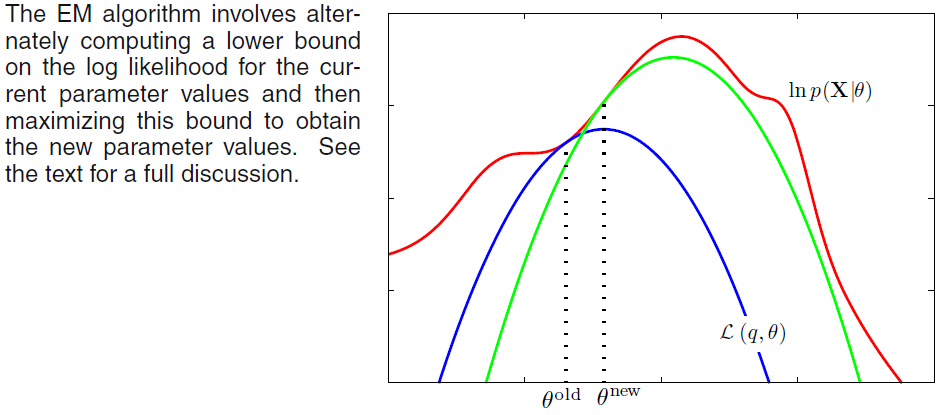
对于E步，当我们固定当前时刻的参数的时候，通过选取q函数为z的后验，使得KL项为0，从而提升L函数，使得L函数达到当前模型的下界。



对于M步，我们根据当前的q分布，最大化，提升L函数，同时KL会因为参数变化导致，不为0。所以，整体上提升，极大化模型下界。



我们可以换一个角度在看一下EM的优化过程。蓝线代表当前参数下的L函数，也就是目标函数的下界，E步的时候计算L函数，M步的时候通过重新计算得到L的最大值。



# 4 EM的应用（续）

参考文献

Bishop, Christopher M. *Pattern recognition and machine learning*. Vol. 4. No. 4. New York: springer, 2006. (prml经典书籍，第9章)

Hofmann, Thomas. "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999. (PLSA开篇之作)