上篇主要介绍了贝叶斯分类器，从贝叶斯公式到贝叶斯决策论，再到通过极大似然法估计类条件概率，贝叶斯分类器的训练就是参数估计的过程。朴素贝叶斯则是“属性条件独立性假设”下的特例，它避免了假设属性联合分布过于经验性和训练集不足引起参数估计较大偏差两个大问题，最后介绍的拉普拉斯修正将概率值进行平滑处理。本篇将介绍另一个当选为数据挖掘十大算法之一的**EM算法**。

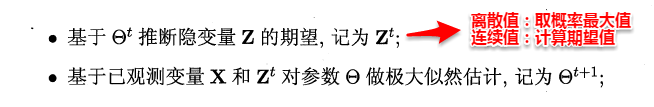
#**8、EM算法**

EM（Expectation-Maximization）算法是一种常用的估计参数隐变量的利器，也称为“期望最大算法”，是数据挖掘的十大经典算法之一。EM算法主要应用于训练集样本不完整即存在隐变量时的情形（例如某个属性值未知），通过其独特的“两步走”策略能较好地估计出隐变量的值。

##**8.1 EM算法思想**

EM是一种迭代式的方法，它的基本思想就是：若样本服从的分布参数θ已知，则可以根据已观测到的训练样本推断出隐变量Z的期望值（E步），若Z的值已知则运用最大似然法估计出新的θ值（M步）。重复这个过程直到Z和θ值不再发生变化。

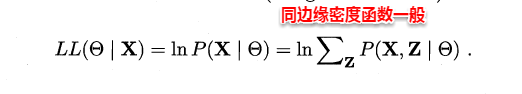
简单来讲：假设我们想估计A和B这两个参数，在开始状态下二者都是未知的，但如果知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。可以考虑首先赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。

[](https://camo.githubusercontent.com/365342a50de9d49894ca570cca2286690ba452e2/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433626635336562322e706e67)

现在再来回想聚类的代表算法K-Means：【首先随机选择类中心=>将样本点划分到类簇中=>重新计算类中心=>不断迭代直至收敛】，不难发现这个过程和EM迭代的方法极其相似，事实上，若将样本的类别看做为“隐变量”（latent variable）Z，类中心看作样本的分布参数θ，K-Means就是通过EM算法来进行迭代的，与我们这里不同的是，K-Means的目标是最小化样本点到其对应类中心的距离和，上述为极大化似然函数。

##**8.2 EM算法数学推导**

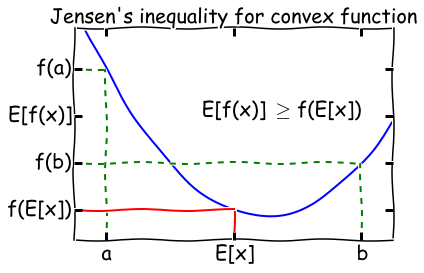
在上篇极大似然法中，当样本属性值都已知时，我们很容易通过极大化对数似然，接着对每个参数求偏导计算出参数的值。但当存在隐变量时，就无法直接求解，此时我们通常最大化已观察数据的对数“边际似然”（marginal likelihood）。

[](https://camo.githubusercontent.com/573fce6d868e7c48aa99adddbfd88a24ec7f4778/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433626664383464322e706e67)

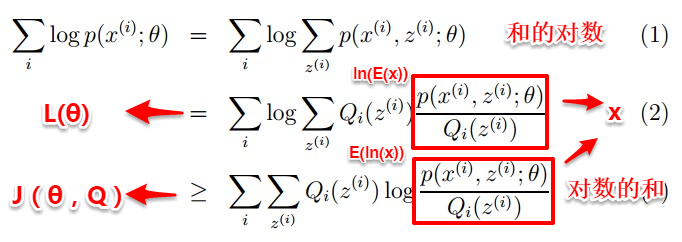
这时候，通过边缘似然将隐变量Z引入进来，对于参数估计，现在与最大似然不同的只是似然函数式中多了一个未知的变量Z，也就是说我们的目标是找到适合的θ和Z让L(θ)最大，这样我们也可以分别对未知的θ和Z求偏导，再令其等于0。

然而观察上式可以发现，和的对数（ln(x1+x2+x3)）求导十分复杂，那能否通过变换上式得到一种求导简单的新表达式呢？这时候 Jensen不等式就派上用场了，先回顾一下高等数学凸函数的内容：

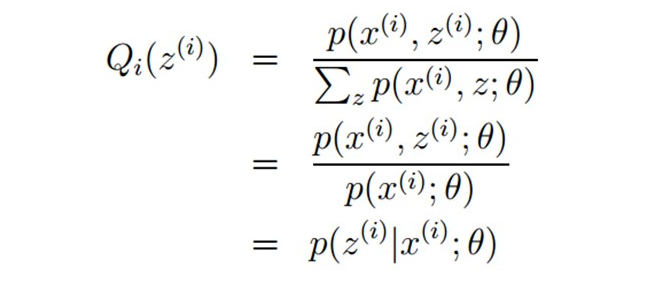
**Jensen's inequality**：过一个凸函数上任意两点所作割线一定在这两点间的函数图象的上方。理解起来也十分简单，对于凸函数f(x)''>0，即曲线的变化率是越来越大单调递增的，所以函数越到后面增长越厉害，这样在一个区间下，函数的均值就会大一些了。

[](https://camo.githubusercontent.com/5863a5f78d84c374c83378dc3c979290ee976051/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433633036346337322e706e67)

因为ln(\*)函数为凹函数，故可以将上式“和的对数”变为“对数的和”，这样就很容易求导了。

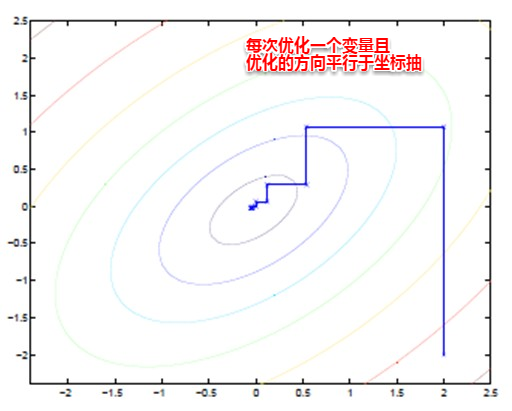
[](https://camo.githubusercontent.com/850a47b6640cf5b6372903c202188bf73d8904a4/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433633334393061642e706e67)

接着求解Qi和θ：首先固定θ（初始值），通过求解Qi使得J（θ，Q）在θ处与L（θ）相等，即求出L（θ）的下界；然后再固定Qi，调整θ，最大化下界J（θ，Q）。不断重复两个步骤直到稳定。通过jensen不等式的性质，Qi的计算公式实际上就是后验概率：

[](https://camo.githubusercontent.com/35170329143d709ebb5313443a29ae4aa8530d25/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433633231323736632e706e67)

通过数学公式的推导，简单来理解这一过程：固定θ计算Q的过程就是在建立L（θ）的下界，即通过jenson不等式得到的下界（E步）；固定Q计算θ则是使得下界极大化（M步），从而不断推高边缘似然L（θ）。从而循序渐进地计算出L（θ）取得极大值时隐变量Z的估计值。

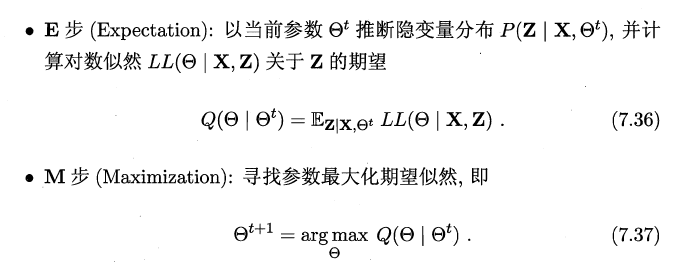
EM算法也可以看作一种“坐标下降法”，首先固定一个值，对另外一个值求极值，不断重复直到收敛。这时候也许大家就有疑问，问什么不直接这两个家伙求偏导用梯度下降呢？这时候就是坐标下降的优势，有些特殊的函数，例如曲线函数z=y^2+x^2+x^2y+xy+...，无法直接求导，这时如果先固定其中的一个变量，再对另一个变量求极值，则变得可行。

[](https://camo.githubusercontent.com/30c6379d31f5d73e532ae0f93c47e54f44efdf73/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433633334653766662e706e67)

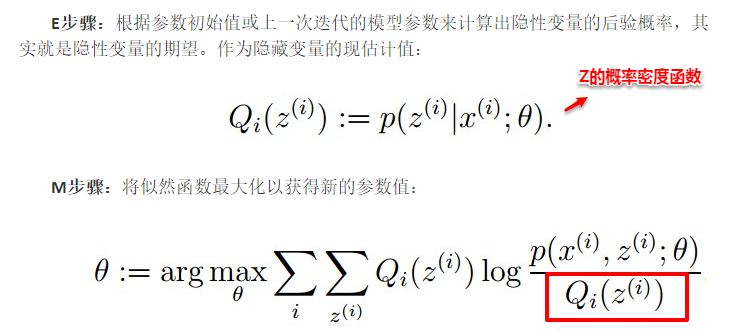
##**8.3 EM算法流程**

看完数学推导，算法的流程也就十分简单了，这里有两个版本，版本一来自西瓜书，周天使的介绍十分简洁；版本二来自于大牛的博客。结合着数学推导，自认为版本二更具有逻辑性，两者唯一的区别就在于版本二多出了红框的部分，这里我也没得到答案，欢迎骚扰讨论~

**版本一：**

[](https://camo.githubusercontent.com/4bb8d3784587ec5171b653ba08b4a3e6f10a315f/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433633065313964622e706e67)

**版本二：**

[](https://camo.githubusercontent.com/47ecbd81bcb345df2e8711b06456ce173cdb598b/68747470733a2f2f692e6c6f6c692e6e65742f323031382f31302f31382f356263383433633334373735622e706e67)