个人资料

zouxy09

从最大似然到EM算法浅解

2013-01-24 13:14 203897人阅读 评论(182) 收藏 举报

分类： OpenCV（28） 计算机视觉（72） 图像处理（54）

访问： 7306377次
积分： 27489
等级： BLOG 7
排名： 第172名

原创： 116篇 转载： 11篇 译文： 1篇
评论： 3684条

个人简介

关注：机器学习、计算机视觉、人机交互和人工智能等领域。
邮箱：zouxy09@qq.com
微博：Erik-zou
交流请发邮件，不怎么看博客私信^-^

从最大似然到EM算法浅解

zouxy09@qq.com

http://blog.csdn.net/zouxy09

机器学习十大算法之一：EM算法。能评得上十大之一，让人听起来觉得挺NB的。什么是NB啊，我们一般说某个人很NB，是因为他能解决一些别人解决不了的问题。神为什么是神，因为神能做很多人做不了的事。那么EM算法能解决什么问题呢？或者说EM算法是因为什么而来到这个世界上，还吸引了那么多世人的目光。

我希望自己能通俗地把它理解或者说明白，但是，EM这个问题感觉真的不太好用通俗的语言去说明白，因为它很简单，又很复杂。简单在于它的思想，简单在于其仅包含了两个步骤就能完成强大的功能，复杂在于它的数学推理涉及到比较繁杂的概率公式等。如果只讲简单的，就丢失了EM算法的精髓，如果只讲数学推理，又过于枯燥和生涩，但另一方面，想把两者结合起来也不是件容易的事。所以，我也没法期待我能把它讲得怎样。希望各位不吝指导。

一、最大似然

扯了太多，得入正题了。假设我们遇到的是下面这样的问题：

假设我们需要调查我们学校的男生和女生的身高分布。你怎么做啊？你说那么多人不可能一个一个去问吧，肯定是抽样了。假设你在校园里随便地活捉了100个男生和100个女生。他们共200个人（也就是200个身高的样本数据，为了方便表示，下面，我说“人”的意思就是对应的身高）都在教室里面了。那下一步怎么办啊？你开始喊：“男的左边，女的右边，其他的站中间！”。然后你就先统计抽样得到的100个男生的身高。假设他们的身高是服从高斯分布的。但是这个分布的均值u和方差∂^2我们不知道，这两个参数就是我们要估计的。记作$\theta=[u, \partial]^T$。

用数学的语言来说就是：在学校那么多男生（身高）中，我们独立地按照概率密度$p(x|\theta)$抽取100了个（身高），组成样本集X，我们想通过样本集X来估计出未知参数θ。这里概率密度$p(x|\theta)$我们知道了是高斯分布$N(u,\partial)$的形式，其中的未知参数是$\theta=[u, \partial]^T$。抽到

文章分类

OpenCV (29)
机器学习 (46)
计算机视觉 (73)
Deep Learning (18)
语音识别与TTS (13)
图像处理 (55)
Linux (15)
Linux驱动 (4)
嵌入式 (18)
OpenAL (3)
Android (1)
C/C++编程 (18)
摄像头相关 (5)
数学 (5)
Kinect (9)
神经网络 (8)
随谈 (2)

文章存档

2015年10月 (4)
2015年04月 (2)
2014年12月 (1)
2014年08月 (1)
2014年05月 (2)

展开

阅读排行

Deep Learning（深度 (602253)
Deep Learning（深度 (424549)
Deep Learning（深度 (405599)
Deep Learning论文笔 (267442)
Deep Learning（深度 (259781)
Deep Learning（深度 (226698)
机器学习中的范数规则 (216310)
Deep Learning（深度 (214674)
从最大似然到EM算法 (203783)
Deep Learning（深度 (183791)

评论排行

Deep Learning论文笔 (272)
从最大似然到EM算法 (182)
Deep Learning（深度 (182)
基于Qt的P2P局域网聊 (165)
时空上下文视觉跟踪 (142)
计算机视觉、机器学习 (120)
机器学习算法与Pytho (99)
机器学习中的范数规则 (93)
Deep Learning（深度 (88)

的样本集是$X=\{x_1,x_2,...,x_N\}$，其中x_i表示抽到的第i个人的身高，这里N就是100，表示抽到的样本个数。

由于每个样本都是独立地从$p(x|\theta)$中抽取的，换句话说这100个男生中的任何一个，都是我随便捉的，从我的角度来看这些男生之间是没有关系的。那么，我从学校那么多男生中为什么就恰好抽到了这100个人呢？抽到这100个人的概率是多少呢？因为这些男生（的身高）是服从同一个高斯分布$p(x|\theta)$的。那么我抽到男生A（的身高）的概率是$p(x_A|\theta)$，抽到男生B的概率是$p(x_B|\theta)$，那因为他们是独立的，所以很明显，我同时抽到男生A和男生B的概率是$p(x_A|\theta)* p(x_B|\theta)$，同理，我同时抽到这100个男生的概率就是他们各自概率的乘积了。用数学家的口吻说就是从分布是$p(x|\theta)$的总体样本中抽取到这100个样本的概率，也就是样本集X中各个样本的联合概率，用下式表示：

$$L(\theta)=L(x_1,\cdots,x_n;\theta)=\prod_{i=1}^{n}p(x_i;\theta),\theta\in\Theta.$$

这个概率反映了，在概率密度函数的参数是θ时，得到X这组样本的概率。因为这里X是已知的，也就是说我抽取到的这100个人的身高可以测出来，也就是已知的了，而theta是未知的了，则上面这个公式只有θ是未知数，所以它是θ的函数。这个函数放映的是在不同的参数θ取值下，取得当前这个样本集的可能性，因此称为参数θ相对于样本集X的似然函数（likehood function）。记为L(θ)。

这里出现了一个概念，似然函数。还记得我们的目标吗？我们需要在已经抽到这一组样本X的条件下，估计参数θ的值。怎么估计呢？似然函数有啥用呢？那咱们先来了解下似然的概念。

直接举个例子：

某位同学与一位猎人一起外出打猎，一只野兔从前方窜过。只听一声枪响，野兔应声到下，如果要你推测，这一发命中的子弹是谁打的？你就会想，只发一枪便打中，由于猎人命中的概率一般大于这位同学命中的概率，看来这一枪是猎人射中的。

这个例子所作的推断就体现了极大似然法的基本思想。

再例如：下课了，一群男女同学分别去厕所了。然后，你闲着无聊，想知道课间是男生上厕所的人多还是女生上厕所的人比较多，然后你就跑去蹲在男厕和女厕的门口。蹲了五分钟，突然一个美女走出来，你狂喜，跑过来告诉我，课间女生上厕所的人比较多，你要不相信你可以进去数数。呵呵，我才没那么蠢跑进去数呢，到时还不得上头条。我问你是怎么知道的。你说：“5分钟了，出来的是女生，女生啊，那么女生出来的概率肯定是最大的了，或者说比男生要大，那么女厕所的人肯定比男厕所的人多”。看到了没，你已经运用最大似然估计了。你通过观察到女生先出来，那么什么情况下，女生会先出来呢？肯定是女生出来的概率最大的时候了，那什么时候女生出来的概率最大啊，那肯定是女厕所比男厕所多人的时候了，这个就是你估计到的参数了。

从上面这两个例子，你得到了什么结论？

回到男生身高那个例子。在学校那么男生中，我一抽就抽到这100个男生（表示身高），而不是其他人，那是不是表示在整个学校中，这100个人（的身高）出现的概率最大啊。那么这个概率怎么表示？哦，就是上面那个似然函数L(θ)。所以，我们就只需要找到一个参数θ，其对应的似然函数L(θ)最大，也就是说抽到这100个男生（的身高）概率最大。这个叫做θ的最大似然估计量，记为：

$$\hat{\theta}=\arg\max l(\theta)$$

有时，可以看到L(θ)是连乘的，所以为了便于分析，还可以定义对数似然函数，将其变成连加的：

$$H(\theta)=\ln L(\theta)=\ln\prod_{i=1}^{n}p(x_i;\theta)=\sum_{i=1}^{n}\ln p(x_i;\theta)$$

Deep Learning（深度 (72)

最新评论

Deep Learning论文笔记之
落雨收衫: introduction一般翻译为引言

标签传播算法（Label Propa
000fly: 楼主，你的mpi4py.mpi怎么安装的，我安装了mpi4py，但是不能用mpi4py.MPI

从最大似然到EM算法浅解
Aoman_Hao: 写的真好，通俗易懂

标签传播算法（Label Propa
yoxiyehei_wzx: 请问这篇源码是不是只适用于二分类问题的标签传播？

Deep Learning（深度学习）
爱上帝的小熊猫: 博主写的很适合我这个刚入门的小白，谢谢

人脸识别之特征脸方法（Eig
浅月歌: 博主好 我是初学者 这第一个代码片段中54句显示矩阵尺寸不一致 T—T

基于3D卷积神经网络的人体
已重置: 受益匪浅

Deep Learning论文笔记之
ee_xiao: @makenothing:addpath('E:\M

Deep Learning论文笔记之
ee_xiao: @makenothing:addpath是添加路径到运行目录，..是叫你把具体路径自己加入，如我ad...

图像分割之（五）活动轮廓
qq_19493721: 大神，能不能求下大米粘连图像分割的代码！改善过的，及算法思想，随便什么都行，只要是与大米粘连图像分割...

好了，现在我们知道了，要求θ，只需要使θ的似然函数L(θ)极大化，然后极大值对应的θ就是我们的估计。这里就回到了求最值的问题了。怎么求一个函数的最值？当然是求导，然后让导数为0，那么解这个方程得到的θ就是了（当然，前提是函数L(θ)连续可微）。那如果θ是包含多个参数的向量那怎么处理啊？当然是求L(θ)对所有参数的偏导数，也就是梯度了，那么n个未知的参数，就有n个方程，方程组的解就是似然函数的极值点了，当然就得到这n个参数了。

最大似然估计你可以把它看作是一个反推。多数情况下我们是根据已知条件来推算结果，而最大似然估计是已经知道了结果，然后寻求使该结果出现的可能性最大的条件，以此作为估计值。比如，如果其他条件一定的话，抽烟者发生肺癌的危险时不抽烟者的5倍，那么如果现在我已经知道有个人是肺癌，我想问你这个人抽烟还是不抽烟。你怎么判断？你可能对这个人一无所知，你所知道的只有一件事，那就是抽烟更容易发生肺癌，那么你会猜测这个人不抽烟吗？我相信你更有可能会说，这个人抽烟。为什么？这就是“最大可能”，我只能说他“最有可能”是抽烟的，“他是抽烟的”这一估计值才是“最有可能”得到“肺癌”这样的结果。这就是最大似然估计。

好了，极大似然估计就讲到这，总结一下：

极大似然估计，只是一种概率论在统计学的应用，它是参数估计的方法之一。说的是已知某个随机样本满足某种概率分布，但是其中具体的参数不清楚，参数估计干次试验，观察其结果，利用结果推出参数的大概值。最大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他概率的样本，所以干脆就把这个参数作为估计的真实值。

求最大似然函数估计值的一般步骤：

（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；

二、EM算法

好了，重新回到上面那个身高分布估计的问题。现在，通过抽取得到的那100个男生的身高和已知的其身高服从高斯分布，我们通过最大化其似然函数，就可以得到了对应高斯分布的参数$\theta=[u, \partial]^T$了。那么，对于我们学校的女生的身高分布也可以用同样的方法得到了。

再回到例子本身，如果没有“男的左边，女的右边，其他的站中间！”这个步骤，或者说我抽到这200个人中，某些男生和某些女生一见钟情，已经好上了，纠缠起来了。咱们也不想那么残忍，硬把他们拉扯开。那现在这200个人已经混到一起了，这时候，你从这200个人（的身高）里面随便给我指一个人（的身高），我都无法确定这个人（的身高）是男生（的身高）还是女生（的身高）。也就是说你不知道抽取的那200个人里面的每一个人到底是从男生的那个身高分布里面抽取的，还是女生的那个身高分布抽取的。用数学的语言就是，抽取得到的每个样本都不知道是从哪个分布抽取的。

这个时候，对于每一个样本或者你抽取到的人，就有两个东西需要猜测或者估计的了，一是这个人是男的还是女的？二是男生和女生对应的身高的高斯分布的参数是多少？

只有当我们知道了哪些人属于同一个高斯分布的时候，我们才能够对这个分布的参数作出靠谱的预测，例如刚开始的最大似然所说的，但现在两种高斯分布的人混在一块了，我们又不知道哪些人属于第一个高斯分布，哪些属于第二个，所以就没法估计这两个分布的参数。反过来，只有当我们对这两个分布的参数作出了准确的估计的时候，才能知道到底哪些人属于第一个分布，那些人属于第二个分布。

这就成了一个先有鸡还是先有蛋的问题了。鸡说，没有我，谁把你生出来的啊。蛋不服，说，没有我，你从哪蹦出来啊。（呵呵，这是一个哲学问题。当然了，后来科学家说先有蛋，因为鸡蛋是鸟蛋进化的）。为了解决这个你依赖我，我依赖你的循环依赖问题，总得有一方要先打破僵局，说，不管了，我先随便整一个值出来，看你怎么变，然后我再根据你的变化调整我的变化，然后如此迭代着不断互相推导，最终就会收敛到一个解。这就是EM算法的基本思想了。

不知道大家能否理解其中的思想，我再来啰嗦一下。其实这个思想无处在不啊。

例如，小时候，老妈给一大袋糖果给你，叫你和你姐姐等分，然后你懒得去点糖果的个数，所以你也就不知道每个人到底该分多少个。咱们一般怎么做呢？先把一袋糖果目测的分为两袋，然后把两袋糖果拿在左右手，看哪个重，如果右手重，那很明显右手这代糖果多了，然后你再在右手这袋糖果中抓一把放到左手这袋，然后再感受下哪个重，然后再从重的那袋抓一小把放进轻的那一袋，继续下去，直到你感觉两袋糖果差不多相等了为止。呵呵，然后为了体现公平，你还让你姐姐先选了。

EM算法就是这样，假设我们想估计知道A和B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。

的取值，这个

EM的意思是“Expectation Maximization”，在我们上面这个问题里面，我们是先随便猜一下男生（身高）的正态分布的参数：如均值和方差是多少。例如男生的均值是1米7，方差是0.1米（当然了，刚开始肯定没那么准），然后计算出每个人更可能属于第一个还是第二个正态分布中的（例如，这个人的身高是1米8，那很明显，他最大可能属于男生的那个分布），这个是属于Expectation一步。有了每个人的归属，或者说我们已经大概地按上面的方法将这200个人分为男生和女生两部分，我们就可以根据之前说的最大似然那样，通过这些被大概分为男生的n个人来重新估计第一个分布的参数，女生的那个分布同样方法重新估计。这个是Maximization。然后，当我们更新了这两个分布的时候，每一个属于这两个分布的概率又变了，那么我们就再需要调整E步……如此往复，直到参数基本不再发生变化为止。

这里把每个人（样本）的完整描述看做是三元组$y_i=\{x_i,z_{i1},z_{i2}\}$，其中，$x_i$是第i个样本的观测值，也就是对应的这个人的身高，是可以观测到的值。z_{i1}和z_{i2}表示男生和女生这两个高斯分布中哪个被用来产生值x_i，就是说这两个值标记这个人到底是男生还是女生（的身高分布产生的）。这两个值我们是不知道的，是隐含变量。确切的说，z_{ij}在x_i由第j个高斯分布产生时值为1，否则为0。例如一个样本的观测值为1.8，然后他来自男生的那个高斯分布，那么我们可以将这个样本表示为{1.8, 1, 0}。如果z_{i1}和z_{i2}的值已知，也就是说每个人我已经标记为男生或者女生了，那么我们就可以利用上面说的最大似然算法来估计他们各自高斯分布的参数。但是它们未知，因此我们只能用EM算法。

咱们现在不是因为那个恶心的隐含变量（抽取得到的每个样本都不知道是从哪个分布抽取的）使得本来简单的可以求解的问题变复杂了，求解不了吗。那怎么办呢？人类解决问题的思路都是想能否把复杂的问题简单化。好，那么现在把这个复杂的问题逆回来，我假设已经知道这个隐含变量了，哎，那么求解那个分布的参数是不是很容易了，直接按上面说的最大似然估计就好了。那你就问我了，这个隐含变量是未知的，你怎么就来一个假设说已知呢？你这种假设是没有根据的。呵呵，我知道，所以我们可以先给这个给分布弄一个初始值，然后求这个隐含变量的期望，当成是这个隐含变量的已知值，那么现在就可以用最大似然求解那个分布的参数了吧，那假设这个参数比之前的那个随机的参数要好，它更能表达真实的分布，那么我们再通过这个参数确定的分布去求这个隐含变量的期望，然后再最大化，得到另一个更优的参数，……迭代，就能得到一个皆大欢喜的结果了。

这时候你就不服了，说你老迭代迭代的，你咋知道新的参数的估计就比原来的好啊？为什么这种方法行得通呢？有没有失效的时候呢？什么时候失效呢？用到这个方法需要注意什么问题呢？呵呵，一下子抛出那么多问题，搞得我适应不过来了，不过这证明了你有很好的搞研究的潜质啊。呵呵，其实这些问题就是数学家需要解决的问题。在数学上是可以稳当的证明的或者得出结论的。那咱们用数学来把上面的问题重新描述下。（在这里可

以知道，不管多么复杂或者简单的物理世界的思想，都需要通过数学工具进行建模抽象才得以使用并发挥其强大的作用，而且，这里面蕴含的数学往往能带给你更多想象不到的东西，这就是数学的精妙所在啊）

三、EM算法推导

假设我们有一个样本集$\{x^{(1)},...,x^{(m)}\}$，包含m个独立的样本。但每个样本i对应的类别$z^{(i)}$是未知的（相当于聚类），也即隐含变量。故我们需要估计概率模型p(x,z)的参数θ，但是由于里面包含隐含变量z，所以很难用最大似然求解，但如果z知道了，那我们就很容易求解了。

对于参数估计，我们本质上还是想获得一个使似然函数最大化的那个参数θ，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，见下式（1）。也就是说我们的目标是找到适合的θ和z让L(θ)最大。那我们也许会想，你就是多了一个未知的变量而已啊，我也可以分别对未知的θ和z分别求偏导，再令其等于0，求解出来不也一样吗？

$$\sum_i \log p(x^{(i)};\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)},z^{(i)};\theta) \quad (1)$$

$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} \quad (2)$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} \quad (3)$$

本质上我们是需要最大化（1）式（对（1）式，我们回忆下联合概率密度下某个变量的边缘概率密度函数的求解，注意这里z也是随机变量。对每一个样本i的所有可能类别z求等式右边的联合概率密度函数和，也就得到等式左边为随机变量x的边缘概率密度），也就是似然函数，但是可以看到里面有“和的对数”，求导后形式会非常复杂（自己可以想象下$\log(f_1(x)+f_2(x)+f_3(x)+...)$复合函数的求导），所以很难求解得到未知参数z和θ。那OK，我们可否对（1）式做一些改变呢？我们看（2）式，（2）式只是分子分母同乘以一个相等的函数，还是有“和的对数”啊，还是求解不了，那为什么要这么做呢？咱们先不管，看（3）式，发现（3）式变成了“对数的和”，那这样求导就容易了。我们注意点，还发现等号变成了不等号，为什么能这么变呢？这就是Jensen不等式的大显神威的地方。

Jensen不等式：

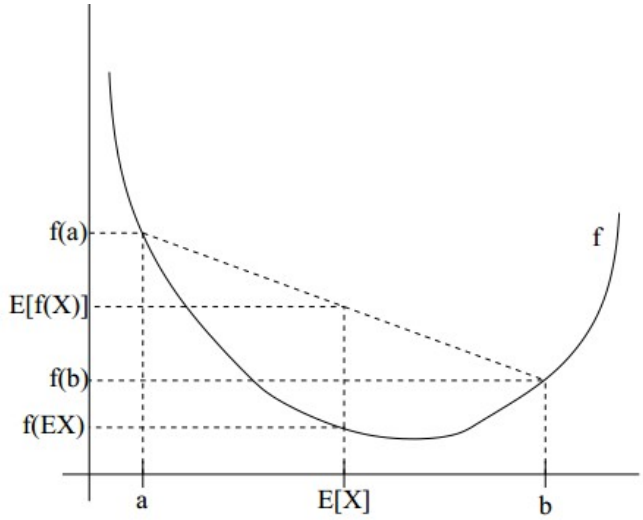
设f是定义域为实数的函数，如果对于所有的实数x。如果对于所有的实数x，f(x)的二次导数大于等于0，那么f是凸函数。当**x**是向量时，如果其hessian矩阵H是半正定的，那么f是凸函数。如果只大于0，不等于0，那么称f是严格凸函数。

Jensen不等式表述如下：

如果f是凸函数，X是随机变量，那么：E[f(X)]>=f(E[X])

特别地，如果f是严格凸函数，当且仅当X是常量时，上式取等号。

如果用图表示会很清晰：



图中，实线f是凸函数，X是随机变量，有0.5的概率是a，有0.5的概率是b。（就像掷硬币一样）。X的期望值就是a和b的中值了，图中可以看到E[f(X)]>=f(E[X])成立。

当f是（严格）凹函数当且仅当-f是（严格）凸函数。

Jensen不等式应用于凹函数时，不等号方向反向。

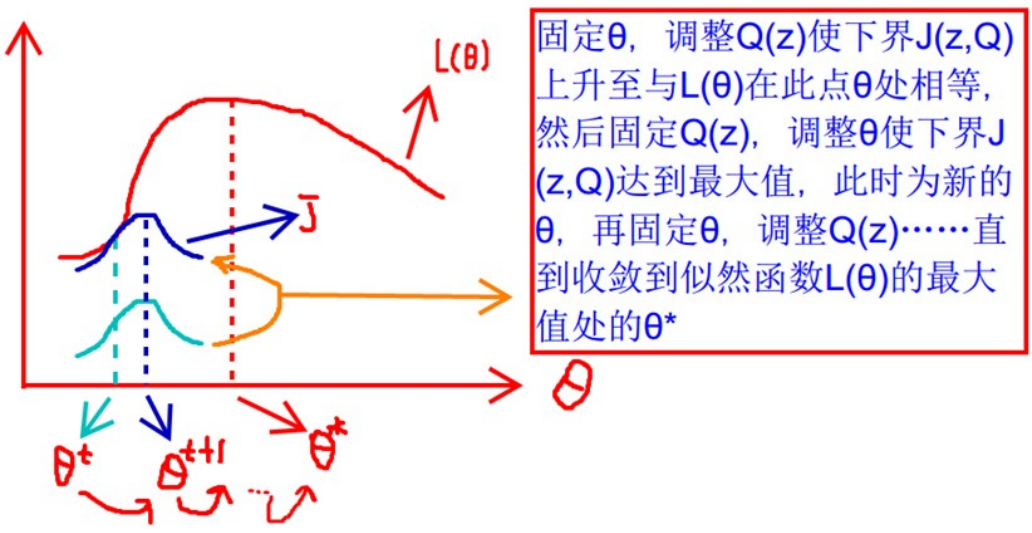
回到公式（2），因为f(x)=log x为凹函数（其二次导数为$-1/x^2<0$）。

（2）式中$\sum_{z^{(i)}} Q_i(z^{(i)}) \left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$是$\left[p(x^{(i)}, z^{(i)};\theta)/Q_i(z^{(i)})\right]$的期望，（考虑到E(X)=∑x*p(x)，f(X)是X的函数，则E(f(X))=∑f(x)*p(x)），又$\sum_z Q_i(z^{(i)}) = 1$，所以就可以得到公式（3）的不等式了（若不明白，请拿起笔，呵呵）：

$$f\left(\mathrm{E}_{z^{(i)}\sim Q_i}\left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]\right) \geq \mathrm{E}_{z^{(i)}\sim Q_i}\left[f\left(\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right)\right]$$

OK，到这里，现在式（3）就容易地求导了，但是式（2）和式（3）是不等号啊，式（2）的最大值不是式（3）的最大值啊，而我们想得到式（2）的最大值，那怎么办呢？

现在我们就需要一点想象力了，上面的式（2）和式（3）不等式可以写成：似然函数L(θ)>=J(z,Q)，那么我们可以通过不断的最大化这个下界J，来使得L(θ)不断提高，最终达到它的最大值。



见上图，我们固定θ，调整Q(z)使下界J(z,Q)上升至与L(θ)在此点θ处相等（绿色曲线到蓝色曲线），然后固定Q(z)，调整θ使下界J(z,Q)达到最大值（θ^t到θ^{t+1}），然后再固定θ，调整Q(z)……直到收敛到似然函数L(θ)的最大值处的θ*。这里有两个问题：什么时候下界J(z,Q)与L(θ)在此点θ处相等？为什么一定会收敛？

首先第一个问题，在Jensen不等式中说到，当自变量X是常数的时候，等式成立。而在这里，即：

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$

再推导下，由于$\sum_z Q_i(z^{(i)}) = 1$（因为Q是随机变量$z^{(i)}$的概率密度函数），则可以得到：分子的和等于c（分子分母都对所有$z^{(i)}$求和：多个等式分子分母相加不变，这个认为每个样例的两个概率比值都是c），则：

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)}|x^{(i)}; \theta) \end{aligned}$$

至此，我们推出了在固定参数θ后，使下界拉升的Q(z)的计算公式就是后验概率，解决了Q(z)如何选择的问题。这一步就是E步，建立L(θ)的下界。接下来的M步，就是在给定Q(z)后，调整θ，去极大化L(θ)的下界J（在固定Q(z)后，下界还可以调整的更大）。那么一般的EM算法的步骤如下：

EM算法（Expectation-maximization）：

期望最大算法是一种从不完全数据或有数据丢失的数据集（存在隐含变量）中求解概率模型参数的最大似然估计方法。

EM的算法流程：

初始化分布参数θ；

重复以下步骤直到收敛：

E步骤：根据参数初始值或上一次迭代的模型参数来计算出隐性变量的后验概率，其实就是隐性变量的期望。作为隐藏变量的现估计值：

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta).$$

M步骤：将似然函数最大化以获得新的参数值：

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

这个不断的迭代，就可以得到使似然函数L(θ)最大化的参数θ了。那就得回答刚才的第二个问题了，它会收敛吗？

感性的说，因为下界不断提高，所以极大似然估计单调增加，那么最终我们会到达最大似然估计的最大值。理性分析的话，就会得到下面的东西：

$$\ell(\theta^{(t+1)}) \geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} \quad (4)$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \quad (5)$$

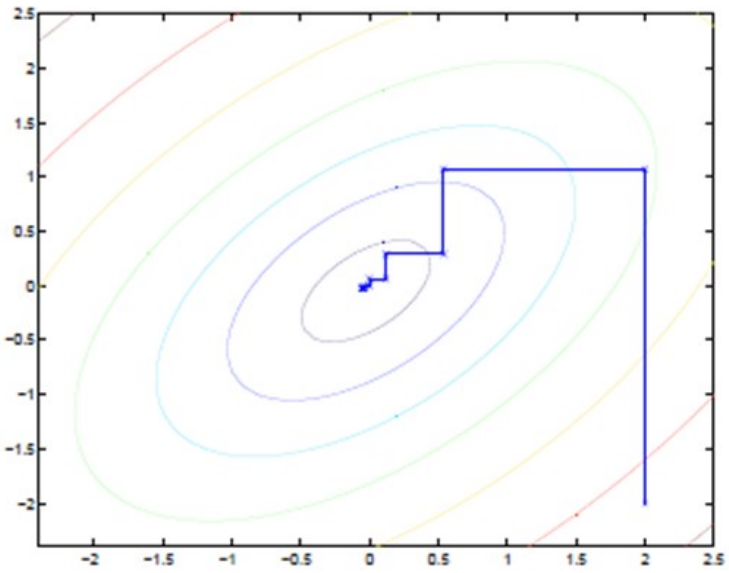
$$= \ell(\theta^{(t)}) \quad (6)$$

具体如何证明的，看推导过程参考：Andrew Ng《The EM algorithm》

http://www.cnblogs.com/jerrylead/archive/2011/04/06/2006936.html

四、EM算法另一种理解

坐标上升法（Coordinate ascent）：



图中的直线式迭代优化的路径，可以看到每一步都会向最优值前进一步，而且前进路线是平行于坐标轴的，因为每一步只优化一个变量。

这犹如在x-y坐标系中找一个曲线的极值，然而曲线函数不能直接求导，因此什么梯度下降方法就不适用了。但固定一个变量后，另外一个可以通过求导得到，因此可以使用坐标上升法，一次固定一个变量，对另外的求极值，最后逐步逼近极值。对应到EM上，**E步**：固定θ，优化Q；**M步**：固定Q，优化θ；交替将极值推向最大。

五、EM的应用

EM算法有很多的应用，最广泛的就是GMM混合高斯模型、聚类、HMM等等。具体可以参考JerryLead的cnblog中的Machine Learning专栏：

（EM算法）The EM Algorithm

混合高斯模型（Mixtures of Gaussians）和EM算法

K-means聚类算法