机器学习算法系列（15）：EM算法

期望最大值（Expectation Maximization，简称EM算法）是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐藏变量。其主要思想就是通过迭代来建立完整数据的对数似然函数的期望界限，然后最大化不完整数据的对数似然函数。本文将尽可能详尽地描述EM算法的原理。并结合高斯混合模型介绍EM算法是如何求解的。

一、定义

EM算法是一种迭代算法，用于含有隐变量（hidden variable）的改了吧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）。故称为期望极大算法（expectation maximization），简称EM算法。

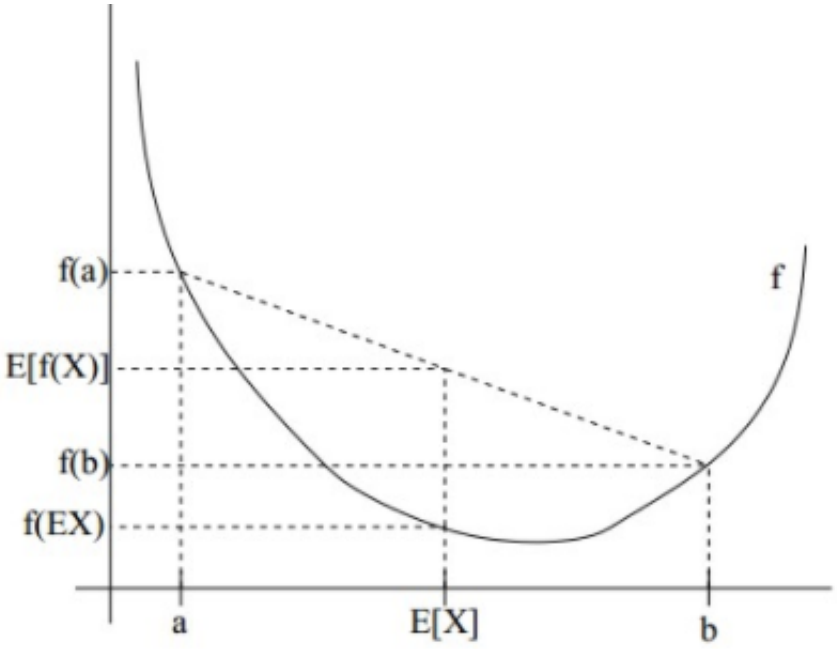
二、Jensen不等式

设f是定义域为实数的函数，如果对于所有的实数x，$f^{''}(x) \geq 0$，那么f是凸函数。当x是向量时，如果其$hessian$矩阵H是半正定的即$H \geq 0$，那么f是凸函数。如果$f^{''}(x) > 0$或$H > 0$，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：$E[f(x)] \geq f(E[x])$。特别地，如果f是严格凸函数，$E[f(x)] \geq f(E[x])$，那么当且仅当$P(x = E[x]) = 1$(也就是说x是常量)，$E[f(x)] = f(E[x])$;
- 如果f是凹函数，x是随机变量，则$E[f(x)] \geq f(E[x])$。当f是（严格）凹函数当且仅当$-f$是（严格）凸函数。

通过下面这张图，我们可以加深理解：



上图中，函数f是凸函数，X是随机变量，有0.5的概率为a，有0.5的概率是b（就像抛硬币一样）。X的期望值就是a和b的中值了，图中可以看到$E[f(x)] \geq f(E[x])$成立。

三、EM思想

3.1 极大似然估计

EM算法推导过程中，会使用到极大似然估计参数。

极大似然估计是一种概率论在统计学的应用。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，参数估计就是通过若干次试验，观察结果，利用结果推出参数的大概值。极大似然估计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小概率的样本，所以干脆就把这个参数作为估计的真实值。

这里再给出求极大似然估计值的一般步骤：

- 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；

关于极大似然估计的实例，可以参考`wikipedia最大似然估计条目`

3.2 EM算法思想

下面介绍EM算法的思想：

给定的训练样本是$x^{(1)}$，$x^{(2)}$，$\cdots$，$x^{(m)}$，样例间相互独立，但每个样本对应的类别$z^{(i)}$是未知的，也即隐含变量。我们想找到每个样例隐含的类别z，能使得$P(x,z)$最大。$P(x,z)$的最大似然估计如下：

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

第一步是对极大似然函数取对数，第二步是对每个样本实例的每个可能的类别z求联合分布概率之和。但是直接求θ一般比较困难，因为有隐藏变量z存在，如果z是一个已知的数，那么使用极大似然估计来估算会很容易。在这种z不确定的情形下，EM算法就派上用场了。

EM算法是一种解决存在隐变量优化问题的有效方法。对于上述情况，由于存在隐变量，不能直接最大化$l(\theta)$，我们可以不断地建立l的下界（E步），然后优化下界（M步），依次迭代，直至算法收敛到局部最优。这就是EM算法的核心思想，简单的归纳一下：

EM算法通过引入隐变量，使用MLE进行迭代求解参数。通常引入隐含变量后会有两个参数，EM算法首先会固定其中的第一个参数，然后使用MLE计算第二个变量值；接着通过固定第二个变量，再使用MLE估计第一个变量值，依次迭代，直至收敛到局部最优解。

四、EM推导

下面来推导EM算法：

对于每一个样例i，让Q_i表示该样例隐含变量z的某种分布，Q_i满足的条件是

$$\sum_{z}Q_i(z)=1\ Q_i(z)\geqslant 0$$

（如果z是连续的，那么Q_i是概率密度函数，需要将求和符号换做积分符号）。比如要将班上学生聚类，假设隐藏变量z是身高，那么就是连续的高斯分布。如果是按照隐藏变量是男女，那么就是伯努利分布。

可以由前面阐述的内容得到下面的公式：

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

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

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

上面三个式子中，式（1）是根据联合概率密度下某个变量的边缘密度求解的（这里把z当做是随机变量）。对每一个样本i的所有可能类别求等式右边的联合概率密度函数和，也就是得到等式左边为随机变量x的边缘概率密度。由于对式（1）直接求导非常困难，我们可以做一个简单的变化，将其分子分母都乘以一个相等的函数$Q_i(Z^{(i)})$，得到式（2）。那么如何从式（2）推导出式（3）呢，这就需要用到之前提到的Jensen不等式。

以下为具体的分析过程：

首先，把（1）式中的log函数看成是一个整体，即令$f(x) = log(x)$，因为$(log(x))'' = -1/x^2 < 0$，根据定理可知其为凹函数。

再根据凹函数的Jensen不等式：$f(E[X]) >= E[f(x)]$。

到这里，我们可以观察到，在式（2）中，当把$log(x)$看成$f(x)$时，后边的

$$\sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}$$

其实就是可以类比为离散型随机变量的期望公式。具体的求解可以参照下图中离散型随机变量的期望公式。

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,…。若$\sum_{k=1}^{\infty} g(x_k)p_k$绝对收敛，则有

$$E(Y) = E[g(X)] = \sum_{k=1}^{\infty} g(x_k)p_k$$

（2）X是连续型随机变量，它的概率密度为$f(x)$，若$\int_{-\infty}^{\infty} g(x)f(x)dx$绝对收敛，则有

$$E(Y) = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

我们可以把$Q_i^{(z^{(i)})}$看成是相应的概率p_i，把

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

看作是$z^{(i)}$的函数$g(z)$，根据期望公式$E[g(x)] = \sum_{i=1}^{\infty} g\left(x_i\right) \cdot p_i$可以得到：

$$E\left(\frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i(z)}\right) = \sum_{z^{(i)}} Q_i(z) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i(z)}$$

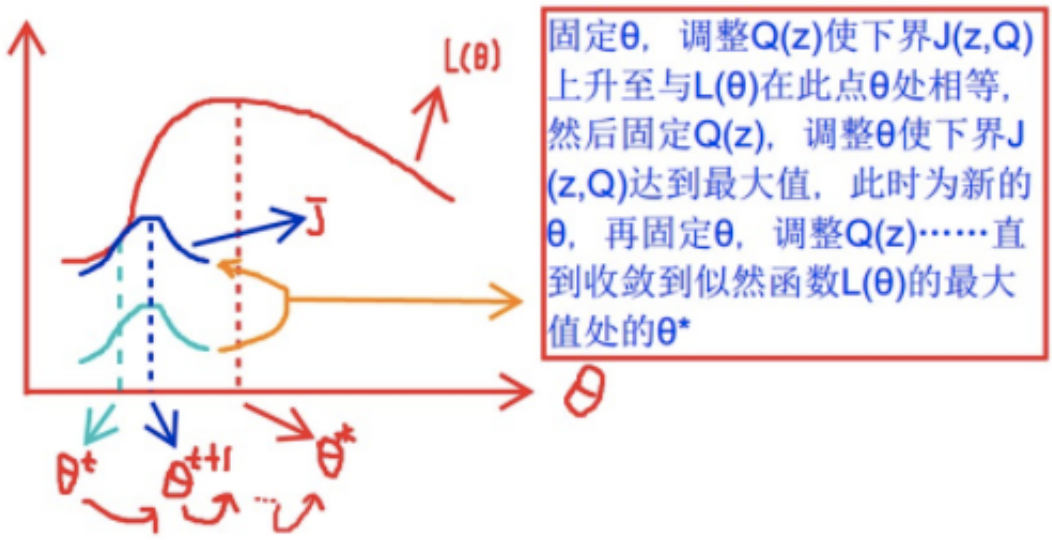
，把上述根据Jensen不等式整合到一起得到：

$$f[E(g(X))] = \log \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}$$

$$\geqslant E[f(g(X))] = \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}$$

这样我们就得到了式（3）。

现在我们把式（2）和式（3）的不等式次而成：似然函数$L(\theta) \geq J(z, Q)$的形式，其中z为隐变量，那么我们可以通过不断地最大化J的下界，来使得$L(\theta)$不断提高，最终达到它的最大值。借助下图来解释下这个过程：



首先我们固定θ，调整$Q(z)$使下界（绿色曲线）$J(z, Q)$沿着绿色虚线上升至与$L(\theta)$在此点θ处相等（绿色曲线至蓝色曲线），然后固定$Q(z)$，调整θ使下界$J(z, Q)$达到最大值(θ_t至θ_{t+1})，然后再固定θ，调整$Q(z)$.....直到收敛到似然函数$L(\theta)$的最大值处的θ^*

这里有两个问题：

- 什么时候下界$J(z, Q)$与$L(\theta)$在此点θ处相等?
- 为什么一定会收敛?

首先来解释下第一个问题。在Jensen不等式中说到，当自变量$X = E(X)$时，即为常数的时候，等式成立。而在这里，为：

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

对该式做个变换，将分母移到等号右边，并对所有的z求和，得到第一个等号；又因为前面提到的$\sum_{z^{(i)}} Q_i\left(z^{(i)}\right) = 1$，得到第二个等号。

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

根据上面两个式子可以得到

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

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

$$= p\left(z^{(i)} | x^{(i)}; \theta\right)$$

到这里，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（条件概率），解决了$Q(z)$如何选择的问题。此步就是EM算法的E步，目的是建立$L(\theta)$的下界。接下来的M步，目的是在给定$Q(z)$后，调整θ，从而极大化$L(\theta)$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：

- 第一步：初始化分布参数θ;
- 第二步：重复E步和M步直到收敛：
  - E步：根据参数的初始值或上一次迭代的模型参数来计算出的因变量的后验概率（条件概率），其实就是隐变量的期望值，来作为隐变量的当前估计值：

$$Q_i\left(z^{(i)}\right)=p\left(z^{(i)}\mid x^{(i)};\theta\right)$$

  - M步：最大化似然函数从而获得新的参数值：

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

通过不断地迭代，然后就可以得到使似然函数$L(\theta)$最大化的参数θ了。

接下来我们看第二个问题。上面多次说到直到收敛，那为什么一定会收敛呢？证明如下：

假定$\theta^{(t)}$和$\theta^{(t+1)}$是EM第t次和t+1次迭代后的结果。如果我们证明了$l(\theta^{(t)})\leq l(\theta^{(t+1)})$，也就是说极大似然估计单调增加，那么最终我们就会得到极大似然估计的最大值。

下面来证明，选定$\theta^{(t)}$后，我们得到E步：

$$Q_i^{(t)}\left(z^{(i)}\right)=p\left(z^{(i)}\mid x^{(i)};\theta\right)$$

这一步保证了在给定$\theta^{(t)}$时，Jensen不等式中的等式成立，也就是

$$l\left(\theta^{(t)}\right)=\sum_i\sum_{z^{(i)}}Q_i\left(z^{(i)}\right)\log\frac{p\left(x^{(i)},z^{(i)};\theta^{(t)}\right)}{Q_i\left(z^{(i)}\right)}$$

然后进行M步，固定$Q_i^{(t)}(z^{(i)})$，并将$\theta^{(t)}$视作变量，对上面的$l(\theta^{(t)})$求导后，得到$\theta^{(t+1)}$,这样经过一些推导会有以下式子成立：

$$l\left(\theta^{(t+1)}\right)\geqslant\sum_i\sum_{z^{(i)}}Q_i\left(z^{(i)}\right)\log\frac{p\left(x^{(i)},z^{(i)};\theta^{(t+1)}\right)}{Q_i\left(z^{(i)}\right)}\cdots\cdots(4)$$

$$\geqslant \sum_i \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t)}\right)}{Q_i\left(z^{(i)}\right)} \cdots\cdots (5)$$

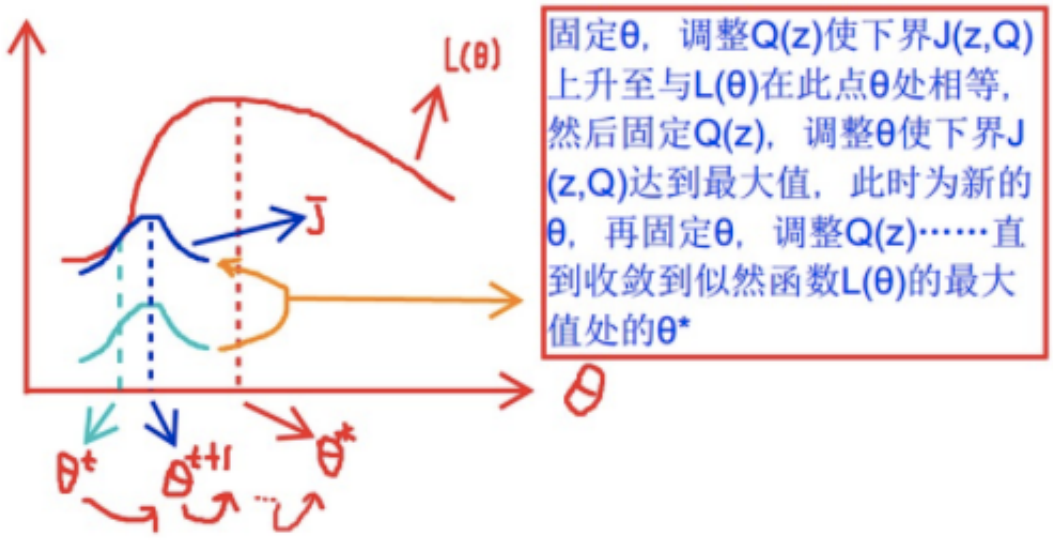
$$= l\left(\theta^{(t)}\right) \cdots\cdots (6)$$

解释第（4）步，得到$\theta^{(t+1)}$时，只是最大化$l(\theta^{(t)})$，也就是$l(\theta^{(t+1)})$的下界，而没有使等式成立，要想使等式成立只有在固定θ，并按E步得到Q_i时才能成立。
况且根据我们前面得到的下式，对于所有的Q_i和θ都成立

$$l\left(\theta^{(t)}\right) \geqslant \sum_i \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta^{(t)}\right)}{Q_i\left(z^{(i)}\right)}$$

第（5）式利用的M步的定义，M步就是将$\theta^{(t)}$调整到$\theta^{(t+1)}$，使得下界最大化。这样（5）、（6）就都证明成立了。

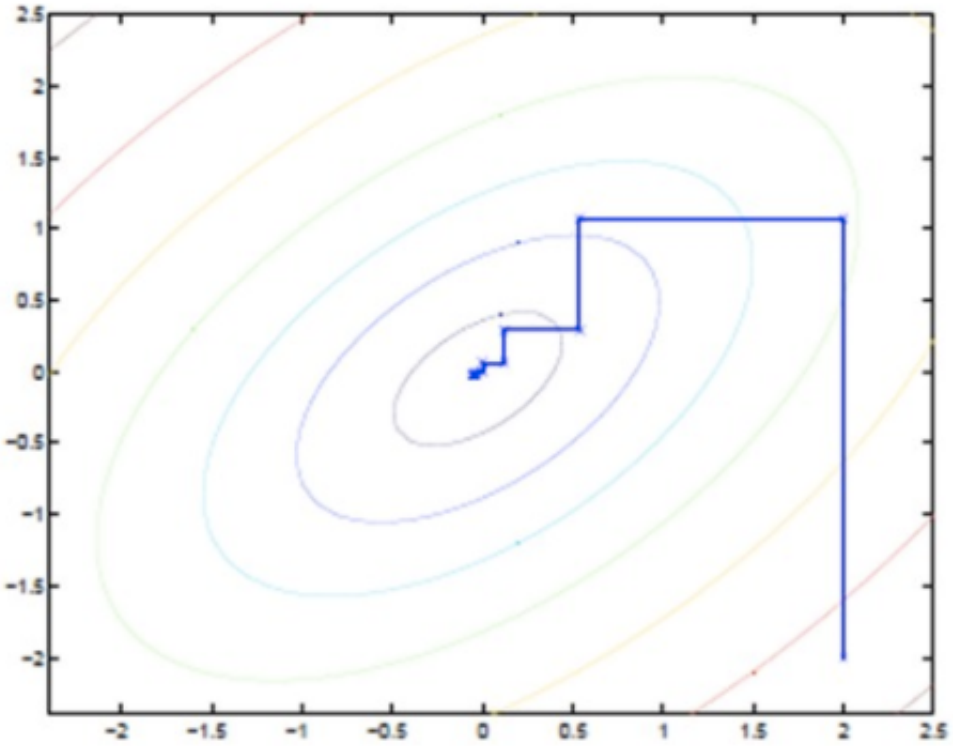
再结合之前那个图解释一下这几步推导：



首先（4）对所有的参数都满足，而其等式成立条件只是在固定θ，并调整好Q时成立，而第（4）步只是固定Q，调整θ，不能保证等式一定成立。对应到图上就是蓝色曲线的峰值与$l(\theta^{(t+1)})$的关系，要使它们相等还必须要固定θ，调整好Q；（4）到（5）就是M步的定义，也就是固定Q，调整$\theta^{(t)}$至$\theta^{(t+1)}$，对应到图上即为蓝色曲线与红色曲线交点处至蓝色曲线峰值。
（5）到（6）是前面E步所保证等式成立条件。也就是说E步会将下界拉到与$l(\theta)$一个特定值（这里为$\theta^{(t)}$）一样的高度，而此时发现下界仍然可以上升，因此经过M步后，下界又被拉升，但达不到与$l(\theta)$另外一个特定值（$\theta^{(t+1)}$）一样的高度，之后E步又将下界拉到了与这个特定值一样的高度，循环往复，直到达到最大值。

这样就证明了$l(\theta)$会单调增加。如果要判断收敛情况，可以这样来做：一种收敛方法是$l(\theta)$不再变化，还有一种就是变化幅度很小，即根据$l(\theta^{(t+1)}) = l(\theta^{(t)})$的值来决定。

从前面的推导中我们知道$l(\theta) \geq J(Q, \theta)$，EM也可以看做是$J$的坐标上升法，如下图所示：



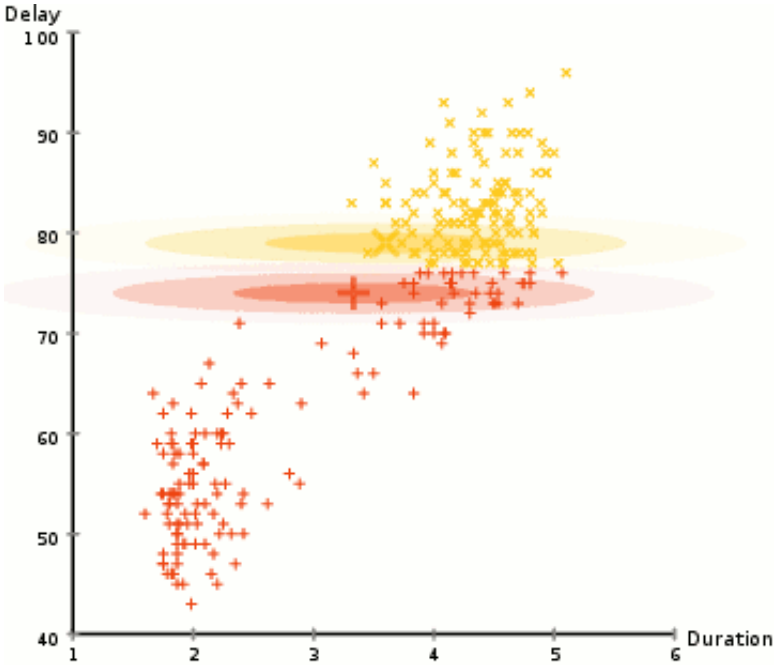
图中的直线式迭代优化的路径，可以看到每一步都会向最优值前进一步，而且前进路线是平行于坐标轴的，因为每一步只优化一个变量。
这犹如在x-y坐标系中找一个曲线的极值，然而曲线函数不能直接求导，因此什么梯度下降方法就不适用了。但固定一个变量后，另外一个可以通过求导得到，因此可以使用坐标上升法，一次固定一个变量，对另外的求极值，最后逐步逼近极值。对应到EM上，E步：固定θ，优化Q；M步：固定Q，优化θ；交替将极值推向最大。

五、EM的应用：混合高斯模型

待补充

六、EM的应用：EM聚类

以下的聚类图来自维基百科，可以生动的看出



待补充

七、参考资料

`The EM Algorithm`

`混合高斯模型和EM算法`

`cs229-notes8`