1. 从最大似然到EM
    1.1 最大似然
        1.1.1 问题描述
        1.1.2 估算参数
        1.1.3 最大似然估计总结
        1.1.4 求最大似然函数估计值的一般步骤：
        1.1.5 最大似然函数的应用
    1.2 EM算法
        1.2.1 问题描述
        1.2.2 EM 算法
        1.2.3 总结
2 EM算法推导
    2.1 基础知识
        2.1.1 凸函数
        2.1.2 Jensen不等式
        2.1.3 期望
    2.2 EM算法的推导
    2.3 EM算法流程
    2.4 EM算法另一种理解
    2.5 EM算法的收敛性思考
    2.6. EM算法应用
3. EM算法案例-两硬币模型

最大似然和EM算法，与其说是一种算法，不如说是一种解决问题的思想，解决一类问题的框架，和线性回归，逻辑回归，决策树等一些具体的算法不同，最大似然和EM算法是更加抽象，是很多具体算法的基础。

1. 从最大似然到EM

1.1 最大似然

1.1.1 问题描述

假设我们需要调查我们学校学生的身高分布。我们先假设学校所有学生的身高服从正态分布 $N(\mu, \sigma^2)$ 。这个分布的均值 μ 和方差 σ^2 未知，如果我们估计出这两个参数，那我们就得到了结果。那么怎样估计这两个参数呢？

我们可以先对学生进行抽样。假设我们随机抽到了200个人（也就是200个身高的样本数据，为了方便表示，下面，“人”的意思就是对应的身高）。然后统计抽样这200个人的身高。根据这200个人的身高估计均值 μ 和方差 σ^2 。

用数学的语言来说就是：为了统计学校学生的身高分布，我们独立地按照概率密度 $p(x|\theta)$ 抽取了 200 个（身高），组成样本集 $X = \{x_1, x_2, \ldots, x_N\}$ (其中 x_i 表示抽到的第 i 个人的身高，这里 N 就是200，表示样本个数)，我们想通过样本集 X 来估计出未知参数 θ。这里概率密度 $p(x|\theta)$ 服从高斯分布 $N(\mu, \sigma^2)$ ，其中的未知参数是 $\theta = [\mu, \sigma]^T$。

那么问题来了怎样估算参数 θ 呢？

1.1.2 估算参数

我们先回答几个小问题：

问题一：抽到这 200 个人的概率是多少呢？

由于每个样本都是独立地从 $p(x|\theta)$ 中抽取的，换句话说这 200 个学生随便捉的，他们之间是没有关系的，即他们之间时相互独立的。假如抽到学生A（的身高）的概率是 $p(x_A|\theta)$，抽到学生B的概率是 $p(x_B|\theta)$，那么同时抽到男生A和男生B的概率是 $p(x_A|\theta) \times p(x_B|\theta)$，同理，我同时抽到这200个学生的概率就是他们各自概率的乘积了，即为他们的联合概率用下式表示：

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} p(x_i|\theta), \quad \theta \in \Theta$$

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 θ 是未知了，则上面这个公式只有 θ 是未知数，所以它是 θ 的函数。

这个函数反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 $L(\theta)$。

为了便于分析，还可以定义对数似然函数，将其变成连加的：

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

问题二：学校那么多学生，为什么就恰好抽到了这 200 个人（身高）呢？

在学校那么学生中，我一抽就抽到这200个学生（身高），而不是其他人，那是不是表示在整个学校中，这 200 个人（的身高）出现的概率最大啊，也就是其对应的似然函数L(θ)最大，即

$$\hat{\theta} = \operatorname{argmax} L(\theta)$$

$\hat{\theta}$这个叫做 θ 的最大似然估计量，即为我们所求的值。

问题三：那么怎么极大似然函数？

求 $L(\theta)$ 对所有参数的偏导数，然后让这些偏导数为 0，假设有 n 个参数，就有 n 个方程组成的方程组，那么方程组的解就是似然函数的极值点了，从而得到对应的 θ 了。

1.1.3 最大似然估计总结

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。

比如说，

- 假如一个学校的学生男女比例为9:1，那么你可以推出，你在这个学校里更大可能性遇到的是男生；
- 假如你不知道那女比例，你走在路上，碰到100个人，发现男生就有90个，这时候你可以推断这个学校的男女比例更有可能为 9:1，这就是最大似然估计。

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，参数估计就是通过若干次试验，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小概率的样本，所以干脆就把这个参数作为估计的真实值。

1.1.4 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；

1.1.5 最大似然函数的应用

应用一 ：回归问题中的极小化平方和

假设线性回归模型具有如下形式: $h(x)=\sum_{i=1}^{d}\theta_j x_j+\epsilon=\theta^T x+\epsilon$ ，其中 $x\in R^{1\times d},\theta\in R^{1\times d}$，误差 $\epsilon\in R$， 当前已知 $X=(x_1,\cdots,x_m)^T\in R^{m\times d},y\in R^{m\times 1}$， 如何求 θ 呢？

- 最小二乘估计：最合理的参数估计量应该使得模型能最好地拟合样本数据，也就是估计值和观测值之差的平方和最小，其推导过程如下所示：

  $J(\theta)=\sum_{i=1}^{n}(h_\theta(x_i)-y_i)$

  求解方法是通过梯度下降算法，通过训练数据不断迭代得到最终的值。

- 最大似然法：最合理的参数估计量应该使得从模型中抽取该 n 组样本观测值的概率最大，也就是似然函数最大。

  假设误差项 $\epsilon\in N(0,\sigma^2)$ ，则 $y_i\in N(\theta x_i,\sigma^2)$

$$p(y_i|x_i;\theta)=\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2})$$

$$\begin{aligned}L(\theta)&=\prod_{i=1}^{m}p(y_i|x_i;\theta)\\&=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2})\end{aligned}$$

$$\begin{aligned}H(\theta)&=log(L(\theta))\\&=\log\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2})\\&=\prod_{i=1}^{m}\log\frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2})\\&=\sum_{i=1}^{m}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i-\theta^T x_i)^2-n\ln\sigma\sqrt{2\pi}\end{aligned}$$

  令 $J(\theta)=-\frac{1}{2}\sum_{i=1}^{m}(y_i-\theta^T x_i)^2$ 则 $arg\max_\theta H(\theta)\Leftrightarrow arg\min_\theta J(\theta)$， 即将极大似然函数等价于极小化平方和。

  这时可以发现，此时的最大化似然函数和最初的最小二乘损失函数的估计结果是等价的。但是要注意这两者只是恰好有着相同的表达结果，原理和出发点完全不同。

应用二：分类问题中极小化交叉熵。

在分类问题中，交叉熵的本质就是似然函数的最大化，逻辑回归的假设函数为：

$$h(x)=\hat{y}=\frac{1}{1+e^{-\theta^T x+b}}$$

根据之前学过的内容我们知道 $\hat{y}=p(y=1|x,\theta)$ ，

当 y=1 时， $p_1=p(y=1|x,\theta)=\hat{y}$

当 y =0 时， $p_0=p(y=0|x,\theta)=1-\hat{y}$

合并上面两式子，可以得到

$p(y|x,\theta)=\hat{y}^y(1-\hat{y})^{1-y}$

$$\begin{aligned}L(\theta)&=\prod_{i=1}^{m}p(y_i|x_i;\theta)\\&=\prod_{i=1}^{m}\hat{y}_i^{y_i}(1-\hat{y}_i)^{1-y_i}\end{aligned}$$

$$\begin{aligned}H(\theta)&=\log(L(\theta))\\&=\log\prod_{i=1}^{m}\hat{y}_i^{y_i}(1-\hat{y}_i)^{1-y_i}\\&=\prod_{i=1}^{m}\log\hat{y}_i^{y_i}(1-\hat{y}_i)^{1-y_i}\\&=\sum_{i=1}^{m}y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\end{aligned}$$

令 $J(\theta)=-H(\theta)=-\sum_{i=1}^{m}y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)$ 则 $arg\max_{\theta}H(\theta)\Leftrightarrow arg\min_{\theta}J(\theta)$， 即将极大似然函数等价于极小化平方和。

1.2 EM算法

1.2.1 问题描述

上面我们先假设学校所有学生的身高服从正态分布 $N(\mu,\sigma^2)$ 。实际情况并不是这样的，男生和女生分别服从两种不同的正态分布，即男生∈ $N(\mu_1,\sigma_1^2)$，女生∈ $N(\mu_2,\sigma_2^2)$，那么该怎样评估学生的身高分布呢？

简单啊，我们可以随便抽100个男生和100个女生，将男生和女生分开，对他们单独进行最大似然估计。分别求出男生和女生的分布。

假如某些男生和某些女生好上了，纠缠起来了。咱们也不想那么残忍，硬把他们拉扯开。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）。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。那怎么办呢？

1.2.2 EM 算法

这个时候，对于每一个样本或者你抽取到的人，就有两个东西需要估计了，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正态分布的参数是多少。

当我们知道了每个人是男生还是女生，我们可以很容易利用最大似然对男女各自的身高的分布进行估计。

反过来，当我们知道了男女身高的分布参数我们才能知道每一个人更有可能是男生还是女生。例如我们已知男生的身高分布为 $N(\mu_1 = 172, \sigma_1^2 = 5^2)$ ，女生的身高分布为 $N(\mu_2 = 162, \sigma_2^2 = 5^2)$ ，一个学生的身高为180，我们可以推断出这个学生为男生的可能性更大。

但是现在我们既不知道每个学生是男生还是女生，也不知道男生和女生的身高分布。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。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。

EM的意思是“Expectation Maximization”，具体方法为：

- 先设定男生和女生的身高分布参数(初始值)，例如男生的身高分布为 $N(\mu_1 = 172, \sigma_1^2 = 5^2)$ ，女生的身高分布为 $N(\mu_2 = 162, \sigma_2^2 = 5^2)$ ，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180，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。
- 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估计分别对男生和女生的身高分布参数进行估计。这个是 Maximization；
- 然后，当我们更新这两个分布的时候，每一个学生属于女生还是男生的概率又变了，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。

1.2.3 总结

上面的学生属于男生还是女生我们称之为隐含参数，女生和男生的身高分布参数称为模型参数。

EM算法解决这个的思路是使用启发式的迭代方法，既然我们无法直接求出模型分布参数，那么我们可以先猜想隐含参数（EM算法的E步），接着基于观察数据和猜测的隐含参数一起来极大化对数似然，求解我们的模型参数（EM算法的M步)。由于我们之前的隐含参数是猜测的，所以此时得到的模型参数一般还不是我们想要的结果。我们基于当前得到的模型参数，继续猜测隐含参数（EM算法的E步），然后继续极大化对数似然，求解我们的模型参数（EM算法的M步)。以此类推，不断的迭代下去，直到模型分布参数基本无变化，算法收敛，找到合适的模型参数。

一个最直观了解EM算法思路的是K-Means算法。在K-Means聚类时，每个聚类簇的质心是隐含数据。我们会假设K个初始化质心，即EM算法的E步；然后计算得到每个样本最近的质心，并把样本聚类到最近的这个质心，即EM算法的M步。重复这个E步和M步，直到质心不再变化为止，这样就完成了K-Means聚类。

2 EM算法推导

2.1 基础知识

2.1.1 凸函数

设是定义在实数域上的函数，如果对于任意的实数，都有：

$$f'' \geq 0$$

那么是凸函数。若不是单个实数，而是由实数组成的向量，此时，如果函数的 Hesse 矩阵是半正定的，即

$$H'' \geq 0$$

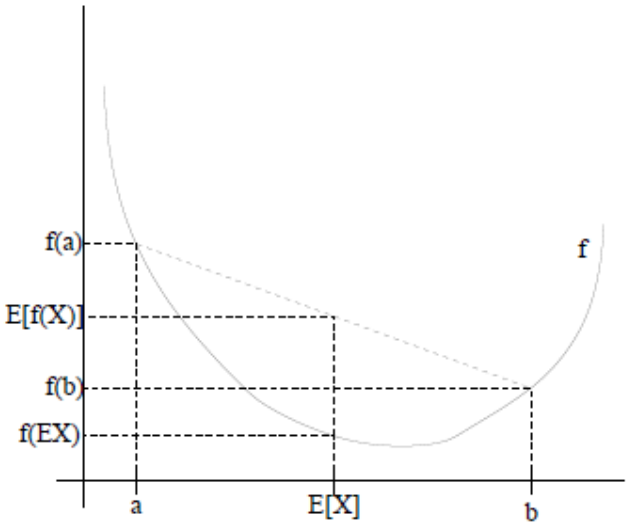
那么是凸函数。特别地，如果 $f'' > 0$ 或者 $H'' > 0$，那么称为严格凸函数。

2.1.2 Jensen不等式

如下图，如果函数 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，X的期望值就是a和b的中值了那么：

$$E[f(x)] \geq f(E(x))$$

特别地，如果函数 f 是严格凸函数，当且仅当：$p(x = E(x)) = 1$ (即随机变量是常量) 时等号成立。



注：若函数 f 是凹函数，Jensen不等式符号相反。

2.1.3 期望

对于离散型随机变量 X 的概率分布为 $p_i = p\{X = x_i\}$，数学期望 E(X) 为：

$$E(X) = \sum_i x_i p_i$$

若连续型随机变量X的概率密度函数为 $f(x)$，则数学期望 E(X) 为：

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

设 $Y = g(X)$， 若 X 是离散型随机变量，则：

$$E(Y) = \sum_i g(x_i) p_i$$

若 $f(x)$ 是连续型随机变量，则：

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

2.2 EM算法的推导

对于 m 个相互独立的样本 $x = (x^{(1)}, x^{(2)}, \dots x^{(m)})$ ，对应的隐含数据 $z = (z^{(1)}, z^{(2)}, \dots z^{(m)})$，此时 (x, z) 即为完全数据，样本的模型参数为 θ, 则观察数据 x 的似然函数为 $P(x|\theta)$，完全数据的似然函数为 $P(x, z|\theta)$。

假如没有隐含变量 z，我们仅需要找到合适的 θ 极大化对数似然函数即可：

$$\theta = arg\max_{\theta} L(\theta) = arg\max_{\theta} \sum_{i=1}^{m} logP(x^{(i)}|\theta)$$

增加隐含变量 z 之后，我们的目标变成了找到合适的 θ 和 z 让对数似然函数最大：

$$\theta, z = arg\max_{\theta,z} L(\theta, z) = arg\max_{\theta,z} \sum_{i=1}^{m} log \sum_{z^{(i)}} P(x^{(i)}，z^{(i)}|\theta)$$

如果对分别对未知的 θ 和 z 分别求偏导，求导后形式会非常复杂（可以想象下 $log(f_1(x) + f_2(x) + f_3(x) + \ldots)$ 复合函数的求导，所以很难求解得到 z 和 θ。那么我们想一下可不可以将 加号从log中提取出来呢?我们可以对这个式子进行缩放如下：

$$\sum_{i=1}^{m} log \sum_{z^{(i)}} P(x^{(i)}，z^{(i)}|\theta) = \sum_{i=1}^{m} log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}，z^{(i)}|\theta)}{Q_i(z^{(i)})} \quad (1)$$

$$\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{P(x^{(i)}，z^{(i)}|\theta)}{Q_i(z^{(i)})} \quad (2)$$

上面第(1)式引入了一个未知的新的分布 $Q_i(z^{(i)})$，满足：

$\sum_z Q_i(z) = 1, Q_i(z) \geq 1$。

第(2)式用到了 Jensen 不等式 (对数函数是凹函数)：

$$log(E(y)) \geq E(log(y))$$

其中：

$E(y) = \sum_i \lambda_i y_i, \lambda_i \geq 0, \sum_i \lambda_i = 1$

$y_i = \frac{P(x^{(i)}，z^{(i)}|\theta)}{Q_i(z^{(i)})}$

$\lambda_i = Q_i(z^{(i)})$

上式实际上是我们构建了 $L(\theta, z)$ 的下界(E步)，下一步要做的就是寻找一个合适的 $Q_i(z)$ 最优化这个下界(M步)。

假设 θ 已经给定，那么 $logL(\theta)$ 的值就取决于 $Q_i(z^{(i)})$ 和 $p(x^{(i)}, z^{(i)})]$ 了。我们可以通过调整这两个概率使下界逼近$logL(\theta)$ 的真实值，当不等式变成等式时，说明我们调整后的下界能够等价于$logL(\theta)$ 了。由 Jensen 不等式可知，等式成立的条件是随机变量是常数，则有：

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

其中 c 为常数，我们得到：

$$P(x^{(i)}，z^{(i)}|\theta) = cQ_i(z^{(i)})$$

$$\sum_z P(x^{(i)}，z^{(i)}|\theta) = c\sum_z Q_i(z^{(i)})$$

由于 $\sum_z Q_i(z^{(i)}) = 1$。 从上面两式，我们可以得到：

$$\sum_z P(x^{(i)}, z^{(i)}|\theta) = c$$

$$Q_i(z^{(i)}) = \frac{P(x^{(i)}, z^{(i)}|\theta)}{c} = \frac{P(x^{(i)}, z^{(i)}|\theta)}{\sum_z P(x^{(i)}, z^{(i)}|\theta)} = \frac{P(x^{(i)}, z^{(i)}|\theta)}{P(x^{(i)}|\theta)} = P(z^{(i)}|x^{(i)}, \theta))$$

如果，$Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}, \theta))$, 则第 (2) 式是我们的包含隐藏数据的对数似然的一个下界。如果我们能极大化这个下界，则也在尝试极大化我们的对数似然。即我们需要最大化下式：

$$arg\max_\theta \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{P(x^{(i)}, z^{(i)}|\theta)}{Q_i(z^{(i)})}$$

去掉上式中为常数的部分，则我们需要极大化的对数似然下界为：

$$arg\max_\theta \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) log P(x^{(i)}, z^{(i)}|\theta)$$

至此，我们推出了在固定参数 θ 后分布 $Q_i(z^{(i)})$ 的选择问题， 从而建立了 $logL(\theta)$ 的下界，这是E步，接下来的M步骤就是固定 $Q_i(z^{(i)})$ 后，调整 θ，去极大化 $logL(\theta)$ 的下界。

2.3 EM算法流程

现在我们总结下EM算法的流程。

输入：观察数据 $x = (x^{(1)}, x^{(2)}, \dots x^{(m)})$ ，联合分布 $p(x, z|\theta)$，条件分布 $p(z|x, \theta)$， 最大迭代次数 J。

1) 随机初始化模型参数 θ 的初值 θ^0

2) $\mathbf{for\ j\ from\ 1\ to\ J}$：

a) E步：计算联合分布的条件概率期望：

$Q_i(z^{(i)}) := P(z^{(i)}|x^{(i)}, \theta))$

b) M步：极大化 $L(\theta)$,得到 θ:

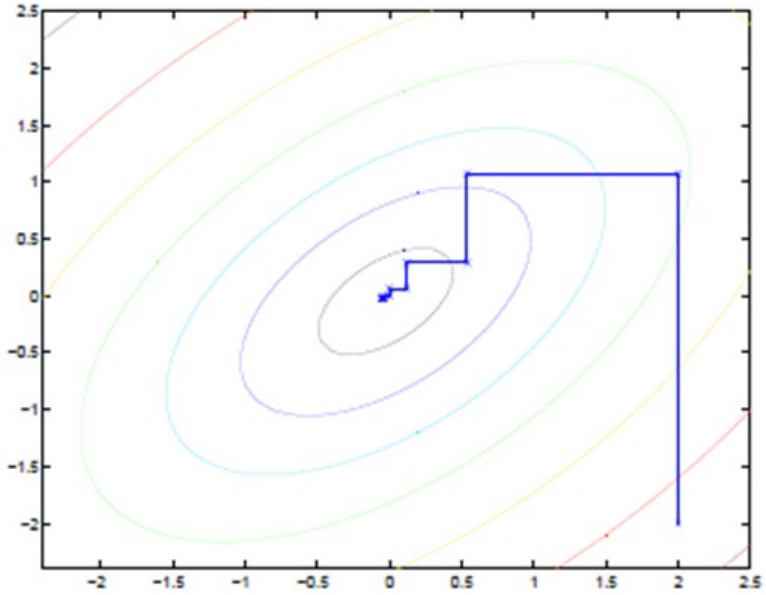
$\theta := arg\max_\theta \sum_{i=1}^m \sum_{z^{(i)}} Q_i(z^{(i)}) log P(x^{(i)}, z^{(i)}|\theta)$

c) 重复E、M步骤直到 θ 收敛

输出：模型参数 θ

2.4 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。

2.5 EM算法的收敛性思考

EM算法的流程并不复杂，但是还有两个问题需要我们思考：

1）EM算法能保证收敛吗？

2）EM算法如果收敛，那么能保证收敛到全局最大值吗？

首先我们来看第一个问题, EM算法的收敛性。要证明EM算法收敛，则我们需要证明我们的对数似然函数的值在迭代的过程中一直在增大。即：

$$\sum_{i=1}^{m} logP(x^{(i)}|\theta^{j+1}) \geq \sum_{i=1}^{m} logP(x^{(i)}|\theta^{j})$$

由于：

$$L(\theta,\theta^{j}) = \sum_{i=1}^{m}\sum_{z^{(i)}} P(z^{(i)}|x^{(i)},\theta^{j}))logP(x^{(i)},z^{(i)}|\theta)$$

令：

$$H(\theta,\theta^{j}) = \sum_{i=1}^{m}\sum_{z^{(i)}} P(z^{(i)}|x^{(i)},\theta^{j}))logP(z^{(i)}|x^{(i)},\theta)$$

上两式相减得到：

$$\sum_{i=1}^{m} logP(x^{(i)}|\theta) = L(\theta,\theta^{j}) - H(\theta,\theta^{j})$$

在上式中分别取 θ 为θ^{j}和 θ^{j+1}，并相减得到：

$$\sum_{i=1}^{m} logP(x^{(i)}|\theta^{j+1}) - \sum_{i=1}^{m} logP(x^{(i)}|\theta^{j}) = [L(\theta^{j+1},\theta^{j}) - L(\theta^{j},\theta^{j})] - [H(\theta^{j+1},\theta^{j}) - H(\theta^{j},\theta^{j})]$$

要证明EM算法的收敛性，我们只需要证明上式的右边是非负的即可。

由于 θ^{j+1} 使得 $L(\theta,\theta^{j})$ 极大，因此有：

$$L(\theta^{j+1},\theta^{j}) - L(\theta^{j},\theta^{j}) \geq 0$$

而对于第二部分，我们有：

$$\begin{aligned} H(\theta^{j+1},\theta^{j}) - H(\theta^{j},\theta^{j}) &= \sum_{i=1}^{m}\sum_{z^{(i)}} P(z^{(i)}|x^{(i)},\theta^{j}) log\frac{P(z^{(i)}|x^{(i)},\theta^{j+1})}{P(z^{(i)}|x^{(i)},\theta^{j})} \quad (3) \\ &\leq \sum_{i=1}^{m} log(\sum_{z^{(i)}} P(z^{(i)}|x^{(i)},\theta^{j})\frac{P(z^{(i)}|x^{(i)},\theta^{j+1})}{P(z^{(i)}|x^{(i)},\theta^{j})}) \quad (4) \\ &= \sum_{i=1}^{m} log(\sum_{z^{(i)}} P(z^{(i)}|x^{(i)},\theta^{j+1})) = 0 \quad (5) \end{aligned}$$

其中第（4）式用到了Jensen不等式，只不过和第二节的使用相反而已，第（5）式用到了概率分布累积为1的性质。

至此，我们得到了：$\sum_{i=1}^{m} logP(x^{(i)}|\theta^{j+1}) - \sum_{i=1}^{m} logP(x^{(i)}|\theta^{j}) \geq 0$ ，证明了EM算法的收敛性。

从上面的推导可以看出，EM 算法可以保证收敛到一个稳定点，但是却不能保证收敛到全局的极大值点，因此它是局部最优的算法，当然，如果我们的优化目标 $L(\theta,\theta^{j})$ 是凸的，则EM算法可以保证收敛到全局最大值，这点和梯度下降法这样的迭代算法相同。至此我们也回答了上面提到的第二个问题。

2.6. EM算法应用

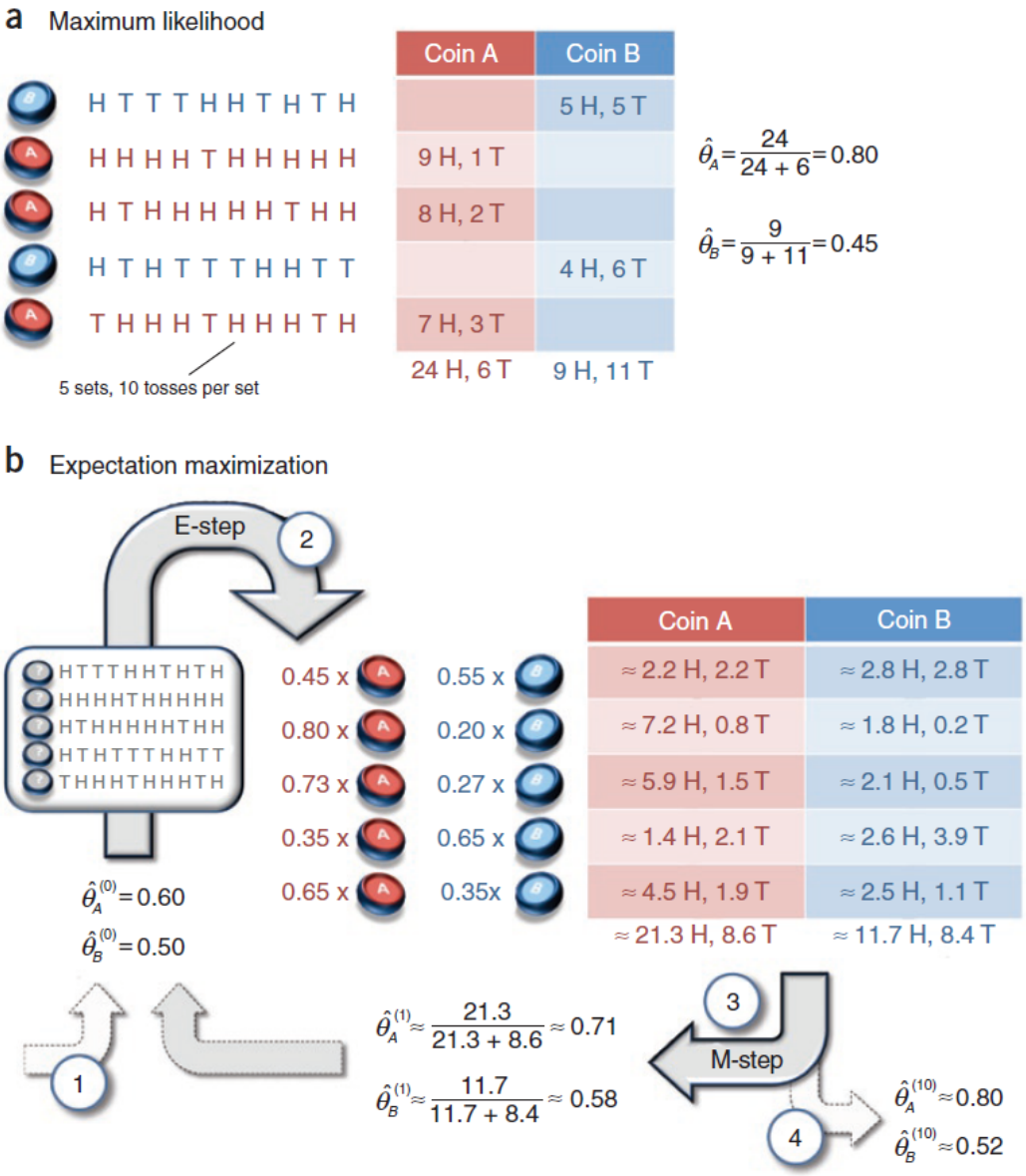
如果我们从算法思想的角度来思考EM算法，我们可以发现我们的算法里已知的是观察数据，未知的是隐含数据和模型参数，在E步，我们所做的事情是固定模型参数的值，优化隐含数据的分布，而在M步，我们所做的事情是固定隐含数据分布，优化模型参数的值。EM的应用包括：

- 支持向量机的SMO算法
- Lasso回归算法坐标轴下降法
- 混合高斯模型
- K-means
- 隐马尔可夫模型

3. EM算法案例-两硬币模型

总结一下：

假设有两枚硬币A、B，以相同的概率随机选择一个硬币，进行如下的掷硬币实验：共做 5 次实验，每次实验独立的掷十次，结果如图中a所示，例如某次实验产生了H、T、T、T、H、H、T、H、T、H(H代表正面朝上)。a是在知道每次选择的是A还是B的情况下进行，b是在不知道选择的硬币情况下进行，问如何估计两个硬币正面出现的概率？



CASE a

已知每个实验选择的是硬币A 还是硬币 B，重点是如何计算输出的概率分布，这其实也是极大似然求导求出来的。

$$\underset{\theta}{argmax} logP(Y|\theta) = log((\theta_B^5(1-\theta_B)^5)(\theta_A^9(1-\theta_A))(\theta_A^8(1-\theta_A)^2)(\theta_B^4(1-\theta_B)^6)(\theta_A^7(1-\theta_A)^3))$$
$$= log((\theta_A^{24}(1-\theta_A)^6)(\theta_B^9(1-\theta_B)^{11}))$$

上面这个式子求导之后发现，5次实验中A正面向上的次数再除以总次数作为即为 $\hat{\theta}_A$，5次实验中B正面向上的次数再除以总次数作为即为 ，即:

$\hat{\theta_A} = \frac{24}{24+6} = 0.80$

$\hat{\theta_B} = \frac{9}{9+11} = 0.45$

CASE b

由于并不知道选择的是硬币 A 还是硬币 B，因此采用EM算法。

E步：初始化 $\hat{\theta}_A^{(0)} = 0.60$ 和 $\hat{\theta}_B^{(0)} = 0.50$ ，计算每个实验中选择的硬币是 A 和 B 的概率，例如第一个实验中选择 A 的概率为：

$P(z = A|y_1, \theta) = \frac{P(z=A, y_1|\theta)}{P(z=A, y_1|\theta)+P(z=B, y_1|\theta)} = \frac{(0.6)^5 * (0.4)^5}{(0.6)^5 * (0.4)^5 + (0.5)^{10}} = 0.45$

$P(z = B|y_1, \theta) = 1 - P(z = A|y_1, \theta) = 0.55$

计算出每个实验为硬币A和硬币B的概率，然后进行加权求和。

M步：求出似然函数下界 $Q(\theta, \theta^i)$， y_j代表第 j 次实验正面朝上的个数，μ_j 代表第 j 次实验选择硬币A的概率，$1 - \mu_j$ 代表第 j 次实验选择硬币B的概率 。

$$
\begin{aligned}
Q(\theta, \theta^i) &= \sum_{j=1}^{5} \sum_{z} P(z|y_j, \theta^i) log P(y_j, z|\theta) \\
&= \sum_{j=1}^{5} \mu_j log(\theta_A^{y_j}(1-\theta_A)^{10-y_j}) + (1-\mu_j) log(\theta_B^{y_j}(1-\theta_B)^{10-y_j})
\end{aligned}
$$

针对L函数求导来对参数求导，例如对 θ_A求导：

$$
\begin{aligned}
\frac{\partial Q}{\partial \theta_A} &= \mu_1(\frac{y_1}{\theta_A} - \frac{10-y_1}{1-\theta_A}) + \cdots + \mu_5(\frac{y_5}{\theta_A} - \frac{10-y_5}{1-\theta_A}) = \mu_1(\frac{y_1 - 10\theta_A}{\theta_A(1-\theta_A)}) + \cdots + \mu_5(\frac{y_5 - 10\theta_A}{\theta_A(1-\theta_A)}) \\
&= \frac{\sum_{j=1}^{5} \mu_j y_j - \sum_{j=1}^{5} 10\mu_j \theta_A}{\theta_A(1-\theta_A)}
\end{aligned}
$$

求导等于 0 之后就可得到图中的第一次迭代之后的参数值:

$\hat{\theta_A^{(1)}} = 0.71$

$\hat{\theta_B^{(1)}} = 0.58$

当然，基于Case a 我们也可以用一种更简单的方法求得：

$\hat{\theta_A^{(1)}} = \frac{21.3}{21.3+8.6} = 0.71$

$\hat{\theta_B^{(1)}} = \frac{11.7}{11.7+8.4} = 0.58$

第二轮迭代：基于第一轮EM计算好的 $\hat{\theta_A^{(1)}}, \hat{\theta_B^{(1)}}$，进行第二轮EM，计算每个实验中选择的硬币是 A 和 B 的概率（E 步），然后在计算M步得到$\hat{\theta_A^{(2)}} = 0.8, \hat{\theta_B^{(2)}} = 0.52$ ，如此继续迭代.....