EM 算法 期望最大化算法

- 最大似然
- EM算法推导
- GMM（高斯混合模型）

最大似然估计

- 一个栗子：假如你去赌场，但是不知道能不能赚钱，你就在门口堵着出来一个人就问一个赚了还是赔了，如果问了5个人都说赚了，那么你就会认为，赚钱的概率肯定是非常大的。

- 已知：（1）样本服从分布的模型，（2）观测到的样本
  求解：模型的参数

  总的来说：极大似然估计就是用来估计模型参数的统计学方法

最大似然数学问题（100名学生的身高问题）

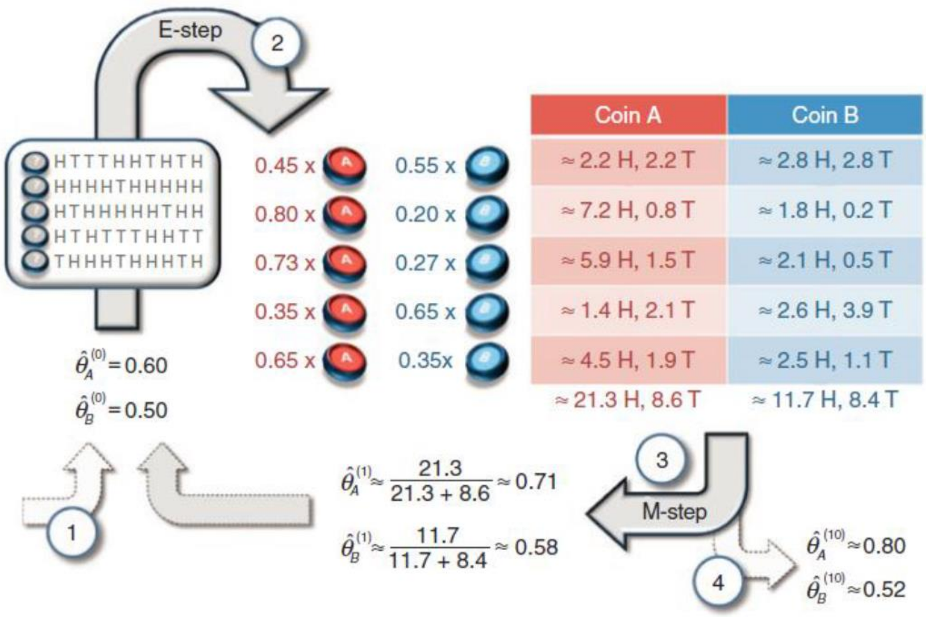
- 样本集X={x1,x2,...,xN} N=100
- 概率密度：p(xi|θ)抽到男生i（的身高）的概率
- θ是服从分布的参数
- 独立同分布：同时抽到这100个男生的概率就是他们各自概率的乘积

- 最大似然函数： $l(\theta)=\sum_{i=1}^{m} logp(x_i;\theta)$ （对数是为了乘法转加法）
- 什么样的参数 θ 能够使得出现当前这批样本的概率最大

- 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。

- 问题又难了一步
  - 现在这100个人中，不光有男生，还有女生（2个类别，2种参数）
  - 男生和女生的身高都服从高斯分布，但是参数不同（均值，方差）
  - 用数学的语言描述：抽取得到的每个样本都不知道是从哪个分布抽取的
  - 求解目标：男生和女生对应的身高的高斯分布的参数是多少
- 加入隐变量
  - 用Z=0或Z=1标记样本来自哪个分布，则Z就是隐变量。
  - 最大似然函数：$l(\theta)=\sum_{i=1}^{m} logp(x_i;\theta)=\sum_{i=1}^{m} log\sum_{Z} p(x_i,z;\theta)$
  - 求解：在给定初始值情况下进行迭代求解

通过迭代进行求解



两个硬币的初始假设的分布
A：0.6几率正面
B：0.5几率正面

投掷出5正5反的概率：
pA=C(10,5)*(0.6^5)*(0.4^5)
pB=C(10,5)*(0.5^5)*(0.5^5)

选择硬币A的概率：
pA/(pA+pB)=0.45
选择硬币B的概率
1- pA=0.55

EM算法推导

问题：样本集{x(1),...,x(m)}，包含m个独立的样本。
其中每个样本i对应的类别z(i)是未知的，所以很难用最大似然求解。

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

上式中，要考虑每个样本在各个分布中的情况。
本来正常求偏导就可以了，但是现在log后面还有求和，这就难解了！

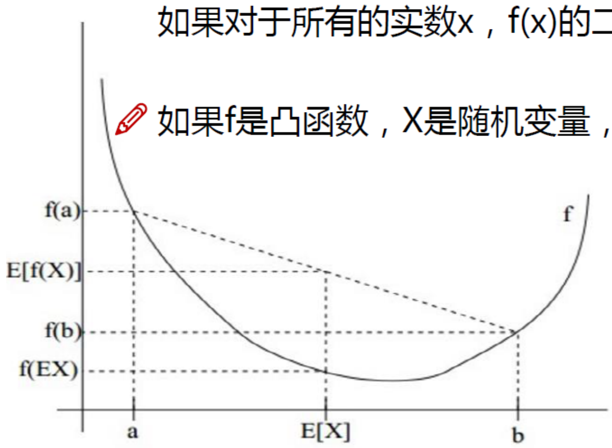
右式分子分母同时乘 $Q(z)$：$log\sum_{Z} p(x_i,z;\theta)=log\sum_{Z} Q(z)\frac{p(x_i,z;\theta)}{Q(z)}$

为嘛这么干呢？说白了就是要凑-Jensen不等式（ Q(z)是Z的分布函数）

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。

如果f是凸函数，X是随机变量，那么：E[f(X)]>=f(E[X])



实线f是凸函数，X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

Jensen不等式应用于凹函数时，不等号方向反向

由于$\sum_{Z} Q(z)\frac{p(x_i,z;\theta)}{Q(z)}$ 是 $\frac{p(x_i,z;\theta)}{Q(z)}$ 的期望

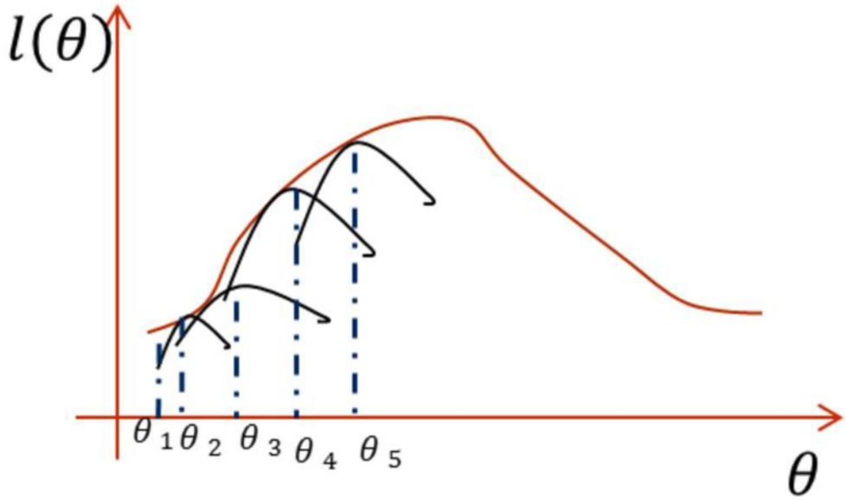
假设 $Y=\frac{p(x_i,z;\theta)}{Q(z)}$ 则：$log\sum_{Z} Q\frac{p(x_i,z;\theta)}{Q}=log\sum_{Y} P(Y)Y=logE(Y)$

可得：$logE(Y)\geq E(logY)=\sum_{Y} P(Y)logY=\sum_{Z} Q(z)log\frac{p(x_i,z;\theta)}{Q(z)}$

结论： $l(\theta)=\sum_{i=1}^{m}\log\sum_{Z}p(x_i,z;\theta) \geq \sum_{i=1}^{m}\sum_{Z}Q(z)\log\frac{p(x_i,z;\theta)}{Q(z)}$

优化下界

迭代到收敛



Jensen不等式

如何能使得等式成立呢？（取等号）

Jensen中等式成立的条件是随机变量是常数：$Y=\frac{p(x_i,z;\theta)}{Q(z)}=C$

Q（z）是z的分布函数：$\sum_Z Q(z)=\sum_Z\frac{p(x_i,z;\theta)}{C}=1$

所有的分子和等于常数C（分母相同）

Q(z)求解

$\sum_Z Q(z)=\sum_Z\frac{p(x_i,z;\theta)}{C}=1$

由上式可得C就是p(xi,z)对z求和

$Q(z)=\frac{p(x_i,z;\theta)}{C}=\frac{p(x_i,z;\theta)}{\sum_Z p(x_i,z;\theta)}=\frac{p(x_i,z;\theta)}{p(x_i)}=p(z|x_i;\theta)$

Q(z)代表第i个数据是来自zi的概率

- EM算法流程
  - 初始化分布参数θ
  - E-step：根据参数θ计算每个样本属于zi的概率(也就是我们的Q)
  - M-Step：根据Q，求出含有θ的似然函数的下界并最大化它，得到新的参数θ
  - 不断的迭代更新下去
- GMM（高斯混合模型）
  - 数据可以看作是从数个 Gaussian Distribution 中生成出来的
  - GMM 由 K 个 Gaussian 分布组成，每个 Gaussian 称为一个“Component”
  - 类似k-means方法，求解方式跟EM一样
  - 不断的迭代更新下去