EM算法（Expection — Maximum）

(1) Jensen不等式

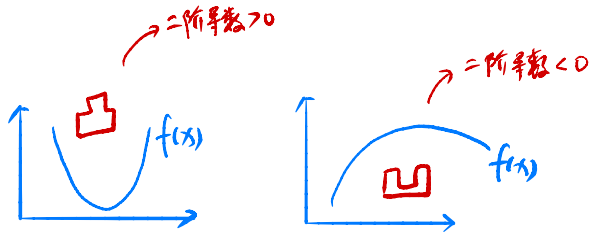
满足概率性质

当 $f(x)$ 是凸函数，且 $\sum_i^k \theta_i = 1$，$\theta_i \in [0,1]$ 时

$$\underbrace{f(\theta_1 x_1 + \cdots + \theta_k x_k)}_{A} \le \underbrace{\theta_1 f(x_1) + \cdots + \theta_k f(x_k)}_{B}$$

当 $f(x)$ 是凹函数，且 $\sum_i^k \theta_i = 1$，$\theta_i \in [0,1]$ 时

$$\underbrace{f(\theta_1 x_1 + \cdots + \theta_k x_k)}_{A} \ge \underbrace{\theta_1 f(x_1) + \cdots + \theta_k f(x_k)}_{B}$$

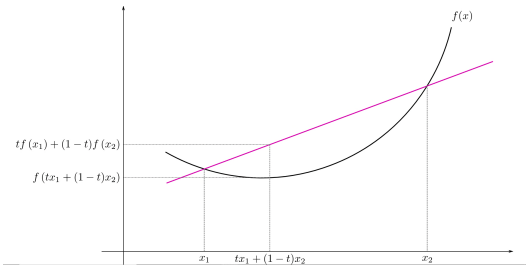


从统计学角度看：$A = f(E(x))$，$B = E[f(x)]$

$\Rightarrow$（f为凸函数时）$A \le B$

$\Rightarrow$（f为凹函数时）$A \ge B$

当一个复杂函数难以直接求解时，我们一般会通过Jensen不等式来构建其下界函数。
然后通过求解下界函数的值来逼近原函数的解。



(2) 最大似然估计（MLE）

$$\max_{\theta} \arg \prod_{i=1}^{m} P(x;\theta)$$

$$\Rightarrow l(\theta) = \sum_{i=1}^{m} \log P(x;\theta)$$

引入隐变量

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

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

$$= \underbrace{\sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \ge \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}}$$

① log函数是凹函数

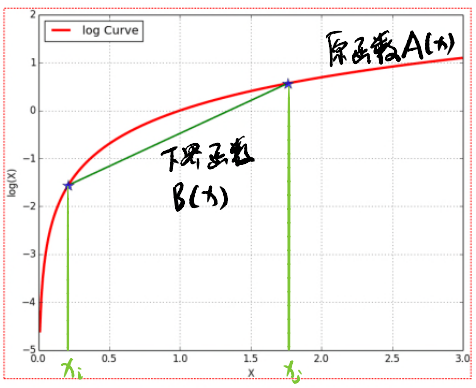
② $\sum_{z^{(i)}} Q_i(z^{(i)}) = 1$

根据Jensen不等式，上式成立 $\Rightarrow f[E(Y)] \ge E[f(Y)]$

$f(Y) = \log(Y)$，$Y = \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$

是我们为了使Jensen不等式成立而给定的约束条件

(3) 寻找最佳下界函数



理想化的最佳下界函数当然是 $B(x) = A(x)$

但是根据Jensen不等式，仅当无论 x 取什么值，函数值 $B(x)$ 只是 $A(x)$ 上的一个点时才成立

$$\Rightarrow \text{即} \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} = C$$

一个常数

无论自变量 $z^{(i)}$ 如何变化，这个表达式的值都为某一个常数

（注意：这里的 $x^{(i)}$，θ 为给定的值）

$$\Rightarrow P(x^{(i)}, z^{(i)};\theta) = C\left[Q_i(z^{(i)})\right]$$

$$\Rightarrow \sum_z P(x^{(i)}, z^{(i)};\theta) = C\left[\underbrace{\sum_z Q_i(z^{(i)})}_{=1}\right] = \sum_z P(x^{(i)}, z^{(i)};\theta) = C$$

将C的值代入表达式 $\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}=C$

对应不同的$z^{(i)}$分类，都会有对应的概率值输出

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

是一个关于随机变量$z^{(i)}$的分布

比如一维水果中有三个类别：苹果、桔子、西瓜，那么

$$P(z^{(i)}=\text{苹果}|x^{(i)};\theta)+P(z^{(i)}=\text{桔子}|x^{(i)};\theta)+P(z^{(i)}=\text{西瓜}|x^{(i)};\theta)=1$$

※ 当$Q_i(z^{(i)})=P(z^{(i)}|x^{(i)};\theta)$时，可取得最佳下界

(4) E-M算法（期望最大化）

Repeat until convergence {

(E-step) For each i, set

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

在给定模型参数θ及训练数据$x^{(i)}$的前提下，关于隐变量$z^{(i)}$的分布

$x^{(i)}$是训练集中第i个数据

θ是模型参数（可先经验或随机进行初始化）

通过这个条件，确保下面的这个下界函数是最佳的

(M-step) Set

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

寻找一个恰当的θ值，使得这个下界函数取最大值

即为给定训练数据$x^{(i)}$及模型参数θ的条件下，服从Q_i分布的隐变量$z^{(i)}$的期望值

}

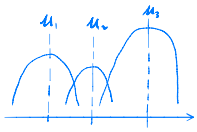
$$=\arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)})\cdot \log P(x^{(i)},z^{(i)};\theta) - Q_i(z^{(i)})\,Q_i(z^{(i)})$$

寻找θ，使得这一部分的值最大即可

$z^{(i)}$在Q_i分布下的信息熵（当给定Q_i和$z^{(i)}$后，它是一个确定的值）

求得$z^{(i)}$和$Q_i(z^{(i)})$后，代入公式后可更新θ值

如此多次迭代，直到隐变量z、模型参数θ不再更新。



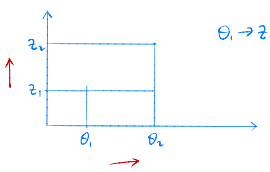
总结：

(1) 含有隐变量的求极值问题，一般都可以使用EM算法求求解。 → 局部最优解

(2) 不同的初始值，可能会带来不同的极值点。 → 因为有些模型可能有多个极值点，比如高斯混合模型(GMM)

(3) EM算法的本质，是通过坐标上升的方法来逼近真实值。

Coordinate Ascent



$\theta_1 \to z_1 \to \theta_2 \to z_1$