输入 $\{X_i\}_{i=1\sim N}$ 样本

定义 $\{Z_i\}_{i=1\sim N}$ 隐含变量

目的：最大化 $E(\theta)=\sum_{i=1}^{N}\log[p(x_i/\theta)]$

$$=\sum_{i=1}^{N}\log[\sum_{z_i}p(x_i,z_i/\theta)]$$ (最大似然估计)

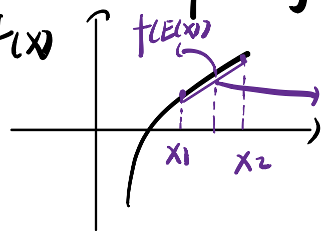
设 $Q_i(z_i)$ 为 z_i 的概率分布

$$\sum_{z_i}Q_i(z_i)=1$$

$$E(\theta)=\sum_{i=1}^{N}\log\left[\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right]$$

根据 Jenson's Inequality

若 $f(x)$



则 $f(E(x))\geq E(f(x))$

$$E(\theta)\geq\sum_{i=1}^{N}\sum_{z_i}Q_i(z_i)\log\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}$$

当且仅当 $Q_i(z_i)$ 与 $p(x_i,z_i/\theta)$ 成比例时等号成立.

当 $Q_i(z_i)=\dfrac{p(x_i,z_i/\theta)}{\sum_{z_i}p(x_i,z_i/\theta)}$ 时

$E(\theta)$ 取最大值. $\sum_{i=1}^{N}\sum_{z_i}Q_i(z_i)\log\dfrac{p(x_i,z_i/\theta)}{Q_i(z_i)}$

EM算法的一般形式

① 随机选择 θ_0

② E-Step:

$$Q_i(Z_i) = \frac{p(x_i, z_i \mid \theta)}{\sum_{z_i} p(x_i, z_i \mid \theta)}$$

③ 固定 $Q_i(Z_i)$，求 $\theta^{(new)}$

$$\theta_{k+1} = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{z_i} Q_i(Z_i) \log \frac{p(x_i, z_i \mid \theta_k)}{Q_i(Z_i)}$$

④ 回到② 循环至收敛

EM算法收敛性的证明

设 $M(\theta) = \sum_{i=1}^{N} \sum_{z_i} Q_i(Z_i) \log \frac{p(x_i, z_i \mid \theta)}{Q_i(Z_i)}$

则②做完后 $E(\theta_k) = M(\theta_k)$

③做完后 $M(\theta_{k+1}) \geq M(\theta_k)$

接着②… $E(\theta_{k+1}) = M(\theta_{k+1}) \geq M(\theta_k) = E(\theta_k)$

$E(\theta) \leq 0$，有上界

以K-均值算法为例.

输入 $\{x_i\}_{i=1\sim N}$ 样本，待求 $\theta = \{u_1, u_2, \dots u_k\}$

① 定义 $\{z_i\}_{i=1\sim N}$ $z_i = \{1, 2, \dots k\}$

② 定义 $p(X, Z \mid \theta) = \frac{1}{(\sqrt{2\pi})^d} \exp\{-\frac{\|X - u_Z\|^2}{2}\}$

$$\begin{cases} \quad \text{当 } z=\arg\min_j \|x-u_j\| \text{ 时} \\ 0 \quad \text{其它} \end{cases}$$

① 随机取 $\theta=\{u_1, u_2, \dots, u_k\}$

② $$Q_i(z_i)=\frac{p(x_i,z_i/\theta_k)}{\sum_{z_i} p(x_i,z_i/\theta_k)}=\begin{cases} 1, & \text{当 } z_i=\arg\min_j \|x_j-u_j\| \text{ 时} \\ 0 & \text{其它} \end{cases}$$

③ $$\theta_{k+1}=\arg\max_{\theta} \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i) \frac{\log p(x_i,z_i/\theta_k)}{Q_i(z_i)}$$

$$u_j^{(new)}=\arg\max_{u_j} \sum_{i=1}^{N}\sum_{z_i=j} \log p(x_i,j/\theta_k)$$

设 $$E(u_j)=\sum_{\substack{i=1\\z_i=j}}^{N} \log p(x_i,j/\theta_k)$$

$$=-\sum_{\substack{i=1\\z_i=j}}^{N} \frac{\|x_i-u_j\|^2}{2} - \text{常数}$$

$$\frac{\partial E}{\partial u_j}=-\sum_{\substack{i=1\\z_i=j}}^{N}(x_i-u_j)=0$$

$$u_j=\frac{\sum_{\substack{i=1\\z_i=j}}^{N} x_i}{\sum_{\substack{i=1\\z_i=j}}^{N} 1}$$

(所有属于第j类的x求均值)

(1) EM算法求局部极值