引入 | 从缺失数据入手

- 观察data X
  缺失data Z
  完全data $Y=(X,Z)$
  参数 θ.

似然：
- $L(\theta|X)=P(X|\theta)$ → 直接求解困难.
  $L(\theta|Y)=P(Y|\theta)$ → 简单， 容易计算. 利用
  $P(Z|Y,\theta)$ → 简单

计算已知完全data下θ的后验期望.　已知$\theta^{(t)}$下，在完全数据下θ的期望.　完全似然　缺失data.

E-step
- $Q(\theta|\theta^{(t)})=\int \log L(\theta|Y)\,p(z|X,\theta^{(t)})$
  $=E[\int \log L(\theta|Y)\,|\,X,\theta^{(t)}]$.

M-step
- $\theta^{(t+1)}=\arg\max\limits_{\theta} Q(\theta|\theta^{(t)})$,

极大似然 ↓ Q函数.
- $\log P(X|\theta)=\int \log P(x,z|\theta)\,dz$
  $=\log\int \frac{P(x,z|\theta)}{q(z)}q(z)\,dz$　令$q(z)=P(z|x,\theta^{(t)})$.
  $\geq \int \log\frac{P(x,z|\theta)}{P(z|x,\theta^{(t)})}\cdot P(z|x,\theta^{(t)})\,dz$
  $=\underbrace{\int \log P(x,z|\theta)\,P(z|x,\theta^{(t)})\,dz}_{Q(\theta,\theta^{(t)})}-\underbrace{\int \log P(z|x,\theta^{(t)})\,dz}_{与\theta无关\ const.}$

案例 k-means

- 观察数据：每个点坐标 X
  未知数据：每个点的标签 Z

- 先初始化 k 个中心 θ

分布的概率 E-step
- 有 θ 中心，然后对 $X \longrightarrow Z$ 组成 $Y=(X,Z)$ 完整数据

参数 M-step
- 有了 $\theta^{(t)}$，和 $Y=(X,Z)$ 数据，
  现在求 $\theta^{(t+1)}$ 使
  $$\log P(Y \mid \theta^{(t)}) \text{ 最大.}$$

举例1

- $Y_1, Y_2 \overset{iid}{\sim} Exp(\theta) = \theta \exp\{-y\theta\}$

Y_2为缺失的数据.

$Y_1 = 5 \qquad Y_2 = ?$

观察data. $p(y_1|\theta) = \theta \exp\{-y_1\theta\}$

$\log p(y_1|\theta) = \log\theta - y_1\theta.$

最大似然估计: $\frac{1}{\theta} - y_1 = 0 \qquad \underline{\theta = \frac{1}{y_1}} = 0.2$

2种方法同解. 但EM解出了缺失数据

完全data. $p(y_1, y_2|\theta) = \theta^2 \exp\{-(y_1+y_2)\theta\}$

$\log p(y_1, y_2|\theta) = 2\log\theta - (y_1+y_2)\theta$

EM.

$p(y_2|y_1, \theta^{(t)}) = p(y_2|\theta^{(t)}) = \theta^{(t)} \exp\{-y_2\theta^{(t)}\}$

$Q(\theta|\theta^{(t)}) = E[\log p(y_1, y_2|\theta) \mid y_1, \theta^{(t)}]$

$= 2\log\theta - y_1\theta - \theta \cdot E[y_2]$

E-step: $\underline{\theta = \frac{2\theta^{(t)}}{5\theta^{(t)}+1}}$

最终 $\theta^* = \frac{2\theta^*}{5\theta^*+1} \Rightarrow \theta^* = 0.2$

举例2 —原论文中

最大似然法

- $\vec{y} = (y_1, y_2, y_3, y_4)^T$ 多项式分布.

  $\frac{1}{2}+\frac{1}{4}\theta$， $\frac{1}{4}(1-\theta)$， $\frac{1}{4}(1-\theta)$， $\frac{1}{4}\theta$

  $\vec{y} = (125, 18, 25, 34)^T$， $n=197$

  $P(\vec{y}\mid\theta) = \frac{n!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{1}{4}\theta)^{y_1}[\frac{1}{4}(1-\theta)]^{y_2}[\frac{1}{4}(1-\theta)]^{y_3}(\frac{1}{4}\theta)^{y_4}$.

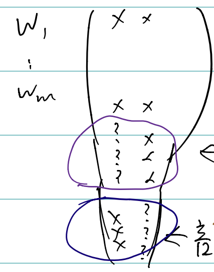
  $\log P(\vec{y}\mid\theta) = y_1\log(\frac{2+\theta}{4}) + y_2\log\frac{1}{4}(1-\theta) + y_3\log(1-\theta)\cdot\frac{1}{4} + y_4\log\frac{1}{4}\theta$

  $\frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta} = 0$

案例

$W=(W_1, W_2)^T$. $\quad W = N(\mu, \Sigma)$ $\quad \mu=\begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}$ $\quad \Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$

$$p(W|\theta) = \frac{1}{(2\pi)|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(W-\mu)^T\Sigma^{-1}(W-\mu)\right\}$$



← 都有 $\quad W_1 \cdots W_m$

← 部分值缺失 $\quad W_{2j}$ ($j=m, m+1, \cdots m+m_1$) 第1个缺失,

← 部分值缺失 $\quad W_{1j}$ ($j=m_1, \cdots, N$) 第2个缺失.

通过后验分布而期望补全数据.

$\vec{y} = (y_1, \cdots, y_m)$

$$\log L(\theta|y) = -n\log(2\pi) - \frac{1}{2}n\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\vec{w}_i-\mu)^T\Sigma^{-1}(\vec{w}_i-\mu)$$

$$|\Sigma| = \begin{vmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{vmatrix} = \Sigma_{11}\Sigma_{22}-\Sigma_{21}\Sigma_{12}$$

$T_i = \sum_{j=1}^{n} w_{ij} \quad (i=1,2,\cdots)$