Кластеризация строк

$\mathcal{N}(\bar x|\mu_k,\Sigma_k)$

ACGTAATGCGT $\in \{ACGT\}^n$

ACGTTATGCGT

$i=$ ACGTAATGCGT ($1 \ldots n$)

CGTAATGGCGT

$$p(\bar x)=\sum_k \underbrace{\pi_k}_{p(C_k)}\, p(\bar x|C_k)$$

$p(\bar x|\bar \mu)=\prod p_j$ ограничения

$$p(\bar x|p_{isk}) = \prod_{i=1}^{n} p_{i,x_i,k}$$

$$p_{isk}=p(x_i=s|C_k), \quad s\in ACGT$$

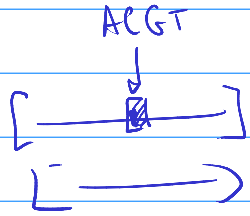
E-шаг: $$\mathbb{E}z_{nk}=\frac{p(C_k)p(\bar x_n|C_k)}{\ldots}=\frac{\pi_k\prod_i p_{i,x_{ni},k}}{\sum_{k'} \ldots_{k'}}$$

$z_{nk}=[\bar x_n\in C_k]$

$$p(D|p\ldots)=\prod_n\Big(\sum_k \pi_k \prod_i p_{i,x_{ni},k}\Big)$$

$$p(X,Z|\_)=\prod_n\prod_k(\ldots)^{z_{nk}}$$

M-шаг: $$Q(\theta,\theta^{(m)})=\mathbb{E}_{Z|\theta^{(m)}}\big[\log p(X,Z|\theta)\big]=$$



$$=\sum_n\sum_k \mathbb{E}z_{nk}\Big(\log\pi_k+\sum_i \log p_{i,x_{ni},k}\Big)\xrightarrow[p_{i,s,k}]{}\max$$

$$=\Big(\sum_{n,k}\mathbb{E}z_{nk}\log\pi_k\Big)+\sum_k\sum_i\Big[\sum_n \mathbb{E}z_{nk}\log p_{i,x_{ni},k}\Big]$$

$\downarrow \pi_k$ max $\qquad$ $\downarrow p_{isk}\ \forall i,k \quad \sum_s p_{isk}=1$ max

M-шаг

$$\pi_k^{(m+1)}=\frac{\sum_n \mathbb{E}z_{nk}^{(m+1)}}{N}$$

$$p_{isk}^{(m+1)}=\frac{\sum_{n:x_{ni}=s}\mathbb{E}z_{nk}^{(m+1)}}{\sum_n \mathbb{E}z_{nk}^{(m+1)}}$$

$$\sum_n \mathbb{E}z_{nk}\log p_{-}=\sum_s\Big[\sum_{n:x_{ni}=s}\mathbb{E}z_{nk}\Big]\log p_{isk}\to\max$$

E-шаг

$$z_{nk}^{(m+1)}:=\frac{\pi_k^{(m)}\prod_i p^{(m)}_{i,x_{ni},k}}{\sum_{k'}\pi_{k'}^{(m)}\prod_i p^{(m)}_{i,x_{ni},k'}}$$

$\log a_k \rightsquigarrow \dfrac{a_k}{\sum a_k}$

$l_k \rightsquigarrow \dfrac{e^{l_k}}{\sum e^{l_n}}$

-20
$-21 \quad -(-20) \to \quad 0, \; -1, \; -5$
-25

$l_k - \max l_k$

$\bar{x}_n \mapsto t_n \in \{0,1\}$

$p(y_n|\bar{x}_n) = \sigma(\bar{w}^T \bar{x}_n)$

presence-only data

$$\boxed{\begin{array}{l} y_n = 1 \Rightarrow z_n = 1 \\ y_n = 0 \to z_n = ? \end{array}}$$

Prospective vs retrospective studies

control group

$$\pi_{ab} = p(x=a, d=b)$$

	$d=0$	$d=1$	
$x=0$	π_{00} 0.70	π_{01} 0.02	$\pi_{0*} = 0.72$
$x=1$	π_{10} 0.26	π_{11} 0.03	$\pi_{1*} = 0.28$
	$\pi_{*0} = 0.95$	$\pi_{*1} = 0.05$	

$p(d \mid \bar{x})$ $\qquad$ $p(d=1 \mid x=0) \approx \pi_{01}$

$$\pi_0 = p(z=1 \mid d=0)$$
$$\pi_1 = p(z=1 \mid d=1)$$

$$p(d \mid \bar{x}, z=1)$$
$$\parallel$$
$$\underbrace{p(z=1 \mid d, \cancel{\bar{x}})}_{\pi_1} p(d \mid \bar{x})$$
$$p(z=1 \mid d, \bar{x}) p(d \mid \bar{x}) + p(z=1 \mid \bar{d}, \bar{x}) p(\bar{d} \mid \bar{x})$$

$\bar{w}, w_0$:

$$p(d=1 \mid \bar{x}) \approx \sigma(\bar{w}^T \bar{x} + w_0)$$

$$\pi_1 = p(z=1 \mid d=1, \bar{x}) p(d=1 \mid \bar{x}) + p(z=1 \mid d=0, \bar{x}) p(d=0 \mid \bar{x})$$

$$p(d=1 \mid z=1, \bar{x}) = \frac{\pi_1 \, p(d=1 \mid \bar{x})}{\pi_1 \cdot p(d=1 \mid \bar{x}) + \pi_0 \, p(d=0 \mid \bar{x})} =$$

$$= \frac{\pi_1 \sigma(\bar{w}^T \bar{x} + w_0)}{\pi_1 \sigma(-) + \pi_0 (1-\sigma)} = \frac{\frac{\pi_1}{\pi_0} e^{\bar{w}^T \bar{x} + w_0}}{1 + \frac{\pi_1}{\pi_0} e^{\bar{w}^T \bar{x} + w_0}} =$$

$$\sigma(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$

$$1 - \sigma = \frac{1}{1+e^a}$$

$$= \frac{e^{\bar{w}^T \bar{x} + w_0 + \log \frac{\pi_1}{\pi_0}}}{1 + e^{\cdots}}$$

$$p(d=1 \mid z=1, \bar{x}) = \sigma\left(\bar{w}^T \bar{x} + w_0 + \log \frac{\pi_1}{\pi_0}\right)$$

$$p(d=1 \mid z=1, \bar{x}) = \sigma\left(\bar{\beta}^T \bar{x} + \alpha\right)$$

$\bar{\beta}^*, \alpha^*$

$$\begin{cases} \bar{w}^* = \bar{\beta}^* \\ \alpha^* = w_0^* + \log \dfrac{\pi_1}{\pi_0} \end{cases}$$