Buổi 1. (13/08/2020)

1. K-means

+) Giả thiết:

- 1 điểm dữ liệu thuộc 1 cụm
- Các điểm dữ liệu trong 1 cụm gần nhau

+) Input:

- $\{x_n\}_{n=1}^{N}$ $\quad x_n \in R^n$
- K: the number of cluster.

+) Output:

- $\{\mu_k\}_{k=1}^{K}$ $\quad (\mu_k \in R^n)$
- $\{z_n^k\}_{n,k=1}^{N,K}$ $\quad z_n^k = \begin{cases} 1 & \text{nếu } n \in \text{cụm } k \\ 0 & \text{_____} \notin \text{_____} \end{cases}$

$$\sum_{k=1}^{K} z_n^k = 1$$

$$\underset{\mu, z}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{k=1}^{K} z_n^k \|x_n - \mu_k\|^2$$

+) Quá trình tối ưu.

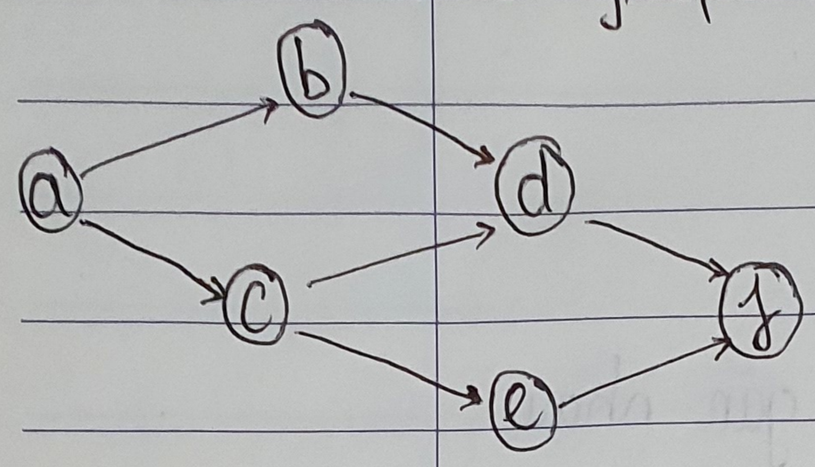
1) Khởi tạo $\{\mu_k\}_{k=1}^{K}$. Tìm $\underset{z_n^k}{\operatorname{argmin}} \sum_{k=1}^{K} z_n^k \|x_n - \mu_k\|^2$

2) Cố định $\{z_n^k\}$. Tìm $\underset{\mu_k}{\operatorname{argmin}} \sum_{n=1}^{N} z_n^k \|x_n - \mu_k\|^2$

$$\frac{\partial f}{\partial \mu_k} = -2 \cdot \sum_{n=1}^{N} z_n^k (x_n - \mu_k) = 0.$$

$$\Leftrightarrow \mu_k = \frac{\sum_{n=1}^{N} z_n^k \cdot x_n}{\sum_{n=1}^{N} z_n^k}.$$

2. Probabilistic graphical models. (PGM)



$a \sim P(a)$

$b \sim P(b|a)$

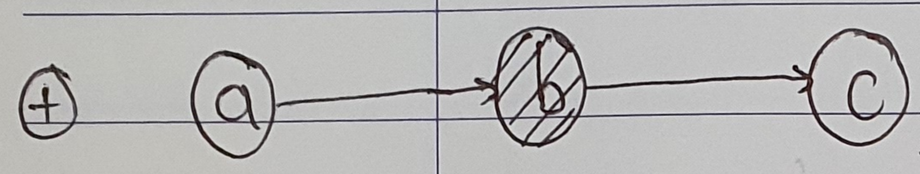
$c \sim P(c|a)$

$d \sim P(d|b,c)$

$e \sim P(e|c)$

$f \sim P(f|e,d)$

$\Rightarrow P(a,b,c,d,e,f) = P(a).P(b|a).P(c|a).P(d|b,c).P(e|c).P(f|e,d)$

+ 

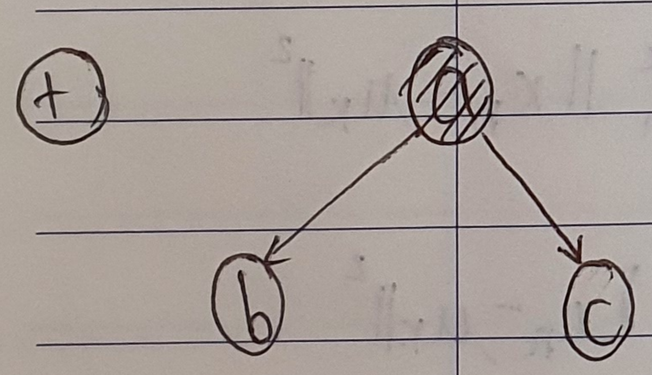
$a \perp c | b$

Hay: $P(a,c|b) = P(a|b).P(c|b)$

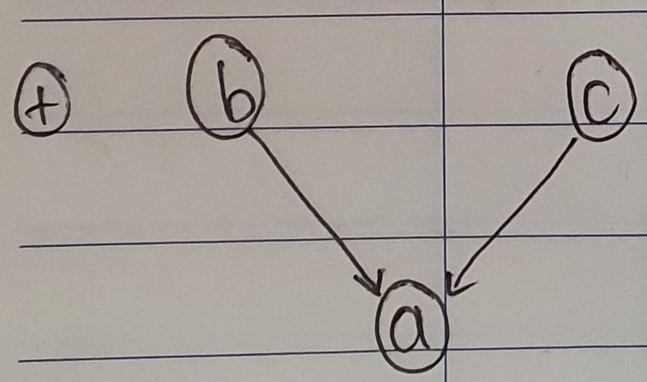
$P(a|b,c) = P(a|b)$.

Chứng minh:

$$P(a,c|b) = \frac{P(a,b,c)}{P(b)} = \frac{P(a).P(b|a).P(c|b)}{P(b)} = P(a|b).P(c|b)$$

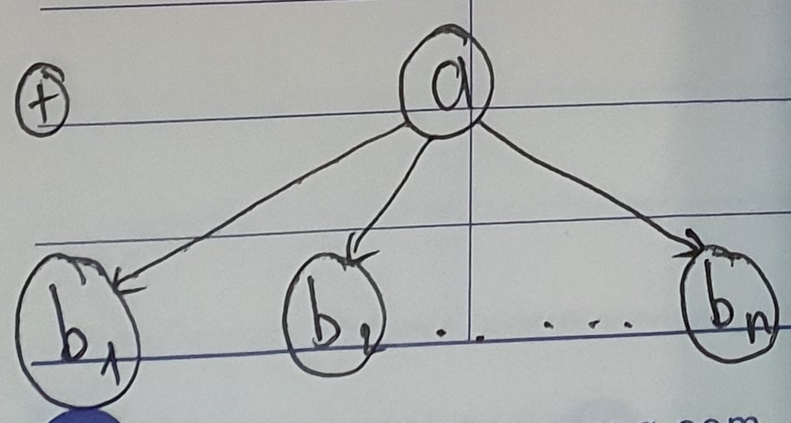
+ 

$b \perp c | a$.

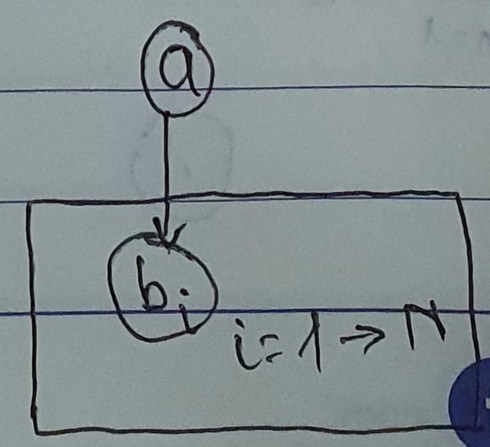
+ 

$b \perp c$

$b \not\perp c | a$

+ 

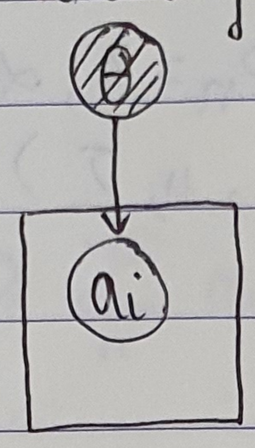
$\Leftrightarrow$



⊕ Exchangeability.

$$P(a,b,c) = P(b,a,c) = P(c,a,b).$$

Khi $a_i, (i=1 \to N)$ là exchangeability thì có thể vẽ:



$\exists \theta$ sao cho: $P(a,b,c|\theta) = P(a|\theta).P(b|\theta).P(c|\theta)$

3. Gaussian Mixture Model (GMM)

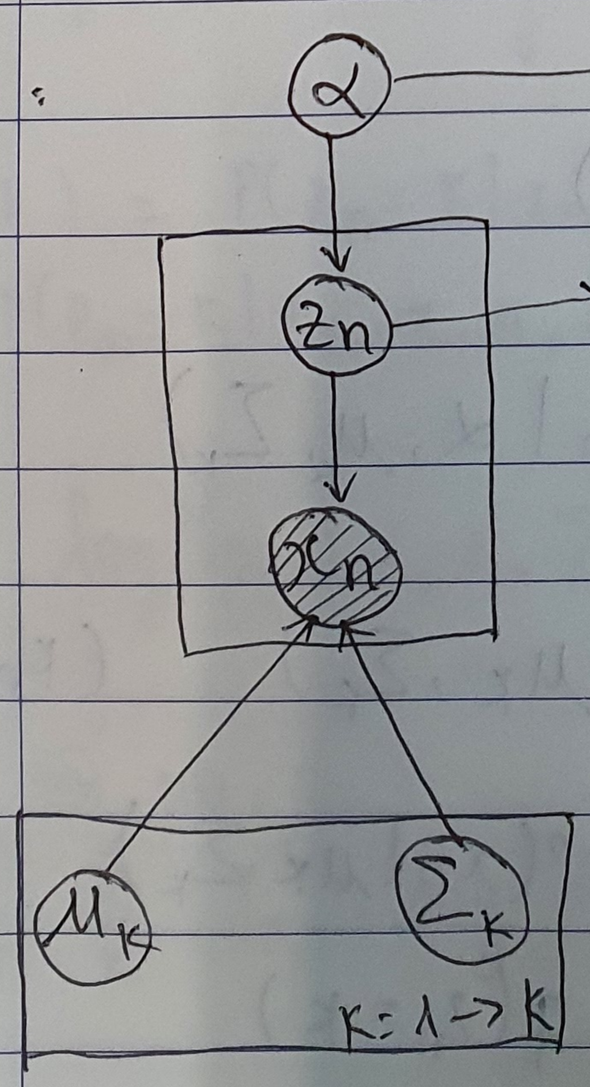
- Input: $\{x_n\}_{n=1}^{N}$ $\quad x \in R^n$

  K: the number of cluster

- Ouput: $z_n \Rightarrow P(x_n = K | z_n)$ $\quad (z_n = \{1, \dots K\})$

  $\{\mu_K, \Sigma_K\}_{K=1}^{K}$

- Giả thiết:



hyperparameter

hidden variable

$z_n \sim P(\alpha)$: Multinomial (α)

$P(z_n = K | \alpha) = \alpha_K \quad (\sum_{K=1}^{K} \alpha_K = 1)$

$x_n \sim P(\dots | z_n, \mu, \Sigma)$

$x_n \sim N(\dots | \mu_{z_n}, \Sigma_{z_n})$

⊛ Tối ưu.

$$P(z_n=k \mid x_n, \mu, \Sigma, \alpha) = \frac{P(z_n=k, x_n \mid \alpha, \mu, \Sigma)}{P(x_n \mid \alpha, \mu, \Sigma)} = \frac{P(z_n=k, x_n \mid \alpha, \mu, \Sigma)}{C}$$

$$\propto P(z_n=k, x_n \mid \alpha, \mu, \Sigma)$$

$$\propto P(z_n=k \mid \alpha, \mu, \Sigma) \cdot P(x_n \mid z_n=k, \alpha, \mu, \Sigma)$$

$$\propto P(z_n=k \mid \alpha) \cdot P(x_n \mid z_n=k, \mu, \Sigma)$$

$$\propto \alpha_k \cdot N(x_n \mid \mu_{z_n}, \Sigma_{z_n})$$

Với $C = P(x_n \mid \alpha, \mu, \Sigma) = \sum_k \alpha_k \cdot N(x_n \mid \mu_k, \Sigma_k)$.

$$\Rightarrow \boxed{P(z_n=k \mid x_n, \mu, \Sigma, \alpha) = \frac{\alpha_k \cdot N(x_n \mid \mu_z, \Sigma_z)}{\sum_k \alpha_k \cdot N(x_n \mid \mu_k, \Sigma_k)} \propto \alpha_k \cdot N(x_n \mid \mu_z, \Sigma_z)}$$

⊛ Maximum likelihood estimation (MLE)

$$\underset{\alpha, \mu, \Sigma}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(x_n \mid \alpha, \mu, \Sigma)$$

$$= \sum_{n=1}^{N} \log \sum_{k=1}^{K} P(z_n=k, x_n \mid \alpha, \mu_k, \Sigma_k)$$

$$= \sum_{n=1}^{N} \log \sum_{k=1}^{K} \alpha_k \cdot P(x_n \mid \mu_k, \Sigma_k). \quad \text{(Khó tối ưu)}$$

$$= \sum_{n=1}^{N} \log \sum_{k=1}^{K} q(z_n=k) \cdot \frac{\alpha_k \cdot P(x_n \mid \mu_k, \Sigma_k)}{q(z_n=k)}$$

$$\geq \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \cdot \log \frac{\alpha_k \cdot P(x_n \mid \mu_k, \Sigma_k)}{q(z_n = k)}$$

(lower bound)

Bất đẳng thức Jensen:

Với $\sum_n a_n = 1$:

$$\log \sum_n a_n \cdot x_n \geq \sum_n a_n \cdot \log x_n.$$

Dấu "=" xảy ra $\Leftrightarrow$: $p(z_n = k) = P(z_n = k \mid x_n, \alpha, \mu_k, \Sigma_k)$

$$= \frac{P(z_n = k, x_n \mid \alpha, \mu_k, \Sigma_k)}{P(x_n \mid \alpha, \mu_k, \Sigma_k)}$$

$$= \frac{\alpha_k \cdot N(x_n \mid \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \alpha_k \cdot N(x_n \mid \mu_k, \Sigma_k)} = \gamma_n^k$$

(Kiểm chứng:

$$VP = \sum_{n=1}^{N} \sum_{k=1}^{K} q(z_n = k) \cdot \log \frac{P(z_n = k, x_n \mid \alpha, \mu_k, \Sigma_k)}{q(z_n = k)}$$

với $q(z_n = k) = P(z_n = k \mid x_n, \alpha, \mu, \Sigma)$ có:

$$VP = \sum_n \sum_k P(z_n = k \mid x_n, \alpha, \mu, \Sigma) \cdot \log \frac{P(z_n = k, x_n \mid \alpha, \mu, \Sigma)}{P(z_n = k \mid x_n, \alpha, \mu, \Sigma)}$$

$$= \sum_n 1 \cdot \log P(x_n \mid \alpha, \mu, \Sigma) = VT \Rightarrow \text{đpcm})$$

Lúc này cần tối ưu:

$$\underset{\alpha, \mu, \Sigma}{\arg\max} \sum_n \sum_k \gamma_n^k \cdot \log \frac{P(z_n = k, x_n \mid \alpha, \mu, \Sigma)}{\gamma_n^k}$$

$$= \underset{\alpha, \mu, \Sigma}{\arg\max} \sum_n \sum_k \gamma_n^k \log \alpha_k \cdot N(x_n | \mu_k, \Sigma_k)$$

$$= \arg\max \sum_n \sum_k \gamma_n^k \left(\log \alpha_k + \log N(x | \mu_k, \Sigma_k) \right).$$

$$\left(\text{Với}, \; N(x|\mu, \Sigma) = \frac{1}{(2\pi)^D \cdot |\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2} \cdot (x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu) \right). \right).$$

(D là số chiều)

$$= \underset{\alpha, \mu, \Sigma}{\arg\max} \sum_n \sum_k \gamma_n^k \cdot \left(\log \alpha_k + \left(-\frac{1}{2} \cdot (x_n - \mu_k)^T \cdot \Sigma_k^{-1} \cdot (x_n - \mu_k) - \frac{1}{2} \log \Sigma_k - \frac{D}{2} \cdot \log 2\pi \right) \right)$$

+) cố định μ, Σ.

$$F(\alpha) = \sum_n \sum_k \gamma_n^k \cdot \log \alpha_k - \lambda \left(\sum_k \alpha_k - 1 \right).$$

$$\frac{\partial F(\alpha)}{\partial \alpha_k} = \frac{\sum_n \gamma_n^k}{\alpha_k} - \lambda = 0.$$

$$\Leftrightarrow \sum_n \gamma_n^k = \alpha_k \cdot \lambda. \Rightarrow \sum_n \sum_k \gamma_n^k = \sum_k \alpha_k \cdot \lambda = \lambda.$$

$$\Rightarrow \sum_n \underbrace{\left(\sum_k \gamma_n^k \right)}_{1} = \lambda \Rightarrow \lambda = \sum_n 1 = N.$$

$$\Rightarrow \boxed{\alpha_k = \frac{\sum_n \gamma_n^k}{N}}$$

+) $$F(\Sigma_k) = \sum_n \gamma_n^k \left((x_n - \mu_k)^T (x_n - \mu_k) \cdot \Sigma_k^{-1} + \log \Sigma_k \right)$$

$$\frac{\partial F(\Sigma_k)}{\partial \Sigma_k} = \sum_n \gamma_n^k \cdot \left(\frac{-(x_n - \mu_k)^T (x_n - \mu_k)}{\Sigma_k^2} + \frac{1}{\Sigma_k} \right) = 0$$

$$\Leftrightarrow \boxed{\Sigma_k = \frac{\sum_n \gamma_n^k (x_n - \mu_k)^T (x_n - \mu_k)}{\sum_n \gamma_n^k}}$$

svmc.recruit@samsung.com SamsungVietnam RnD

+) $F(\mu_k) = \sum_n \gamma_n^k (x_n - \mu_k)^T . (x_n - \mu_k) . \Sigma_k^{-1}$

$\frac{\partial F(\mu_k)}{\partial \mu_k} = -2 . \sum_n \gamma_n^k (x_n - \mu_k) . \Sigma_k^{-1} = 0$

$\Leftrightarrow \boxed{\mu_k = \frac{\sum_n \gamma_n^k . x_n}{\sum_n \gamma_n^k}}$

Bước 2.

- $\int p(x) . q(x) dx = \mathop{E}_{x \sim q(x)} . P(x) = \mathop{E}_{x \sim p(x)} q(x)$. ($x$ liên tục)
- $\sum p(x) . q(x) = \mathop{E}_{x \sim p(x)} q(x) = \mathop{E}_{x \sim q(x)} p(x)$ (x rời rạc).

Nhắc lại,

$$\sum_n \log(x_n | \theta) = \sum_n \log \sum_k \frac{p(x_0, z_n | \theta) . q(z_n)}{q(z_n)}$$

$$\geq \sum_n \left[\sum_k q(z_n) . \log \frac{p(x_0, z_n | \theta)}{q(z_n)} \right].$$

$$= \sum_n \left(E_{q(z_n)} . \log \frac{p(x_0, z_n | \theta)}{q(z_n)} \right).$$

$$= \sum_n \left(E_{q(z_n)} . \log P(x_0 | z_n, \theta) - E_{q(z_n)} . \log \frac{q(z_n)}{p(z_n | \theta)} \right)$$

$$= \sum_n E_{q(z_n)} . \log P(x_0 | z_n, \theta) - KL(q(z_n) \| p(z_n, \theta)).$$

KL divergence. (Dùng để đo sự khác biệt của 2 phân phối)

$$KL(q\|p) = E_q \log \frac{q}{p}.$$

+) $KL(q\|q) = 0$ (1)

+) $KL(q\|p) \geq 0$. (2)

+) $KL(q\|p) \neq KL(p\|q)$.

(KL ~~càng gần~~ nhỏ thì phân phối càng gần nhau)

Chứng minh (2): $\sum_n q(x) \cdot \log \frac{q(x)}{p(x)} \geq 0$.

$\Leftrightarrow 0 \geq \int q(x) \cdot \log \frac{p(x)}{q(x)}$.

$VP \leq \log \int q(x) \cdot \frac{p(x)}{q(x)} = \log \int p(x) = \log 1 = 0. \Rightarrow$ đpcm.

+) $KL(p\|q) + KL(q\|r) \geq KL(p\|r)$

Hay: $E_p \cdot \log \frac{p}{q} + E_q \cdot \log \frac{q}{r} \geq E_p \cdot \log \frac{p}{r}$

$\Leftrightarrow E_p \cdot \left(\log \frac{p}{q} + \log \frac{r}{p}\right) + E_q \cdot \log \frac{q}{r} \geq 0.$

$\Leftrightarrow E_p \cdot \log \frac{r}{q} + E_q \cdot \log \frac{q}{r} \geq 0.$

variational inference

likelihood prior

④ $P(z \mid x, \theta) \propto P(x \mid z, \theta) \cdot P(z \mid \theta)$

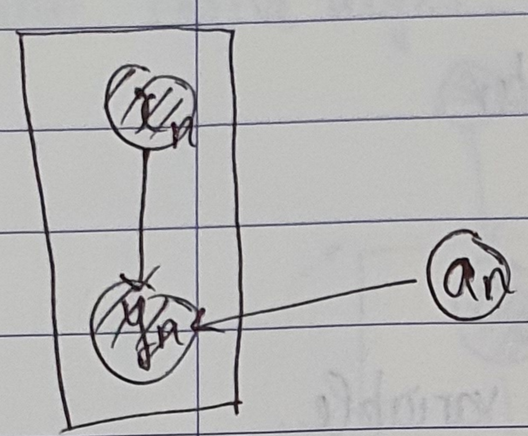
posterior

⊕ Mô hình Linear Regression.

$y_n = \sum_n x_n \cdot a_n$.

$y_n \sim N(\cdot; x^T a, \gamma)$

hằng số.



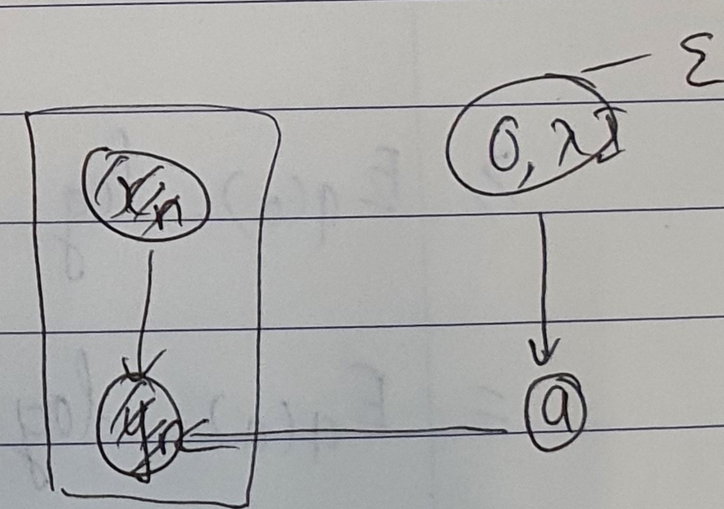
MLE. $\arg\max \sum_n \log P(y_n \mid x_n, a)$

$= \sum_n \log N(y_n, x_n^T a, \gamma)$

$= \arg\min (y_n - x^T a)^2$.

⊕ Mô hình Ridge Regression.

MAP (Maximum a posterior)



$\log P(a \mid \underline{x}, \underline{y}, \underline{\Sigma})$

Data expert knowledge.

$P(a \mid D, \Sigma) \propto P(D \mid a) \cdot P(a \mid \Sigma)$

posterior likelihood prior

(Fully bayssian inference)

- Ước lượng MAP chỉ ~ point estimation

svmc.recruit@samsung.com SamsungVietnam RnD

$a^* = \cancel{\log}\ \text{argmax} \log P(a \mid x, y, \varepsilon) = \text{argmax} \log P(y \mid x, a) . P(a \mid \varepsilon)$

$= \text{argmax} \log P(y \mid x, a) + \log (a \mid \varepsilon)$

$= \sum_n (y - x^T a)^2 + \lambda . \|a\|^2.$

- Frequentist approach.

  $\log (D \mid a)$ a: ~~hog~~ hyper parameter.

- Bayssian approach.

  + MAP (point estimation) với a: random variable.

    $P(a \mid D, \varepsilon)$

  + Variational inference:

    MLE: $P(\text{Data} \mid \varepsilon)$

Có, $\log P(D \mid \varepsilon) = \log \int \frac{P(D, a \mid \varepsilon)}{q(a)} . p(a)$

$\geq E_{q(a)} . \log \frac{P(D, a \mid \varepsilon)}{q(a)}$

$= E_{q(a)} . \log P(D \mid a, \varepsilon) - KL(q(a) \| p(a))$

$= E_{q(a)} . \log \frac{P(D) . P(a \mid D, \varepsilon)}{q(a)}$

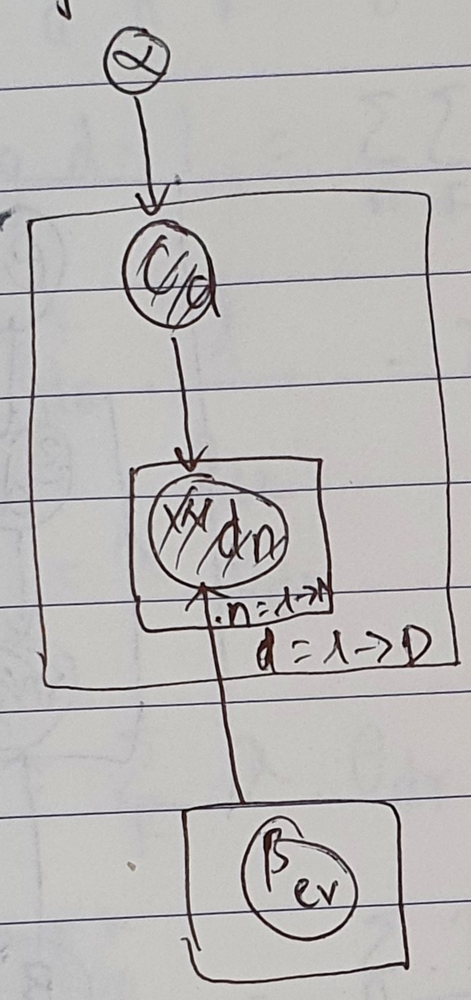
$= E_{q(a)} . \log P(D) - KL(q(a) \| P(a \mid D, \varepsilon))$

$= \log P(D) - KL(q(a) \| P(a \mid D, \varepsilon))$

=> Bản chất của variational inference: là tìm:

$$\arg\min KL(\underbrace{q(a)}_{\text{approximate posterior}} \,||\, \underbrace{p(a|D, \xi)}_{\text{true posterior}})$$

- Mô hình Naive Bayes. (BVN, Gauss).



$C_d | \alpha \sim Mult(\alpha)$

$p(C_d|\alpha) = \alpha_c \quad (\sum_c \alpha_c = 1)$

$w_{dn} | C_d, \beta \sim Multi(\beta_{cv})$.

$$[\beta]_{C \times V} \begin{cases} \sum_v \beta_{cv} = 1 \\ 0 \le \beta_{cv} \le 1 \end{cases}$$

$w_{dn} = \{1, \dots, V\}$.

$P(w_{dn} = j | C_d = c, \beta) = \beta_{cj}$

MLE,

$\log P(C_d, w_{dn} | \alpha, \beta)$

$= \log P(C_d | \alpha) + \log P(w_{dn} | C_d, \beta)$

Probabilistic latent semantic indensity (PLSI) | Topic models
Latent Dirichlet allocation (LDA)

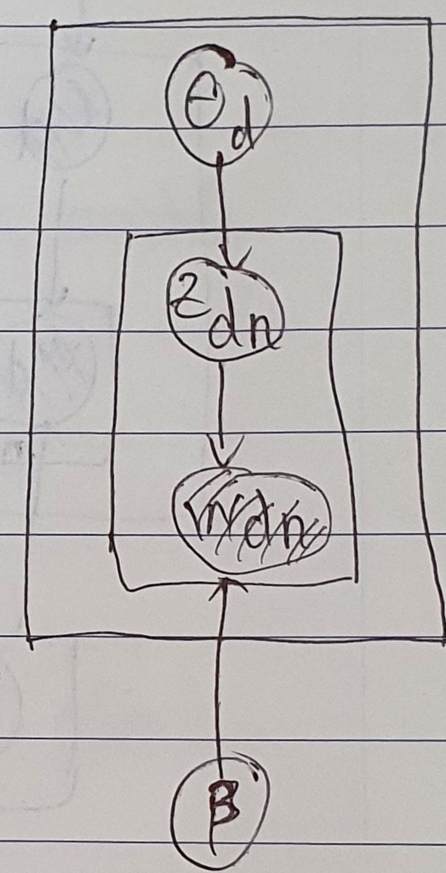
$[D]_{D \times V}$ $[\theta]_{D \times K}$ $[\beta]_{K \times V}$

K, the number of ~~topic~~ hidden topic.

$$\beta_K \begin{cases} \sum_K \beta_{KV} = 1 \\ 0 \le \beta_{KV} \le 1. \end{cases}$$

+) Mô hình PLSI.

z_{dn}: topic index; $z_{dn} = \{1, \dots K\}$.



$P(z_{dn} = k \mid \theta_d) = \theta_{dk}$

$P(w_{dn} = j \mid z_{dn} = k, \beta) = \beta_{kj}$

MLE:

$\log P(w_d \mid \theta_d, \beta)$

$= \cancel{\log} \sum_d \sum_n \log P(w_{dn} \mid \theta_d, \beta) = \sum_d \sum_n \log \sum_k P(w_{dn}, z_{dn} = k \mid \theta_d, \beta)$

$= \sum_d \sum_n \log \sum_k \frac{P(w_{dn}, z_{dn} = k \mid \theta_d, \beta)}{q(z_{dn} = k)} \cdot q(z_{dn} = k)$

$\geq \sum_d \sum_n \sum_k \cancel{\log}\, q(z_{dn} = k) \cdot \log \frac{P(z_{dn} = k \mid \theta_k) \cdot P(w_{dn} \mid z_{dn} = k, \beta)}{q(z_{dn} = k)}$

- θ, β: được khởi tạo ngẫu nhiên.

+) E. step:

$\tau_{dn}^k = q(z_{dn} = k) = P(z_{dn} = k \mid w_{dn}, \theta_d, \beta)$

$\propto P(z_{dn}, w_{dn} \mid \theta_d, \beta)$

$\propto P(z_{dn} = k \mid \theta_d) \cdot P(w_{dn} \mid z_{dn} = k, \beta)$

$\propto \theta_{dk} \cdot \beta_{k, w_{dn}}$

① M- step.

$$ELBO = \sum_d \sum_n \sum_k \gamma_{dn}^k \cdot \log \frac{\theta_{dk} \cdot \beta_{k,w_{dn}}}{\gamma_{dn}^k}.$$

$$= \sum_d \sum_n \sum_k \gamma_{dn}^k (\log \theta_{dk} + \log \beta_{k,w_{dn}})$$

- Tối ưu θ_{dk}: $\sum_d \sum_n \sum_k \gamma_{dn}^k \log \theta_{dn}$; $\sum_k \theta_{dk} = 1.$

tối ưu theo từng d: $L_1 = \sum_n \sum_k \gamma_{dn}^k \cdot \log \theta_{dk} - \lambda \cdot (\sum_k \theta_{dk} - 1).$

$$\frac{\partial L_1}{\partial \theta_{dk}} = \cancel{\sum_n \sum_k} \sum_n \gamma_{dn}^k \cdot \frac{1}{\theta_{dk}} - \lambda = 0$$

$$\Leftrightarrow \sum_n \gamma_{dn}^k = \lambda \cdot \theta_{dk}.$$

$$\Rightarrow \sum_k \sum_n \gamma_{dn}^k = \sum_k \lambda \cdot \theta_{dk} = \lambda. = N.$$

$$\Rightarrow \theta_{dk} = \frac{\sum_n \gamma_{dn}^k}{N} = \frac{\sum_n \gamma_{dn}^k}{\lambda}.$$

- Tối ưu $\beta_{k,w_{dn}}$: $\sum_d \sum_n \sum_k \gamma_{dn}^k \cdot \log \beta_{k,w_{dn}}$; $\sum_{j=1}^{V} \beta_{kj} = 1.$

tối ưu theo từng k: $L_2 = \sum_d \sum_n \gamma_{dn}^k \cdot \log \beta_{k,w_{dn}} - \lambda \cdot (\sum_{j=1}^{V} \beta_{kj} - 1).$

trick: $\beta_{k,w_{dn}} = \prod_{j=1}^{V} (\beta_{kj})^{I(w_{dn}=j)}$. với: $I(w_{dn}=j) = \begin{cases} 1 : \text{if } w_{dn} = j \\ 0 \text{ else.} \end{cases}$

$$\Rightarrow L_2 = \sum_d \sum_n \gamma_{dn}^k \cdot \log \prod_{j=1}^{V} (\beta_{kj})^{I(w_{dn}=j)} - \lambda \cdot (\sum_{j=1}^{K} \beta_{kj} - 1).$$

$$= \sum_d \sum_n \gamma_{dn}^k \cdot \sum_{j=1}^{V} I(w_{dn}=j) \cdot \log \beta_{kj} - \lambda \cdot (\sum_{j=1}^{K} \beta_{kj} - 1)$$

$$\frac{\partial L_2}{\partial \beta_{ki}} = \sum_d \sum_n \gamma_{dn}^k \cdot \cancel{\sum_{j=1}^{V}} \cdot I(w_{dn}=j) \cdot \frac{1}{\beta_{kj}} - \lambda = 0.$$

$$\Leftrightarrow \sum_d \sum_n \cancel{\sum_{j=1}^{V}} \gamma_{dn}^k \cdot I(w_{dn}=j) = \lambda \cdot \beta_{kj}.$$

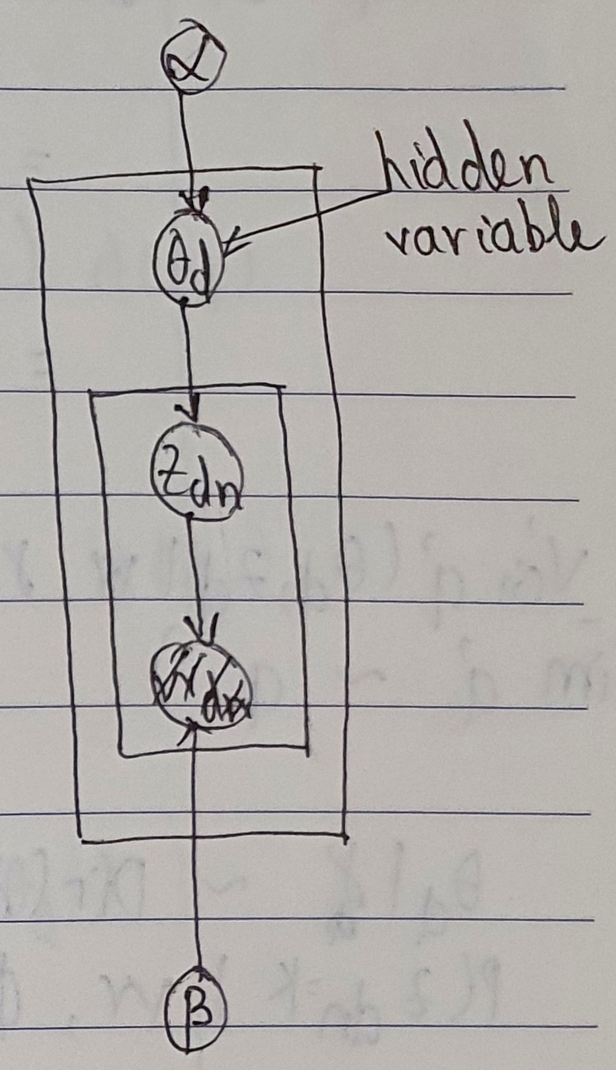
$$\Leftrightarrow \cancel{\sum_k} \cdot \sum_d \sum_n \sum_{j=1}^{V} \gamma_{dn}^k \cdot I(w_{dn}=j) = \sum_{j=1}^{V} \lambda \cdot \beta_{kj} = \lambda.$$

$$\Rightarrow \beta_{kj} = \frac{\sum_d \sum_n \gamma_{dn}^k \cdot I(w_{dn}=j)}{\sum_d \sum_n \sum_{j=1}^{V} \gamma_{dn}^k \cdot I(w_{dn}=j)} \propto \sum_d \sum_n \gamma_{dn}^k \cdot I(w_{dn}=j)$$

svmc.recruit@samsung.com SamsungVietnam RnD

Buổi 3 : (27/08/2020)

Latent Dirichlet Allocation (LDA)



K: the number of hidden topic.

$\theta \sim Dir(\alpha)$: $\alpha = (\alpha_1, \dots \alpha_K)$

$z_{dn} \sim Multi(\theta)$; $P(z_{dn} | \theta) = \theta_k$.

$w_{dn} | z_{dn} \sim Multi(\beta)$; $P(w_{dn} = j | z_{dn} = k, \beta) = \beta_{kj}$

$(\beta = R^{K \times V})$

$$P(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1}$$

$$\text{MLE}: \sum_{d=1}^{D} \log P(w_d | \alpha, \beta)$$

$$= \sum_{d=1}^{D} \log \int_{\theta_d} \sum_{z_d} \frac{P(w_d, z_d, \theta_d | \alpha, \beta)}{q(\theta_d, z_d)} \cdot q(\theta_d, z_d)$$

$$\geq \sum_{d} \int_{\theta_d} \sum_{z_d} q(\theta_d, z_d) \cdot \log \frac{P(w_d, z_d, \theta_d | \alpha, \beta)}{q(\theta_d, z_d)}$$

$$(q(\theta_d, z_d) = P(\theta_d, z_d | w_d, \alpha, \beta))$$

$$= \sum_{d} E_{q(\theta_d, z_d)} \cdot \log \frac{P(w_d, z_d, \theta_d | \alpha, \beta)}{q(\theta_d, z_d)}$$

Việc tìm $q(\theta_d, z_d)$ khó => mean field variational inference.

- mean field variational inference.

$$q'(\theta_d, z_d) = q'(\theta_d) \cdot \prod_{n=1}^{N} q'(z_{dn})$$

$$= q'(\theta_d \mid \gamma_d, \gamma) \prod_{n=1}^{N} q'(z_{dn} \mid w_d, \phi)$$

$$= Dir(\gamma_d) \cdot \prod_{n=1}^{N} Multi\ q(z_{dn} \mid w_d, \phi).$$

Với $q'(\theta_d, z_d \mid w, \gamma, \phi) = \arg\min_{q} KL(q \,||\, p(\theta_d, z_d \mid w_d, \alpha, \beta))$

(Tìm $q' \sim q$)

approximate posterior true posterior

$\theta_d \mid \gamma_d \sim Dir(\gamma)$; $\gamma \in \mathbb{R}^K$

$P(z_{dn} = k \mid w, \phi_{dn}) = \phi_{dnk}$; $\sum_k \phi_{dnk} = 1$

$$ELBO = \sum_d \left[E_{q(\theta_d, z_d)} \cdot \log p(w_d, \theta_d, z_d \mid \alpha, \beta) - E_{q(\theta_d, z_d)} \cdot \log q(\theta_d, z_d) \right]$$

$$= \sum_d \left[E_{q(\theta_d, z_d)} \cdot \log(\theta_d \mid \alpha) + \sum_n \left(E_{q(\theta_d, z_d)} \cdot \log(z_{dn} \mid \theta_d) + E_{q(\theta_d, z_d)} \log(w_{dn} \mid z_{dn}, \beta) \right) \right.$$
$$\left. - E_{q(\theta_d, z_d)} \cdot \log q(\theta_d, z_d) \right]$$

$$= \sum_d \left[\overset{(1)}{E_{q(\theta_d)} \cdot \log(\theta_d \mid \alpha)} + \sum_n \overset{(2)}{E_{q(\theta_d, z_d)} \cdot \log(z_{dn} \mid \theta_d)} + \sum_n \overset{(3)}{E_{q(z_d)} \cdot \log(w_{dn} \mid z_{dn}, \beta)} \right.$$
$$\left. - \overset{(4)}{E_{q(\theta_d)} \cdot \log q(\theta_d)} - \sum_n \overset{(5)}{E_{q(z_{dn})} \cdot \log q(z_{dn})} \right]$$

Giải thích biến đổi 2:

$$E_{q(\theta_d, z_d)} \cdot \log q(\theta_d|\alpha) = \int_{\theta_d} \sum_{z_d} q(\theta_d, z_d) \cdot \log q(\theta_d|\alpha) \, d(\theta_d)$$

$$= \int_{\theta_d} \sum_{z_{d_1}} \cdots \sum_{z_{d_n}} q(\theta_d) \cdot \prod_n q(z_{dn}) \cdot \log q(\theta_d|\alpha) \, d(\theta_d)$$

$$= \underbrace{\sum_{z_{d_1}} \cdots \sum_{z_{dn}} \prod_n q(z_{dn})}_{=1} \cdot \int_{\theta_d} q(\theta_d) \cdot \log q(\theta_d|\alpha) \, d(\theta_d)$$

$$= E_{q(\theta_d)} \cdot \log(\theta_d|\alpha)$$ (Các b.thức ≠ tương tự)

+) xét (1): $E_{q(\theta_d)} \cdot \log(\theta_d|\alpha) = E_{q(\theta_d|\gamma)} \cdot \log(\theta_d|\alpha)$

~~=(4)+(3)~~ Có: $\log(\theta_d|\alpha) = \log \dfrac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \cdot \prod_k \theta_{dk}^{\alpha_k - 1}$

$$= \sum_k (\alpha_k - 1) \cdot \log \theta_{dk} + \log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k)$$

Có: $E_{q(\theta_d)}[\log \theta_{dk}] = \psi(\gamma_{dk}) - \psi(\sum_k \gamma_{dk})$.

$$\Rightarrow E_{q(\theta_d)} \cdot \log(\theta_d|\alpha) = \sum_k (\alpha_k - 1) \cdot \left(\psi(\gamma_k) - \psi(\sum_k \gamma_k)\right) + \log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k)$$

SamsungVietnam RnD
...ung.recruit@samsung.com

+) Xét 2: $\sum_n E_{q(\theta_d, z_{dn})} \cdot \log(z_{dn} | \theta_d)$

Có, $\log(z_{dn} | \theta_d) = \log \prod_k \theta_{dk}^{I(z_{dn}=k)}$

$= \sum_k I(z_{dn}=k) \cdot \log \theta_{dk}$.

$\Rightarrow E_{q(\theta_d, z_{dn})} \cdot \log(z_{dn} | \theta_d) = E_{q(\theta_d, z_{dn})} \cdot \sum_k I(z_{dn}=k) \cdot \log \theta_{dk}$

$= \sum_k E_{q(\theta_d)} \cdot \log \theta_{dk} \, E_{q(z_{dn})} \cdot I(z_{dn}=k)$.

$= \sum_k \left(\psi(\gamma_k) - \psi\left(\sum_k \gamma_k\right) \right) \cdot \phi_{dnk}$.

+) Xét 3: $E_{q(z_{dn})} \cdot \log(w_{dn} | z_{dn}, \beta)$

Có, $\log(w_{dn} | z_{dn}, \beta) = \log \prod_k \prod_{j=1}^{V} (\beta_{kj})^{I(w_{dn}=j, z_{dn}=k)}$

$= \sum_k \sum_{j=1}^{V} (\log \beta_{kj}) \cdot I(w_{dn}=j, z_{dn}=k)$

$\Rightarrow E_{q(z_{dn})} \cdot \log(w_{dn} | z_{dn}, \beta) = \sum_k \sum_{j=1}^{V} E_q I(w_{dn}=j, z_{dn}=k) \log \beta_{kj}$

$= \sum_k \sum_{j=1}^{V} \phi_{dnk} \cdot I(w_{dn}=j) \cdot \log \beta_{kj}$

+) Xét 4: $E_{q(\theta_d)} \cdot \log q(\theta_d)$.

Có, $q(\theta_d) = \dfrac{\Gamma(\sum \gamma_k)}{\prod_k \Gamma(\gamma_k)} \cdot \prod_k \theta_{dk}^{\gamma_k - 1}$

$\Rightarrow E_q \cdot \log q(\theta_d) = \sum_k (\gamma_k - 1) \cdot (\Psi(\gamma_k) - \Psi(\sum_k \gamma_k))$

$+ \log \Gamma(\sum_k \gamma_k) - \sum_k \log \Gamma(\gamma_k)$

+) Xét 5: $\sum_n E_{q(z_{dn})} \cdot \log q(z_{dn})$

Có, $q(z_{dn}) = \prod_k \phi_{dnk}^{I(z_{dn}=k)}$

$\Rightarrow E_{q(z_{dn})} \cdot \log q(z_{dn}) = \sum_k E_{q(z_{dn})} \cdot I(z_{dn} = k) \cdot \log \phi_{dnk}$

$= \sum_k \phi_{dnk} \cdot \log \phi_{dnk}$

$\Rightarrow$ Hàm mục tiêu:

$$\sum_d \Big[\sum_k (\alpha_k - 1)(\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)) + \log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k)$$

$$+ \sum_n \sum_k \sum_v \underbrace{I(w_{dn} = j)}_{\phi_{dnk}} \cdot \log \beta_{kj} + \sum_n \sum_k \phi_{dnk} \cdot (\Psi(\gamma_k) - \Psi(\sum_k \gamma_k))$$

$$- \sum_k (\gamma_k - 1)(\Psi(\gamma_k) - \Psi(\sum_k \gamma_k)) - \log \Gamma(\sum_k \gamma_k) + \sum_k \log \Gamma(\gamma_k)$$

$$- \sum_n \sum_k \phi_{dnk} \cdot \log \phi_{dnk} \Big]$$

Local variable θ_d, z_{dn}	$q(\theta_d \mid \gamma_d)$ $q(z_{dn} \mid \phi_{dn})$	$\Rightarrow$ inference
Global parameter	α, β	$\Rightarrow$ learn

SamsungVietnam RnD

④ Mô giả: init α, β.

while (convergence): init: $\phi_{nk} = 1/K$; $\gamma_{dk} = \alpha_k + N/K$.

for each doc:

while (convergence):

for each word:

for each topic:

update ϕ^{t+1}_{dnk}; end for

normalize ϕ^{t+1}_{dnk} to sum to 1

$\gamma^{t+1}_{dk} = \alpha_k + \sum_n \phi^{t+1}_{dn}$ # update γ_{dk}.

learning

training

Sau khi learning => có $[\beta]_{K \times V}$

$$E(\phi_d) = \frac{\gamma_{dk}}{\sum_k \gamma_{dk}}.$$

④ Tối ưu: E-step.

- Tối ưu ϕ:

$$\sum_d \Big(\sum_n \sum_k \sum_v \phi_{dnk} . I(w_{dn} = j) . \log \beta_{kj}) + \sum_n \sum_k \phi_{dnk} . \big(\psi(\gamma_{dk}) + \psi(\sum_k \gamma_{dk}) \big) \Big)$$

$$- \sum_n \sum_k \phi_{dnk} . \log \phi_{dnk} \Big) \quad \text{với} \quad \sum_k \phi_{dnk} = 1.$$

- Tối ưu theo từng d, n:

$$L(\phi_{dnk}) = \sum_k \sum_v \phi_{dnk} . I(w_{dn} = j) . \log \beta_{kj} + \sum_k \phi_{dnk} . \big(\psi(\gamma_{dk}) + \psi(\sum_k \gamma_{dk}) \big)$$

$$- \sum_k \phi_{dnk} . \log \phi_{dnk} - \lambda . \big(\sum_k \phi_{dnk} - 1 \big)$$

$$\frac{\partial L(\phi_{dnk})}{\partial \phi_{dnk}} = \sum_v I(w_{dn} = j) . \log \beta_{kj} + \psi(\gamma_{dk}) + \psi(\sum_k \gamma_{dk}) - (\log \phi_{dnk} + 1) - \lambda = 0$$

$$\Leftrightarrow \log \phi_{dnk} = \sum_v I(w_{dn} = j) . \log \beta_{kj} + \psi(\gamma_{dk}) + \psi(\sum_k \gamma_{dk}) - 1 - \lambda$$

$$\Leftrightarrow \phi_{dnk} = \frac{\exp\big(\sum_v I(w_{dn} = j) . \log \beta_{kj} + \psi(\gamma_{dk}) + \psi(\sum_k \gamma_{dk}) \big)}{\exp(1 + \lambda)}$$

$$\Rightarrow \boxed{\phi_{dnk} \propto \beta_{kj} . \exp\big(\psi(\gamma_{dk}) + \psi(\sum_k \gamma_{dk}) \big)}$$

- Tối ưu γ_{dk}: Tối ưu theo từng d.

$$L(\gamma_{dk}) = \sum_k (\alpha_k - 1) \big(\psi(\gamma_{dk}) - \psi(\sum_k \gamma_{dk}) \big) + \sum_n \phi_{dnk} . \big(\psi(\gamma_{dk}) - \psi(\sum_k \gamma_{dk}) \big)$$

$$- \log \Gamma(\sum_k \gamma_{dk}) + \log \Gamma(\gamma_{dk}) - \sum_k (\gamma_{dk} - 1) . \big(\psi(\gamma_{dk}) - \psi(\sum_k \gamma_{dk}) \big)$$

SamsungVietnam RnD
svmc.recruit@samsung.com

$$= \sum_k \left(\psi(\gamma_{dk}) - \psi\left(\sum_k \gamma_{dk}\right)\right)\left(\alpha_k + \sum_n \phi_{dnk} - \gamma_{dk}\right) - \log \Gamma\left(\sum_k \gamma_{dk}\right) + \log \Gamma(\gamma_{dk})$$

$$\frac{\partial L}{\partial \gamma_{dk}} = \psi'(\gamma_{dk}).\left(\alpha_k + \sum_n \phi_{dnk} - \gamma_{dk}\right) - \psi'\left(\sum_k \gamma_{dk}\right).\sum_k \left(\alpha_k + \sum_n \phi_{dnk} - \gamma_{dk}\right) = 0$$

$$\Leftrightarrow \boxed{\gamma_{dk} = \alpha_k + \sum_n \phi_{dnk}}$$

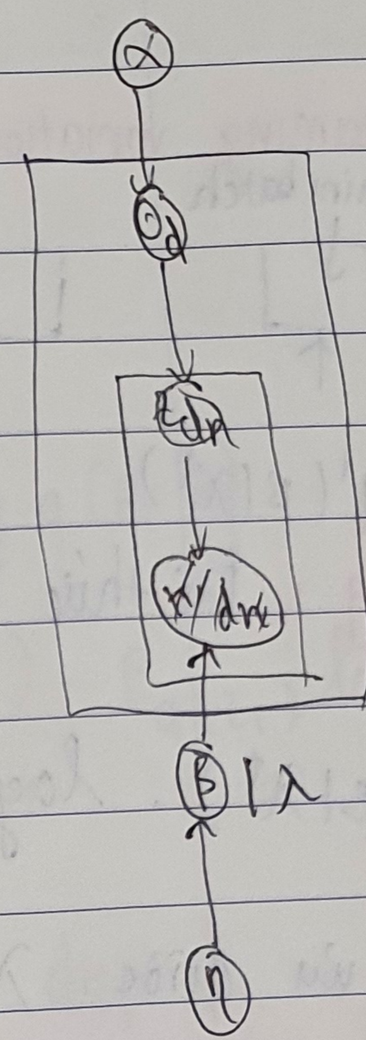
M-step.

$$\beta_{kj} \propto \sum_d \sum_n \phi_{dnk} . I(w_{dn} = j)$$

Buổi 4 (03/09/2020)

$\log(w \mid \alpha, \eta) = \dots$

$q(\theta, z, \beta) = \prod_k^K \underbrace{q(\beta_k \mid \lambda_k)}_{Dir(\lambda_k)} \prod_d (\dots)$



tối ưu được:

local
$$\begin{cases} \gamma_{dk} = \alpha_k + \sum_k \phi_{dnk} \\ \phi_{dnk} \propto Exp(E(\log \phi_{dk})) . Exp(E(\log \beta_k w_{dn})) \end{cases}$$

$$\lambda_{kj} = \eta_{kj} + \sum_k \sum_n \phi_{dnk} . I(w_{dn} = j)$$

Online LDA.

Stochastic ~~variable~~ variational inference (SVI)

- Init λ.

for each mini batch
  for each doc:
    repeat:
      ϕ
      γ
    until
  end for
end mini batch

Local variable

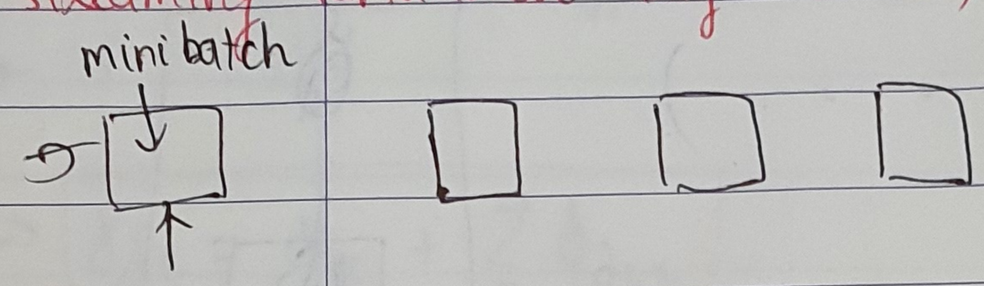
$$\tilde{\lambda}_{kj} = \eta_k + \frac{D_1}{d_{minibat}} \sum_{d}^{minibatch} \sum_n \phi_{dk} . I(w_k = j)$$

→ Cần phải biết toàn bộ dữ liệu

(D_1 là toàn bộ data)

$\lambda^{T} = (1-\rho_t)\cdot\lambda^{t-1} + \rho_t \cdot \tilde{\lambda}^{t}$ với $\rho_t = (\tau + t)^{-\kappa}$.

③ Streaming variational Bayes (SVB)



$q^{1}(\beta|\lambda^{1})$

Ý tưởng: tri thức học được ở batch trước => làm prior ở batch sau.

$E_{q(\beta|\lambda^{t})} \log P(D|\beta) - KL(q(\beta|\lambda^{t}) \| p(\beta|\lambda^{t-1}))$.

=> Tối ưu được, $\lambda^{t}_{kj} = \lambda^{t-1}_{kj} + \sum_{\text{minibatch } n} \sum \phi_{dnk} I(w_{dn} = j)$

explot prior knowledge ~~variational~~ inference in streaming

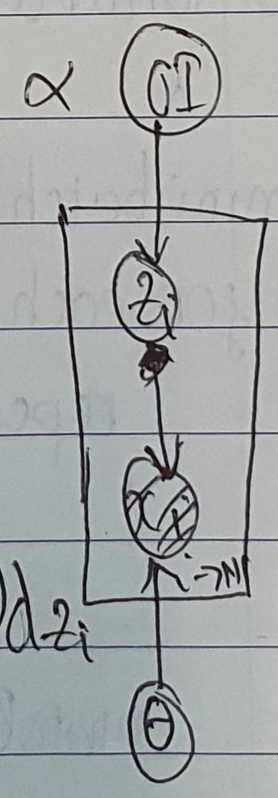
$\boxed{\text{Proposal}}$

④ Variational autoencoder (VAE)

$z_i \sim N(0, I)$

$x_i \sim Ber(f(z_i, \theta))$; $x_{ij} \sim ber(p_{ij})$

$p_i = f_i(z_i, \theta)$



⊕ Autoencoder variational inference (AEVB)

$$\sum_i \log P(x_i|\alpha,\theta) = \log \int_{\theta_i} \int_{z_i} P(x_i, z_i, \theta_i | \alpha)\, dz_i$$

$$= \sum_i \log \int_{\theta_i} \int_{z_i} \frac{P(x_i, z_i, \theta_i | \theta, \alpha)}{q(z_i, \theta_i)} \cdot q(z_i, \theta_i)$$

$$\geq \sum_i E_{q(z_i, \cancel{\theta_i})} \log \frac{p(x_i, z_i, \cancel{\theta_i} \mid \alpha, \theta)}{q(z_i, \cancel{\theta_i})}$$

$$\cancel{q(z_i, \theta_i) = q(z_i, \theta_i \mid \alpha, x_i)}$$

$$q(z_i) = q(z_i, \cancel{\theta} \mid \alpha, \theta_i)$$

$$= \sum_i \left[E_{q(z_i)} \log p(x_i, z_i \mid \alpha, \theta) - E_{q(z_i)} \log q(z_i) \right]$$

$$= \sum_i \left[E_{q(z_i)} \left(\log p(z_i \mid \alpha) + \log p(x_i \mid z_i, \theta) \right) - E_{q(z_i)} \log q(z_i) \right]$$

$$\sum_i \log p(x_i \mid \alpha, \theta) = \sum_i \log \int_{z_i} p(x_i, z_i \mid \alpha, \theta) \, dz_i$$

$$\geq \sum_i E_{q(z_i)} \log \frac{p(x_i, z_i \mid \alpha, \theta)}{q(z_i)} = \sum_i \left(\underbrace{E_{q(z_i)} \log p(x_i \mid z_i, \theta)}_{\text{Reconstructer loss}} - \underbrace{KL(q(z_i) \| p(z_i \mid \alpha))}_{\text{Regularization.}} \right)$$

$$q(z_i \mid x_i) \approx p(z_i \mid x_i, \alpha, \theta)$$

$$p(z_i \mid x_i) = N(z_i, \underbrace{\mu_\phi(x_i)}, \underbrace{\Sigma_\phi(x_i)})$$

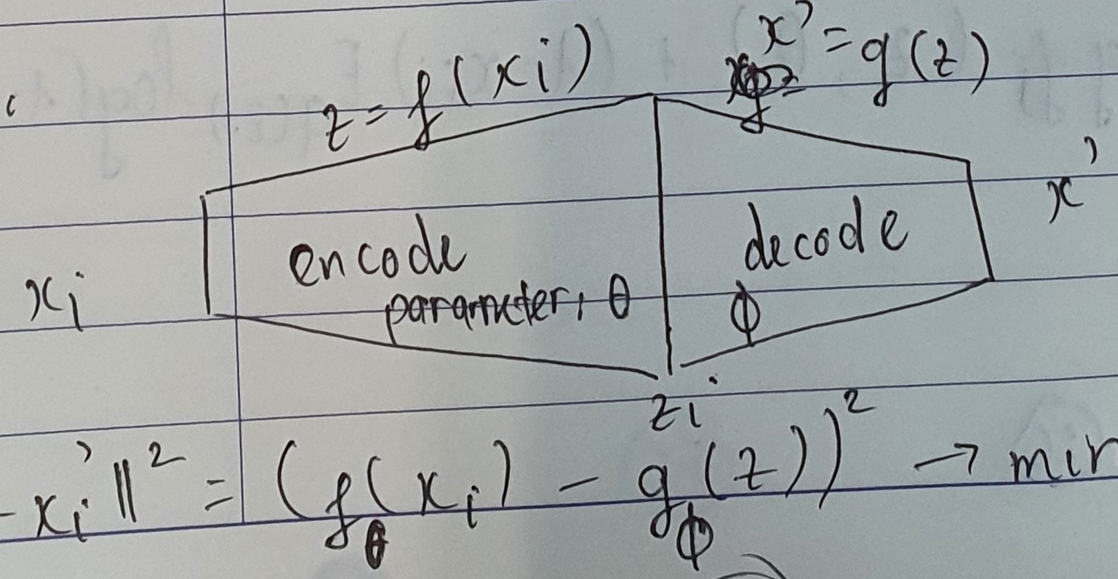
Hàm số => tìm tham số ϕ sẽ tìm đc μ, Σ

$\mu, \Sigma = x_i$ → [ ] → μ, Σ

weight, ϕ

Mạng

AE.



$$\|x_i - x_i'\|^2 = (f_\theta(x_i) - g_\phi(z))^2 \to \min$$

decoder: $x_i' \sim ber(g(z_i, \theta))$

encoder: $z_i \sim N(\mu_\phi(x_i, \phi), \Sigma_\phi(x_i, \phi))$

Ưu điểm VAE so với LDA,

Khi inference 1 doc mới: VAE, chỉ cần cho x_i vào "mạng" => đẩy ra μ, Σ.
LDA: phải tối ưu μ, Σ.

Tối ưu:

+) Xét $KL(q(z_i | \mu_\phi(x_i), \Sigma_\phi(x_i)) | N(z_i | 0, I))$

$= F(\mu_\phi(x_i), \Sigma_\phi(x_i), 0, I)$ -> Có công thức tường minh.

+) Xét $\sum_i E_{q(z|\mu_\phi(x_i), \Sigma_\phi(x_i))} \cdot P(x_i | f(z_i, \phi))$

$$= \sum_j^{\alpha} E_{q(z|\dots)} \log f_j(z_i, \theta)^{x_{ij}} \cdot (1 - f_j(z_i, \theta))^{(1-x_{ij})}$$

(α là số chiều mà $f(z_i | \phi)$ output.)

$$= \sum_j^{\alpha} E_{q(z|\dots)} \left[x_{ij} \cdot \log f_j(z_i, \theta) + (1 - x_{ij}) \cdot \log(1 - f_j(z_i, \theta)) \right]$$

$$= \sum_j^{\alpha} x_{ij} \cdot E_{q(z|\dots)} \cdot \log f_j(z_i, \theta) + (1 - x_{ij}) \cdot E_{q(z\dots)} \log(1 - f_j(z_i, \theta))$$

$$\theta^t = \theta^{t-1} + \eta \frac{\partial L}{\partial \theta}(\theta^{t-1}, \phi^{t-1})$$

$$\phi^t = \phi^{t-1} + \eta \frac{\partial L}{\partial \phi}(\theta^{t-1}, \phi^{t-1})$$

reparameterization trick | ~~Reinforce~~

$z \sim N(\ , \mu, \Sigma)$

$\varepsilon \sim N(0, \Sigma)$

→ reparameter $z = \mu + \Sigma . \varepsilon$

ε là ma trận đg chéo

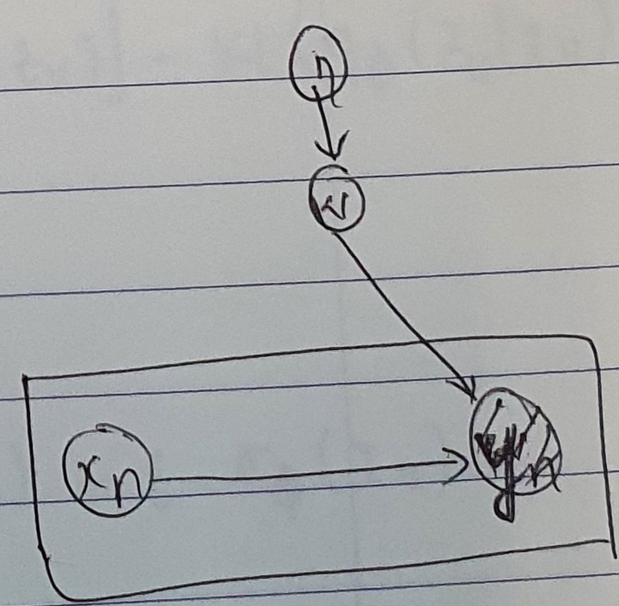
Reinforce.

$$E_{q(z \mid \mu_\phi(x_i), \Sigma_\phi(x_i))} . \log f(z_i, \theta)$$

$$\simeq \frac{1}{L} . \sum_{l=1}^{L} \log f(\mu_\phi(x_i) + \Sigma_\phi(x_i) . \varepsilon_l ; \theta)$$

BNN.

$$\log \int_w \frac{p(x, y, w \mid \eta)}{q(w)} . q(w)\, dw$$



$$\geq E_{q(w)} . \log P(x, y, w, \mid \eta) - KL(q(w) \mid p(w))$$

$q(w) = N(\mu, \Sigma)$

svmc.recruit@samsung.com
SamsungVietnam RnD