Variational Inference

z: latent variables.

x :- data

$P(z,x)$:- model

Goal :- approximate $P(z|x)$ by $q(z) = \prod_{i=1}^{M} q(z_i)$

$q(z) \leftarrow \underset{q(z)}{\operatorname{argmin}} KL(q \| p(z|x))$ [Variational inference]

or $\underset{q(z)}{\operatorname{argmin}} KL(p(z|x) \| q)$ [Expectation-propagation]

$$q(z) \longleftarrow \underset{q(z)}{\operatorname{argmin}} \; KL(q \,\|\, p(z|x))$$

$$= \underset{q(z)}{\operatorname{argmin}} \; - E_{q(z)} \log \frac{P(z|x)}{q(z)}$$ ← not known

$$= \underset{q(z)}{\operatorname{argmin}} \; - E_{q(z)} \log \frac{P(z,x)/P(x)}{q(z)}$$

$$= \underset{q(z)}{\operatorname{argmin}} \; - E_{q(z)} \log \frac{P(z,x)}{q(z)}$$ ← defined by model

Hence $$q(z) \longleftarrow \underset{q(z)}{\operatorname{argmax}} \; \underbrace{\mathbb{E}_{q(z)} \log P(z,x)}_{\text{Cross-entropy:- fit of } q(z) \text{ to } P(z,x)} \; - \underbrace{\overbrace{\mathbb{E}_{q(z)} \log q(z)}^{\text{width of } q(z)}}_{\text{entropy}}$$

"Find $q(z)$ which best fits $P(z|x)$ and is as wide as possible" $\Longrightarrow$ avoids regions where $P(z|x)=0$

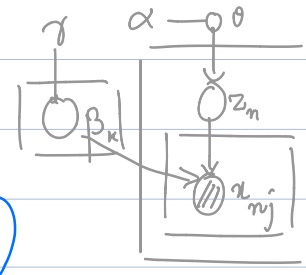
i.e $q(z)$ is also zero.

For $q(z) = \prod_{i=1}^{M} q(z_i)$

Solution:-

$$\ln q(z_i) = \mathbb{E}_{\prod_{j\neq i} q(z_j)} \ln P(z,x) + \text{const}$$

Variational inference for Multinomial Mixture Model



$$\ln P(Z, X \mid \text{hyper-parameters})$$

$$= \ln P(\beta_1, \dots, \beta_K, \theta, z_1, x_1, \dots, z_N, x_n \mid \alpha, \gamma)$$

using conditional independences

$$= \sum_{k=1}^{K} \ln P(\beta_k \mid \gamma) + \ln P(\theta \mid \alpha) + \underbrace{\sum_{n=1}^{N} \ln P(z_n, x_n \mid \beta_1, \dots, \beta_K, \theta)}_{\longrightarrow A}$$

$$\ln q(z_n) = \mathbb{E}_{q_{-z_n}} \left[A \right]$$

$q_{-z_n} \longrightarrow$ all q's except $q(z_n)$

$$= \sum_{k=1}^{K} \mathbb{E}_{q(\beta_k)} \ln P(\beta_k \mid \gamma) + \mathbb{E}_{q(\theta)} \ln P(\theta \mid \alpha)$$

$$+ \sum_{l \neq n} \mathbb{E}_{q(z_l) \prod_k q(\beta_k)\, q(\theta)} \left[\ln P(z_l, x_l \mid \beta_1, \dots, \beta_K, \theta) \right]$$

no z_n

$$+ \mathbb{E}_{\prod_{k=1}^{K} q(\beta_k)\, q(\theta)} \ln P(z_n \mid \theta) + \mathbb{E}_{\prod q(\beta_k)\, q(\theta)} \ln P(x_n \mid z_n, \beta_1, \dots, \beta_K, \theta)$$

$$= \text{Constant} + \mathbb{E}_{q(\theta)} \ln \prod_{k=1}^{K} (\theta_k)^{z_{n,k}} +$$

$$\mathbb{E}_{\prod_k q(\beta_k)} \ln \prod_{w=1}^{W} \left(\prod_{k=1}^{K} \beta_{k,w}^{z_{n,k}} \right)^{m_{n,w}}$$

$m_{n,w} :\ \sum_{j=1}^{J_N} x_{n,j,w}$

$\#\{$ wth word in n'th document $\}$

$$= \sum_{k=1}^{K} z_{n,k} \mathbb{E}_{q(\theta)} \ln \theta_k + \sum_{k=1}^{K} z_{n,k} \sum_{w=1}^{W} m_{n,w} \mathbb{E}_{q(\beta_k)} \ln \beta_{k,w}$$

$$\ln q(z_n) = \sum_{k=1}^{K} z_{n,k} \underbrace{\left(\mathbb{E}_{q(\theta)} \ln \theta_k + \sum_{w=1}^{W} m_{n,w} \mathbb{E}_{q(\beta_k)} \ln \beta_{k,w} \right)}_{\ln v z_{n,k}}$$

$$+ \text{const}$$

compare

Compare to :- $\ln \text{Multinomial}(t \mid \pi)$

$$\propto \sum_{k=1}^{K} t_k \ln \pi_k$$

Hence

$$q(z_n) = \text{Multinomial}(z_n \mid v z_n)$$

Where :-

$$\nu z_{n,k} \propto \exp\left\{ \underbrace{E_{q(\theta)} \ln \theta_k}_{?} + \sum_{w=1}^{W} m_{n,w} \underbrace{E_{q(\beta_k)} \ln \beta_{k,w}}_{?} \right\}$$

What is $E_{q(\theta)} \ln \theta_k$, $E_{q(\beta_k)} \ln \beta_{k,w}$?

$$\ln q(\theta) = E_{q_{-\theta}} [(A)]$$

$$= \sum_{k=1}^{K} E_{q(\beta_k)} \ln P(\beta_k | \gamma) + \ln P(\theta | \alpha) + \sum_{n=1}^{N} E_{q_{-\theta}} \ln P(z_n, x_n | \theta, ..)$$

$$\Downarrow \quad \prod_k \prod_w (\beta_{kw})^{m_{nw} z_{nk}}$$

$$= \underbrace{\sum_{k=1}^{K} E_{q(\beta_k)} \ln P(\beta_k | \gamma)}_{} + \ln P(\theta | \alpha) +$$

$$\sum_{n=1}^{N} E_{q(z_n)} \ln P(z_n | \theta) + \underbrace{\sum_{n=1}^{N} \sum_{k=1}^{K} E_{q(z_n)} z_{n,k} \left(\sum_{w=1}^{W} m_{n,w} E_{q(\beta_k)} \ln \beta_{kw} \right)}_{}$$

Ignoring terms independent of θ

$$= \ln P(\theta | \alpha) + \sum_{n=1}^{N} \sum_{k=1}^{K} E_{q(z_n)} z_{n,k} \ln \theta_k$$

$$= \underbrace{\ln \left(\frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \right) + \sum_{k=1}^{K} (\alpha - 1) \ln \theta_k}_{\ln Dir(\theta \mid \alpha)} +$$

$$\sum_{k=1}^{K} \ln \theta_k \left(\sum_{n=1}^{N} \underbrace{E_{q(z_n)} z_{n,k}}_{\nu z_{n,k} \text{ from above}} \right)$$

Conjugacy in action

$$\ln q(\theta) = \sum_{k=1}^{K} \underbrace{\left(\sum_{n=1}^{N} \nu z_{n,k} + \alpha_k - 1 \right)}_{\nu\theta_k} \ln \theta_k + \text{const}$$

$$q(\theta) = Dir(\theta \mid \nu\theta)$$

Similarly :-

$$q(\beta_k) = Dir(\beta_k \mid \nu\beta_k)$$

where $$\nu\beta_{k,w} = \sum_{n=1}^{N} \nu z_{n,k} \, m_{n,w} + \gamma - 1$$

In $\nu z_{n,k}$:- $$E_{q(\theta)} \ln \theta_k = \psi(\nu\theta_k) - \psi\left(\sum_k \nu\theta_k\right)$$ {proof below}

$$E_{q(\beta_k)} \ln \beta_{k,w} = \psi(\nu\beta_{k,w}) - \psi\left(\sum_w \nu\beta_{k,w}\right)$$

Summary :-

* 1) Knowing functional form of q is not needed

Imp

* 2) Variational distribution takes the form of prior for conjugate models

Given 2), we have an alternative way to derive same results

* Write ELBO :- $E_q \ln P(x, z) - E_q \ln q$

Choose :-

$$q(z_n) = \text{Multinomial}(z_n \mid \nu z_n)$$
$$q(\theta) = \text{Dir}(\theta \mid \nu\theta)$$
$$q(\beta_k) = \text{Dir}(\beta_k \mid \nu\beta_k)$$

$$\nu z_n, \nu\theta, \nu\beta_k \leftarrow \underset{\nu z_n, \nu\theta, \nu\beta_k}{\text{argmax}} \; E_q \ln P(x, z) - E_q \ln q$$

Solve :-

$$\nabla_{\nu z_n} \text{ELBO} = 0, \quad \nabla_{\nu\theta} \text{ELBO} = 0, \quad \nabla_{\nu\beta_k} \text{ELBO} = 0$$

$\forall\; n = 1, \dots, N$, $k = 1, \dots, K$, Same answers as above!

PROOF :-

$$\mathbb{E}_{q(\theta)} \ln \theta_k = \psi(\nu\theta_k) - \psi\left(\sum_{k=1}^{k} \nu\theta_k\right)$$

$$q(\theta) = \mathrm{Dir}(\theta \mid \nu\theta)$$

for an exponential family :-

$$\int g(\eta)\, h(x) \exp(\eta^T u(x))\, dx = 1$$

$$\int h(x) \exp(\eta^T u(x))\, dx = \frac{1}{g(\eta)}$$

$$\ln \int h(x) \exp(\eta^T u(x))\, dx = -\ln g(\eta)$$

$$= \nabla_\eta \ln \int h(x) \exp(\eta^T u(x))\, dx = -\nabla_\eta \ln g(\eta)$$

$$\frac{1}{\int h(x) \exp(\eta^T u(x))\, dx} \int h(x) \exp(\eta^T u(x))\, u(x)\, dx = -\nabla_\eta \ln g(\eta)$$

$$g(\eta) \int h(x) \exp(\eta^T u(x))\, u(x)\, dx = -\nabla_\eta \ln g(\eta)$$

$$\mathbb{E}_{p(x|\eta)}\, u(x) = -\nabla_\eta \ln g(\eta)$$

$$-\nabla_{\eta} \ln g(\eta) = \mathbb{E}_{p(x|\eta)} u(x).$$

true for any exponential family distribution

for $\mathrm{Dir}(\theta \mid v\theta)$, $u(\theta) = [\ln \theta_1, \dots, \ln \theta_K]$

$$\ln g(v\theta) = \ln \Gamma\left(\sum_k v\theta_k \right) - \sum_k \ln \Gamma(v\theta_k)$$

and $\mathbb{E}_{\mathrm{Dir}(\theta|v\theta)} \ln \theta_k = \psi(v\theta_k) - \psi\left(\sum_k v\theta_k\right)$

ψ: derivative of $\ln \Gamma$

called as digamma function.