Correlated Supervised LDA

Veshnu Ramakrishnan

Apr 30

1 Model definition

The model is given by Figure . I use a combination of correlated topic model and supervised topic model to achieve my goals. The joint distribution is given by equation 1. Taking log on both sides, integrating and using jensen’s inequality we get equation 2. The objective function for variational EM is given by equation 3. The variational posterior of the model is given by 4.

\[ p(w_{1:N}, y, \mu, \Sigma, \beta, \alpha, \sigma^2) = p(w_{1:N}|\beta_{1:K}, z_{1:N}) * p(z_{1:N}|\eta_d) * p(\eta_d|\mu, \Sigma) * p(y|\sigma^2, \alpha, z_{1:N}) \]

\[ \log p(w_{1:N}, y|\mu, \Sigma, \beta, \alpha, \sigma^2) = \sum_{n=1}^{N} E_q[\log p(w_{1:N}|\beta_{1:K}, z_{1:N})] + \]

\[ \sum_{n=1}^{N} E_q[\log p(z_n|\eta_d)] + \]

\[ E_q[\log p(\eta_d|\mu, \Sigma)] + \]

\[ E_q[\log p(y|\sigma^2, \alpha, z_{1:N})] + \]

\[ H(q) \]  

\[ L(\mu, \Sigma, \beta_{1:K}, \alpha, \sigma^2; w_{1:D}, y) \geq \sum_{d=1}^{D} E_{q_d}[\log p(\eta_d, z_d, w_d, \eta_d|\mu, \Sigma, \beta_{1:K}, \alpha, \sigma^2)] + H(q_d) \]

\[ q(\eta_{1:K}, z_{1:N}|\lambda_{1:K}, \nu_{1:K}, \phi_{1:N}) = \prod_{i=1}^{K} q(\eta_i|\lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n|\phi_n) \]
2 Variational Equation Derivation

2.1 First part of equation 2

In this section I solve equation \( E_q[\log p(w_{1:N}|\beta_{1:K}, z_{1:N})] \)

\[
\log p(w_{1:N}|\beta_{1:K}, z_{1:N}) = \log z_n^T \beta w_n
\]

\[
E_q[\log p(w_{1:N}|\beta_{1:K}, z_{1:N})] = \sum_{i=1}^{K} \log z_n^T \beta w_n q(z_n|\phi_n) = \phi_n^T \log \beta w_n
\] (5)

2.2 Second part of equation 2

I try to solve the equation \( \sum_{n=1}^{N} E_q[\log p(z_n|\eta_d)] \). Probability distribution of \( z_n \) is given by \( p(z_n|\eta_d) = \frac{\exp(\eta_d z_n)}{\sum_{i=1}^{K} \exp(\eta_i)} \). The derivation is given in the following equations 6

\[
\log p(z_n|\eta_d) = \eta_d z_n - \log \sum_{i=1}^{K} \exp(\eta_i)
\] (6)

\[
E_q[\log p(z_n|\eta_d)] = E_q[\eta_d z_n] - E_q[\log \sum_{i=1}^{K} \exp(\eta_i)]
\]
\[ E_q[q_dz_n] = \sum_{z_n} \int \eta_d q(\eta_d | \lambda, \nu^2) d\eta_n q(z_n | \phi_n) \]
\[ = \sum_{z_n} E_q[\eta] z_n q(z_n | \phi_n) dz_n \]
\[ = \sum_{z_n} \lambda z_n q(z_n | \phi_n) dz_n \]
\[ = \lambda^T \phi_n \]

Using concave-conjugate of log function we can derive the second part of this equation. \( \log x \leq \zeta^{-1} x + \log \zeta - 1 \)
\[ E_q[\log \sum_{i=1}^K \exp(\eta_i)] \leq \zeta^{-1} \sum_{i=1}^K (E_q[\exp \eta_i]) + \log(\zeta) - 1 \]
\[ E_q[\exp \eta_i] = \exp\{\lambda_i + \nu_i^2/2\} \]

With these two equations we get the whole second equation
\[ E_q[\log p(z_n | \eta_d)] = \lambda^T \phi_n - \zeta^{-1} (\sum_{i=1}^K \exp\{\lambda_i + \nu_i^2/2\}) - \log(\zeta) + 1 \]

2.3 Third part of equation 2
In this section I try to solve \( E_q[\log p(\eta_d | \mu, \Sigma)] \).
\[
E_q[\log p(\eta_d | \mu, \Sigma)] = \frac{1}{2} \log |\Sigma^{-1}| - \frac{K}{2} \log 2\pi - \frac{1}{2} E_q[(\eta - \mu)^T \Sigma^{-1} (\eta - \mu)] \\
E_q[(\eta - \mu)^T \Sigma^{-1} (\eta - \mu)] = Tr(diag(\nu^2)\Sigma^{-1}) + (\lambda - \mu)^T \Sigma^{-1} (\lambda - \mu) \]
2.4 Fourth part of equation 2

Solve $E_q[\log p(y|\sigma^2, \alpha, z_{1:N})]$ 

\[ p(y|\sigma^2, \alpha, z_{1:N}) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left( -\frac{(\alpha^T \bar{z} - y)^2}{2\sigma^2} \right) \]

\[ \log p(y|\sigma^2, \alpha, z_{1:N}) = -\frac{1}{2} \log (2\pi\sigma^2) - \frac{(\alpha^T \bar{z} - y)^2}{2\sigma^2} \]

\[ (\alpha^T \bar{z} - y)^2 = y^2 + (\alpha^T \bar{z})^2 - 2y\alpha^T \bar{z} \]

\[ \sum_z (\alpha^T \bar{z} - y)^2 q(z_{1:N}|\phi_{1:N}) = \sum_z y^2 q(z_{1:N}|\phi_{1:N}) \]

\[ + \sum_z (\alpha^T \bar{z})^2 q(z_{1:N}|\phi_{1:N}) \]

\[ - \sum_z 2y\alpha^T \bar{z} q(z_{1:N}|\phi_{1:N}) \]

\[ = y^2 + \alpha^T E[\bar{z}\bar{z}^T] \alpha - 2y\alpha^T E[\bar{z}] \]

\[ E_q[\log p(y|\sigma^2, \alpha, z_{1:N})] = -\frac{1}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} (y^2 + \alpha^T E[\bar{z}\bar{z}^T] \alpha - 2y\alpha^T E[\bar{z}] \]

\[ E_q[\bar{z}] = \frac{1}{N} \sum_{n=1}^{N} \phi_n \]

\[ E_q[\bar{z}\bar{z}^T] = \frac{1}{N^2} \left( \sum_{n=1}^{N} \sum_{m \neq n} \phi_n \phi_m^T + \sum_{n=1}^{N} \text{diag}\{\phi_n\} \right) \]

(11)

2.5 Fifth part of equation 2

Solve $H(q)$

\[ \sum_{i=1}^{K} \frac{1}{2} \left( \log \nu_i^2 + \log 2\pi + 1 \right) - \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{n,i} \log \phi_{n,i} \]  (12)

3 Variational E-Step

In this section I derive update equations for all variational parameters $\phi, \lambda, \nu$ and $\zeta$

3.1 Maximization with respect to $\phi$

Writing objective function with just $\phi$ terms we get

\[ L(\phi_n) = \phi_n^T \log \beta w_n + \lambda^T \phi_n - \frac{1}{2\sigma^2} (\alpha^T E[\bar{z}\bar{z}^T] \alpha - 2y\alpha^T E[\bar{z}] \]

\[ - \sum_{i=1}^{K} \phi_{n,i} \log \phi_{n,i} \]  (13)
These are one word update. Differentiating this with respect to $\phi_n$ is given by equation 14

$$
\alpha^T E[\bar{z} \bar{z}^T] \alpha = \alpha^T \frac{1}{N^2} \left( \sum_{n=1}^{N} \sum_{m \neq n} \phi_n \phi_m^T + \sum_{n=1}^{N} \text{diag} \{ \phi_n \} \right) \alpha
$$

$$
= \frac{1}{N^2} \left( \alpha^T \left( \sum_{n=1}^{N} \sum_{m \neq n} \phi_n \phi_m^T \right) \alpha + \alpha^T \sum_{n=1}^{N} \text{diag} \{ \phi_n \} \alpha \right)
$$

$$
\frac{d (\alpha^T E[\bar{z} \bar{z}^T] \alpha)}{d \phi_n} = \frac{1}{N^2} \left( \alpha^T \sum_{m \neq n} \phi_m^T \alpha + \alpha \alpha \right)
$$

$$
\frac{dL(\phi_n)}{d\phi_n} = \log \beta w_n + \lambda + \frac{y \alpha}{N \sigma^2} - \frac{1}{N^2} \left( \alpha^T \sum_{m \neq n} \phi_m^T \alpha + \alpha \alpha \right) - \log \phi_n
$$

$$
\phi_n = \exp \left( \log \beta w_n + \lambda + \frac{y \alpha}{N \sigma^2} - \frac{1}{N^2} \left( \alpha^T \sum_{m \neq n} \phi_m^T \alpha + \alpha \alpha \right) \right)
$$

(14)

3.2 Maximization with respect to $\lambda$

Writing the objective function with just $\lambda$ terms we get.

$$
L(\lambda) = \sum_{n=1}^{N} \lambda^T \phi_n - N \zeta^{-1} \left( \sum_{i=1}^{K} \exp \left( \lambda_i + \nu_i^2/2 \right) \right) - \frac{1}{2} (\lambda - \mu)^T \Sigma^{-1} (\lambda - \mu)
$$

(15)

These terms are for one document. Differentiating the above equation W.R.T $\lambda$.

$$
\frac{L(\lambda)}{d\lambda} = \sum_{n=1}^{N} \phi_n - N \zeta \left( \exp \left( \lambda + \nu^2/2 \right) \right) - \frac{1}{2} \Sigma^{-1} (\lambda - \mu)
$$

(16)

There is no closed form solution for this equation.

3.3 Maximization with respect to $\nu$

Writing the objective function with just $\nu$ terms we get.

$$
L(\nu) = -\frac{N}{\zeta} \left( \sum_{i=1}^{K} \exp \left( \lambda_i + \nu_i^2/2 \right) \right) - Tr(diag(\nu^2) \Sigma^{-1}) + \sum_{i=1}^{K} \frac{1}{2} (\log \nu_i^2)
$$

(17)

These terms are for one document. Differentiating the above equation W.R.T $\nu$.

$$
\frac{L(\nu)}{d\nu} = -\frac{N}{2\zeta} \left( \exp \left( \lambda + \nu^2/2 \right) \right) - \frac{1}{2} diag(\Sigma^{-1}) + \frac{1}{2} \nu^2
$$

(18)
3.4 Maximization with respect to $\zeta$

Writing the objective function with just $\zeta$ terms we get.

$$L(\zeta) = -\frac{N}{\zeta} \left( \sum_{i=1}^{K} \exp(\lambda_i + \nu_i^2/2) \right) - N \log(\zeta)$$  \hspace{1cm} (19)

These terms are for one document. Differentiating the above equation W.R.T $\nu$.

$$\frac{L(\zeta)}{d\zeta} = \frac{N}{\zeta^2} \left( \sum_{i=1}^{K} \exp(\lambda_i + \nu_i^2/2) \right) - \frac{N}{\zeta}$$

$$\zeta = \sum_{i=1}^{K} \exp(\lambda_i + \nu_i^2/2)$$  \hspace{1cm} (20)

4 Variational M-Step

Here I try to update all parameters $\mu, \Sigma, \sigma^2, \alpha$ and $\beta$

4.1 Maximizing w.r.t $\mu$

Writing the objective function with just $\mu$ terms we get.

$$L(\mu) = -\frac{1}{2} (\lambda_d - \mu)^T \Sigma^{-1} (\lambda_d - \mu)$$  \hspace{1cm} (21)

These terms are for one document. Differentiating the above equation W.R.T $\nu$.

$$\frac{L(\mu)}{d\mu} = -\frac{1}{2} \Sigma^{-1} (\lambda_d - \mu)$$

$$\mu = \lambda_d$$  \hspace{1cm} (22)

For the whole dataset.

$$\frac{L(\mu)}{d\mu} = -\sum_{d=1}^{D} \frac{1}{2} \Sigma^{-1} (\lambda_d - \mu)$$

$$\mu = \frac{1}{D} \sum_{d=1}^{D} \lambda_d$$  \hspace{1cm} (23)

4.2 Maximizing w.r.t $\Sigma$

Writing the objective function with just $\mu$ terms we get.

$$L(\Sigma) = \frac{1}{2} \log(\Sigma^{-1}) - \frac{1}{2} \text{Tr}(\text{diag}(\nu^2) \Sigma^{-1}) - \frac{1}{2} (\lambda - \mu)^T \Sigma^{-1} (\lambda - \mu)$$  \hspace{1cm} (24)

These terms are for one document. Differentiating the above equation W.R.T $\Sigma^{-1}$.

$$\frac{L(\Sigma^{-1})}{d\Sigma^{-1}} = \frac{1}{2 \Sigma^{-1}} - \frac{1}{2} I \nu^2 - \frac{1}{2} (\lambda - \mu)(\lambda - \mu)^T$$

$$\Sigma = I \nu^2 + (\lambda - \mu)(\lambda - \mu)^T$$  \hspace{1cm} (25)
For the whole dataset.

\[
\frac{L(\Sigma^{-1})}{d\Sigma^{-1}} = \sum_{d=1}^{D} \frac{1}{2\Sigma^{-1}} - \frac{1}{2} \nu_d^2 - \frac{1}{2} (\lambda_d - \mu) (\lambda_d - \mu)^T
\]

\[
\Sigma = \frac{1}{D} \sum_{d=1}^{D} \nu_d^2 + (\lambda_d - \mu) (\lambda_d - \mu)^T
\]

(26)

4.3 Maximizing w.r.t \(\sigma^2\)

Writing the objective function with just \(\sigma^2\) terms we get.

\[
L(\sigma^2) = -\frac{1}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} (y^2 + \alpha^T E[\tilde{z}\tilde{z}^T]\alpha - 2y\alpha^T E[\tilde{z}])
\]

These terms are for one document. Differentiating the above equation W.R.T \(\sigma^2\).

\[
\frac{dL(\sigma^2)}{d\sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} (y^2 + \alpha^T E[\tilde{z}\tilde{z}^T]\alpha - 2y\alpha^T E[\tilde{z}])
\]

\[
\sigma^2 = 2\pi (y^2 + \alpha^T E[\tilde{z}\tilde{z}^T]\alpha - 2y\alpha^T E[\tilde{z}])
\]

(27)

4.4 Maximizing w.r.t \(\alpha\)

Writing the objective function with just \(\alpha\) terms we get.

\[
L(\alpha) = -\frac{1}{2\sigma^2} (\alpha^T E[\tilde{z}\tilde{z}^T]\alpha - 2y\alpha^T E[\tilde{z}])
\]

These terms are for one document. Differentiating the above equation W.R.T \(\alpha\).

\[
\frac{dL(\alpha)}{d\alpha} = -\frac{1}{2\sigma^2} (\alpha^T E[\tilde{z}\tilde{z}^T] - 2yE[\tilde{z}]^T)
\]

\[
\alpha = 2y * (E[\tilde{z}\tilde{z}^T]^{-1})^T E[\tilde{z}]
\]

(29)

(30)

4.5 Maximizing w.r.t \(\beta\)

Writing the objective function with just \(\beta\) terms we get. Note the \(\lambda\) here is a lagrange parameter not to be confused with the variational mean. \(u\) is the unit vector of all ones of size \(V\), which is the vocabulary size.

\[
L(\beta) = \sum_{n=1}^{N} \phi_n^T \log \beta w_n
\]

\[
L(\beta_k) = \sum_{n=1}^{N} \phi_{n,k} \log \beta_k w_n - \lambda_k (\beta_k u - 1)
\]

These terms are for one document. Differentiating the above equation W.R.T \(\beta_k\). Note \(w_n\) is an identity vector.

\[
\beta_k = \frac{1}{N} \sum_{n=1}^{N} \phi_{n,k} w_n
\]

(31)

(32)