

Filtered Latent Dirichlet Allocation: Variational Bayes Algorithm

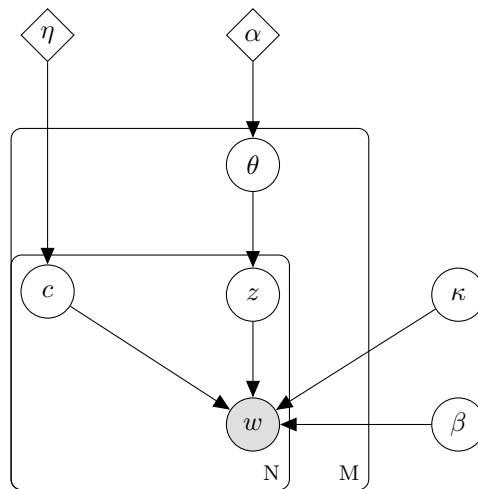
Eric Proffitt

July 14, 2016

Filtered latent Dirichlet allocation (fLDA) is a natural extension of latent Dirichlet allocation that organically captures and filters out corpus-specific stop words which are not otherwise caught by generic stop word lists.

Probabilistic Graphical Model:

θ_d	$\sim \text{Dirichlet}(\alpha)$
c_{d_n}	$\sim \text{Binomial}(\eta)$
z_{d_n}	$\sim \text{Categorical}(\theta_d)$
w_{d_n}	$\sim \text{Categorical}(\kappa) \quad \text{if } c_{d_n} = 0$
w_{d_n}	$\sim \text{Categorical}(\beta_{z_{d_n}}) \quad \text{if } c_{d_n} = 1$



The marginal log-likelihood for fLDA is:

$$\log p(w|\alpha, \beta, \kappa, \eta) = \sum_{d=1}^M \int \sum_{c_d} \sum_{z_d} p(z_d, c_d, \theta_d | w_d, \beta, \kappa, \alpha, \eta) \log \left[\frac{p(w_d, z_d, c_d, \theta_d | \beta, \kappa, \alpha, \eta)}{p(z_d, c_d, \theta_d | w_d, \beta, \kappa, \alpha, \eta)} \right] d\theta$$

The mean-field approximation for the distribution over latent variables is:

$$\prod_{d=1}^M p(z_d, c_d, \theta_d | w_d, \beta, \kappa, \alpha, \eta) \approx \prod_{d=1}^M q(z_d | \phi_d) q(c_d | \tau_d) q(\theta_d | \gamma_d)$$

$$q(z_{d_n} | \phi_{d_n}) = \text{Categorical}(\phi_{d_n})$$

$$q(c_{d_n} | \tau_{d_n}) = \text{Binomial}(\tau_{d_n})$$

$$q(\theta_d | \gamma_d) = \text{Dirichlet}(\gamma_d)$$

Using the KL-Divergence and the mean-field approximation, the variational lower bound for the fLDA marginal log-likelihood, in expected value form, is:

$$\begin{aligned} \log p(w|\beta, \kappa, \alpha, \eta) &\geq \sum_{d=1}^M \left[\mathbb{E}_q[\log p(w_d, z_d, c_d, \theta_d | \beta, \kappa, \alpha, \eta)] - \mathbb{E}_q[\log q(z_d, c_d, \theta_d | \phi_d, \tau_d, \gamma_d)] \right] \\ &= \sum_{d=1}^M \left[\mathbb{E}_q[\log p(w_d | z_d, c_d, \beta, \kappa)] + \mathbb{E}_q[\log p(z_d | \theta_d)] + \mathbb{E}_q[\log p(c_d | \eta)] + \mathbb{E}_q[\log p(\theta_d | \alpha)] \right. \\ &\quad \left. - \mathbb{E}_q[\log q(z_d | \phi_d)] - \mathbb{E}_q[\log q(c_d | \tau_d)] - \mathbb{E}_q[\log q(\theta_d | \gamma_d)] \right] \end{aligned}$$

In particular:

$$\begin{aligned} \bullet \sum_{d=1}^M \mathbb{E}_q[\log p(w_d | z_d, c_d, \beta, \kappa)] &= \sum_{d=1}^M \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{d_n} | z_{d_n}, c_{d_n}, \beta, \kappa)] \\ &= \sum_{d=1}^M \sum_{n=1}^{N_d} \mathbb{E}_q[\log(\beta_{z_{d_n}, w_{d_n}}^{c_{d_n}} \cdot \kappa_{w_{d_n}}^{1-c_{d_n}})] \\ &= \sum_{d=1}^M \sum_{n=1}^{N_d} [\mathbb{E}_q[c_{d_n}] \cdot \mathbb{E}_q[\log \beta_{z_{d_n}, w_{d_n}}] + \mathbb{E}_q[1 - c_{d_n}] \log \kappa_{w_{d_n}}] \\ &= \sum_{d=1}^M \sum_{n=1}^{N_d} [\tau_{d_n} (-\log \kappa_{w_{d_n}} + \sum_{i=1}^K \phi_{d_{in}} \log \beta_{i w_{d_n}}) + \log \kappa_{w_{d_n}}] \end{aligned}$$

- $\sum_{d=1}^M \mathbb{E}_q[\log p(c_d|\eta)] = \sum_{d=1}^M [N_d \log(1 - \eta) + \log(\frac{\eta}{1 - \eta}) \sum_{n=1}^{N_d} \tau_n]$
- $\sum_{d=1}^M \mathbb{E}_q[\log q(c_d|\tau_d)] = \sum_{d=1}^M \sum_{n=1}^{N_d} [\tau_{d_n} \log \tau_{d_n} + (1 - \tau_{d_n}) \log(1 - \tau_{d_n})]$

The remaining expected values are identical to those found in Blei's paper, *Latent Dirichlet Allocation* (2003).

The relevant update equations are as follows:

- $\eta = \frac{\sum_{d=1}^M \sum_{n=1}^{N_d} \tau_{d_n}}{\sum_{d=1}^M N_d}$
- $\kappa_j \propto \sum_{d=1}^M \sum_{n=1}^{N_d} (1 - \tau_{d_n}) w_{d_n}^j$
- $\tau_{d_n} = \frac{\eta}{\eta + (1 - \eta) \kappa_{w_{d_n}} \prod_{i=1}^K \beta_{i w_{d_n}}^{-\phi_{d_{in}}}}$
- $\phi_{d_{in}} \propto \beta_{i w_{d_n}}^{\tau_{d_n}} \exp(\psi(\gamma_i) - \psi(\sum_{l=1}^K \gamma_l))$
- $\beta_{ij} \propto \sum_{d=1}^M \sum_{n=1}^{N_d} \tau_{d_n} \phi_{d_{in}} w_{d_n}^j$

The resulting probability vector κ is the probability distribution over the vocabulary used for drawing stop words in the associated generative process. Therefore those words with the highest probability in κ are those most likely to be corpus-specific stop words.