Filtered Latent Dirichlet Allocation: Variational Bayes Algorithm

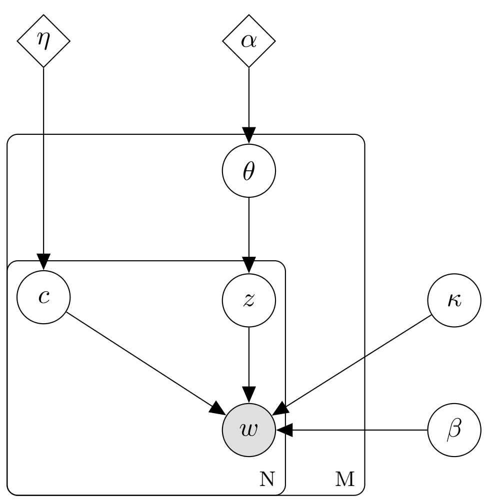
Eric Proffitt

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:

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



The marginal log-likelihood for fLDA is:

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

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

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

$$\begin{aligned} q(z_{d_n}|\phi_{d_n}) &= \text{Categorical}(\phi_{d_n}) \\ q(c_{d_n}|\tau_{d_n}) &= \text{Binomial}(\tau_{d_n}) \\ q(\theta_d|\gamma_d) &= \text{Dirichlet}(\gamma_d) \end{aligned}$$

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

$$\log p(w|\alpha, \beta, \kappa, \eta) \geq \sum_{d=1}^{M} \Big[\mathrm{E}_q[\log p(w_d, z_d, \theta_d | \alpha, \beta, \kappa, \eta)] - \mathrm{E}_q[\log q(z_d, c_d, \theta_d | \phi_d, \tau_d, \gamma_d)]\Big]$$

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

In particular:

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

- $\sum_{d=1}^{M} \mathrm{E}_q[\log p(c_d|\eta)] = \sum_{d=1}^{M} \big[N_d \log(1-\eta) + \log(\frac{\eta}{1-\eta}) \sum_{n=1}^{N_d} \tau_n \big]$

- $\sum_{d=1}^{M} \mathrm{E}_q[\log q(c_d|\tau_d)] = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \big[\tau_{d_n} \log \tau_{d_n} + (1-\tau_{d_n}) \log(1-\tau_{d_n}) \big]$

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

The relevant update equations are as follows:

- $\eta = \frac{\sum_{d=1}^{M} \sum_{n=1}^{N_d} \tau_{d_n}}{\sum_{d=1}^{M} N_d}$

- $\kappa_j \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} (1-\tau_{d_n}) w_{d_n}^{j}$

- $\tau_{d_n} = \frac{\eta}{\eta + (1-\eta) \kappa_{w_{d_n}} \prod_{i=1}^{K} \beta_{i w_{d_n}}^{-\phi_{d_{in}}}}$

- $\phi_{d_{in}} \propto \beta_{i w_{d_n}}^{\tau_{d_n}} \exp\big(\psi(\gamma_i) - \psi(\sum_{l=1}^{K} \gamma_l) \big)$

- $\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \tau_{d_n} \phi_{d_{in}} w_{d_n}^{j}$

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.