CS598JHM: **Advanced NLP** (Spring '10)

Lecture 7: Variational inference for LDA

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

http://www.cs.uiuc.edu/class/sp10/cs598jhm

Variational inference for LDA

Another approximate inference method for inferring the posterior of the hidden variables given the data:

$$p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}, \alpha, \eta) = \frac{p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}{\int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{\vec{z}} p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}$$

References (and figures in today's slides):
- D. Blei and J. Lafferty. **Topic Models.** In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications.* Taylor and Francis, 2009.
- D. Blei, A. Ng, and M. Jordan. **Latent Dirichlet allocation**. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

Variational inference

Approximate the intractable posterior $p(H \mid D)$ with a tractable distribution $q(H \mid D, V)$

$q(H \mid D, V)$ is from a family of simpler distributions defined by a set of free variational parameters V

Variational inference:
Find those parameters V which minimize the KL divergence $KL(q(H \mid D,V) \| p(H \mid D))$ to the true posterior

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

- We can do this without having to compute the actual posterior
- We can't do this exactly, but we can do it up to a constant that is independent of the variational parameters (constant=log likelihood of data under the model)
- The variational parameters V we'll find will depend on the data D

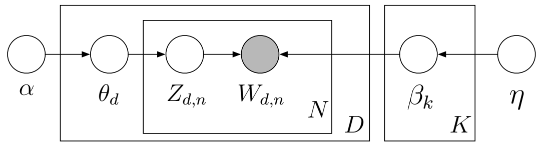
Mean field variational distribution for LDA

Assumptions:
- All variables are independent of each other.
- Each variable has its own variational parameter

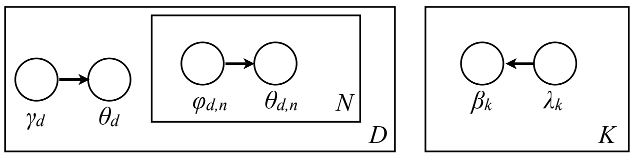
The model:
- **Probability of topic z given document** d: $q(\theta_d \mid \gamma_d)$
  Each document has its own Dirichlet prior γ_d
- **Probability of word w given topic** z: $q(\beta_z \mid \lambda_z)$
  Each topic has its own Dirichlet prior λ_z
- **Probability of topic assignment to word** $w_{d,n}$: $q(z_{d,n} \mid \varphi_{d,n})$
  Each word position $word[d][n]$ has its own prior $\varphi_{d,n}$

A graphical model

LDA:



The variational approximation:



The variational posterior

$$q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) = \prod_{k=1}^{K} q(\vec{\beta}_k \mid \vec{\lambda}_k) \prod_{d=1}^{D} \left(q(\vec{\theta}_{d} \mid \vec{\gamma}_d) \prod_{n=1}^{N} q(z_{d,n} \mid \vec{\phi}_{d,n}) \right)$$

Inference = minimizing KL divergence:

$$\arg \min_{\vec{\gamma}_{1:D}, \vec{\lambda}_{1:K}, \vec{\phi}_{1:D,1:N}} \mathrm{KL}(q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) || p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}))$$

The objective function L turns out to be

$$\mathcal{L} = \sum_{k=1}^{K} \mathrm{E}[\log p(\vec{\beta}_k \mid \eta)] + \sum_{d=1}^{D} \mathrm{E}[\log p(\vec{\theta}_d \mid \vec{\alpha})] + \sum_{d=1}^{D} \sum_{n=1}^{N} \mathrm{E}[\log p(Z_{d,n} \mid \vec{\theta}_d)]$$
$$+ \sum_{d=1}^{D} \sum_{n=1}^{N} \mathrm{E}[\log p(w_{d,n} \mid Z_{d,n}, \vec{\beta}_{1:K})] + \mathrm{H}(q),$$

Inference

Inference = minimizing KL divergence:

$$\arg \min_{\vec{\gamma}_{1:D}, \vec{\lambda}_{1:K}, \vec{\phi}_{1:D,1:N}} \mathrm{KL}(q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) || p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}))$$

The objective function L turns out to be the sum of the expectation of the log probabilities of the posterior under the variational parameters and the entropy of q

$$\mathcal{L} = \sum_{k=1}^{K} \mathrm{E}[\log p(\vec{\beta}_k \mid \eta)] + \sum_{d=1}^{D} \mathrm{E}[\log p(\vec{\theta}_d \mid \vec{\alpha})] + \sum_{d=1}^{D} \sum_{n=1}^{N} \mathrm{E}[\log p(Z_{d,n} \mid \vec{\theta}_d)]$$
$$+ \sum_{d=1}^{D} \sum_{n=1}^{N} \mathrm{E}[\log p(w_{d,n} \mid Z_{d,n}, \vec{\beta}_{1:K})] + \mathrm{H}(q),$$

Variational EM

Initialization:
- Define an initial distribution q

Iterate:
Update each variational parameter
with the expectation of the true posterior
under the variational distribution

Relation between true and variational parameters
- True posterior =
  Dirichlet(hyperparameter + observed frequencies)
- Variational posterior =
  Dirichlet(hyperparameter + expectation of observed frequencies)

Variational inference algorithm

One iteration of mean field variational inference for LDA

(1) For each topic k and term v:

(8) $$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

(2) For each document d:

(a) Update γ_d:

(9) $$\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

(b) For each word n, update $\vec{\phi}_{d,n}$:

(10) $$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{\Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\textstyle\sum_{v=1}^{V} \lambda_{k,v}^{(t+1)})\right\},$$

where Ψ is the digamma function, the first derivative of the $\log \Gamma$ function.