Latent Dirichlet Allocation

Yuriy Sverchkov

Intelligent Systems Program
University of Pittsburgh

October 13, 2011

Outline

Outline

Probabilistic Latent Semantic Analysis (pLSA)
- A quick review

Latent Dirichlet Allocation (LDA)
- The LDA model

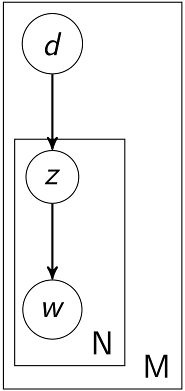
Variational inference

Learning
- Variational EM

Applications, Extensions

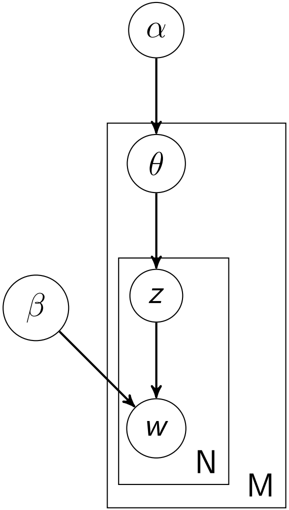
pLSA: A quick review

- Generative model:
  - M documents, to generate each document, do N times:
  - Pick one of k topics,
  - Pick one of V words according to the topic.
- Parameters:
  - $P(z|d)$: $M(k-1)$ parameters defining the distribution of topics for each training document.
  - $P(w|z)$: $k(V-1)$ parameters defining the distribution of words for each topic.
- Number of parameters grows linearly with the training data: A sign of overfitting!
- Not truly generative: learns topic mixtures only for training documents.



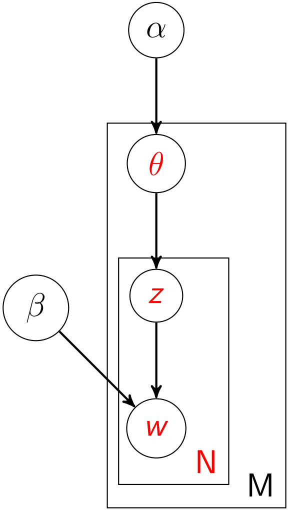
Outline

Latent Dirichlet Allocation



- Main idea: a fully Bayesian version of pLSI: adds a prior distribution over topic mixtures.
- Notation:
  - $\mathbf{w}_m$ is a document: a sequence of N words.
  - A word w_{mn} is a V-vector with 1 in one coordinate and 0 in the rest.
  - θ is a k-vector representing the mixture of topics for a document.
  - A topic z_n is a k-vector with 1 in one coordinate and 0 in the rest.
  - β is a $k \times V$ matrix defining $P(w^j = 1 | z^i = 1)$.

LDA: Generative model



For each of M documents $\mathbf{w}_m$,

1. Choose $N_m \sim \text{Poisson}(\xi)$.
2. Choose $\theta_m \sim \text{Dir}(\alpha)$.
3. For each of the N words w_{mn}:
   - 3.1 Choose a topic $z_{nm} \sim \text{Multinomial}(\theta)$.
   - 3.2 Choose a word w_{nm} with probability $p(w_{nm}|z_{nm}, \beta)$.

Inference

Main inference problem: Estimate topic mixture of a document **w**:

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)}$$

The denominator is intractable to compute:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(w_n | z_n, \beta) p(z_n | \theta) \right) d\theta$$

$$= \int \left(\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1^k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} \underbrace{(\theta_i \beta_{ij})^{w_n^j}}_{\beta \text{ and } \theta \text{ coupled!}} \right) d\theta$$

Outline

Introduction to variational inference

Goal: Estimate posterior $P(X|D)$ by a distribution $Q(X)$

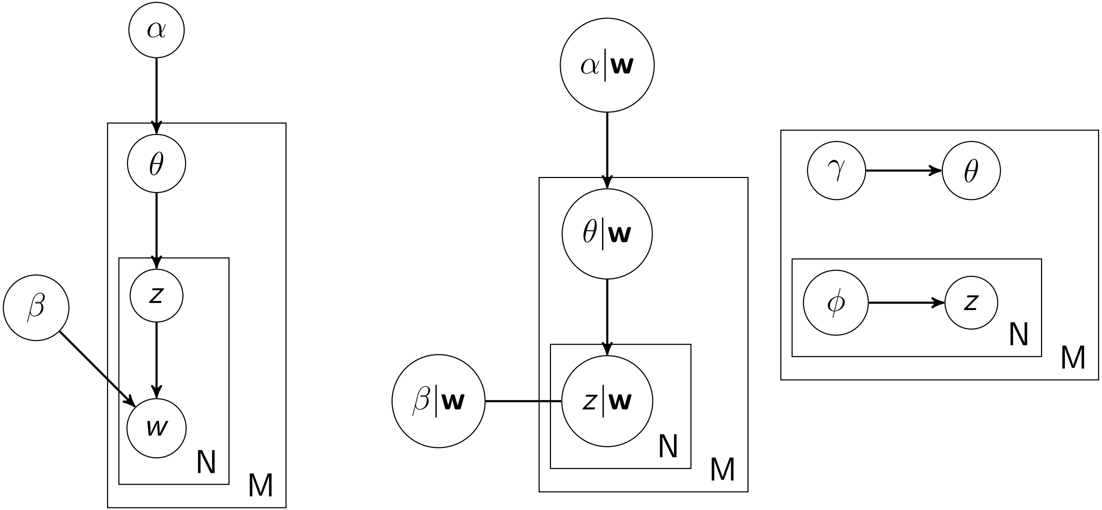
X latent variables

D observed variables (data)

Minimize the KL divergence:

$$
\begin{aligned}
D_{KL}(Q|P) &= \mathop{\mathrm{E}}_{Q(X)} \log \frac{Q(X)}{P(X|D)} \\
&= \mathop{\mathrm{E}}_{Q(X)} \log \left(Q(X) \frac{P(D)}{P(X,D)} \right) \\
&= \underbrace{\log P(D)}_{\text{log-evidence}} + \underbrace{\mathop{\mathrm{E}}_{Q(X)} (\log Q(X) - \log P(X,D))}_{-\mathcal{L}(Q) \text{ variational free energy}}
\end{aligned}
$$

The variational distribution for LDA



Variational free energy for LDA

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{n} q(z_n|\phi_n)$$

$$\begin{aligned}\mathcal{L}(\gamma, \phi; \alpha, \beta) &= -\mathop{\mathrm{E}}_q \log q(\theta, \mathbf{z}|\gamma, \phi) + \mathop{\mathrm{E}}_q \log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) \\ &= \mathop{\mathrm{E}}_q \log p(\theta|\alpha) + \mathop{\mathrm{E}}_q \log p(\mathbf{z}|\theta) + \mathop{\mathrm{E}}_q \log p(\mathbf{w}|\mathbf{z}, \beta) - \mathop{\mathrm{E}}_q \log q(\theta|\gamma) - \mathop{\mathrm{E}}_q \log q(\mathbf{z}|\phi)\end{aligned}$$

We can compute each term (no coupling between parameters): $\mathcal{L}$ is a tractably computable lowerbound for the posterior.

Outline

Learning: Variational EM

- We have a model with latent variables: The natural choice for learning is EM.
- Optimizing the likelihood directly is intractable.
- Solution: optimize the negative variational free energy instead.
- E-Step: $\text{argmax}_{\gamma,\phi}\, \mathcal{L}(\gamma, \phi; \alpha, \beta)$.
- M-step: $\text{argmax}_{\alpha,\beta}\, \mathcal{L}(\gamma, \phi; \alpha, \beta)$.

Outline

- Extensions:
  - Smoothing: Dirichlet prior on β
  - Supervised extensions: sLDA, MedLDA – Add a class variable that is dependent on the latent topic.
- Applications:
  - Document Modeling: Classify by latent topic.
  - Collaborative filtering: e.g. Predicting a user's rating based on other users.