Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Bayesian Non-Parametrics

Advanced Machine Learning for NLP

Jordan Boyd-Graber
TEXT ANALYSIS

What about text?

- Gaussian distributions can't model text
- So typically use multinomial distribution as the base distribution
- Remember multinomial:

$$P(N \mid n, \theta) = \frac{n!}{\prod_j N_j!} \prod_j \theta_j^{N_j} \tag{1}$$

What about text?

- Gaussian distributions can't model text [or can they?]
- So typically use multinomial distribution as the base distribution
- Remember multinomial:

$$P(N \mid n, \theta) = \frac{n!}{\prod_j N_j!} \prod_j \theta_j^{N_j} \tag{1}$$

Dirichlet Process Multinomial Mixture Model

- Break off sticks

$$V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{2}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{3}$$

Dirichlet Process Multinomial Mixture Model

- Break off sticks

$$V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{2}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{3}$$

- Draw atoms

$$\Phi_1, \Phi_2, \cdots \sim_{\text{iid}} \text{Dir}(\beta) \tag{4}$$

Dirichlet Process Multinomial Mixture Model

- Break off sticks

$$V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{2}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{3}$$

- Draw atoms

$$\Phi_1, \Phi_2, \cdots \sim_{\text{iid}} \text{Dir}(\beta) \tag{4}$$

- Merge into complete distribution (i.e., $\Theta \sim \text{DP}(\alpha, \beta \vec{u})$)

$$\Theta = \sum_k C_k \delta_{\Phi_k} \tag{5}$$

Dirichlet Process Multinomial Mixture Model

- Break off sticks

$$V_1, V_2, \cdots \sim_{\text{iid}} \text{Beta}(1, \alpha) \tag{2}$$

$$C_k \equiv V_k \prod_{j=1}^{k-1} (1 - V_k) \tag{3}$$

- Draw atoms

$$\Phi_1, \Phi_2, \cdots \sim_{\text{iid}} \text{Dir}(\beta) \tag{4}$$

- Merge into complete distribution (i.e., $\Theta \sim \text{DP}(\alpha, \beta \vec{u})$)

$$\Theta = \sum_k C_k \delta_{\Phi_k} \tag{5}$$

- Draw document word counts (for $n = 1 \dots D_n$)

$$\phi_d \sim \Theta \tag{6}$$

$$w_{d,n} \sim \phi_d \tag{7}$$

Extending DPMM for text: HDP

- Topic models can use multiple topics per document
- Mixture model can only use one
- HDP is the non-parametric extension

Hierarchical Dirichlet Process

- Draw a global distribution over topics (e.g., $H \equiv \text{Dir}(\alpha)$)

$$G_0 \sim \text{DP}(\gamma, H) \tag{8}$$

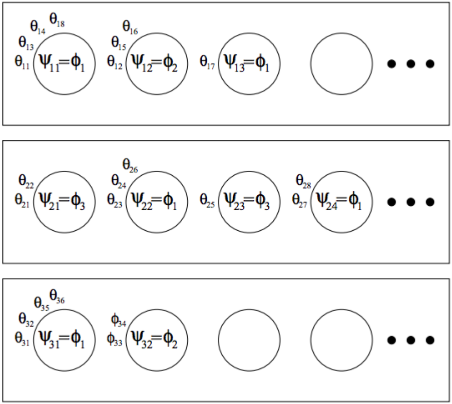
- For each document d, draw distribution over topics

$$\phi_d \sim \text{DP}(\alpha, G_0) \tag{9}$$

    - For each word $w_{d,n}$ in the document, draw it from document distribution

$$w_{d,n} \sim \phi_d \tag{10}$$

Chinese Restaurant Franchise



t: Assignment at global table

z: Assignment at document table

Gibbs Sampling

$$p(z_{dn}=k, t_{dn}=j \mid \boldsymbol{z}^{-ji}, \boldsymbol{t}^{-ji}) \propto \begin{cases} \frac{n_{d,k}}{n_{d,\cdot}+\alpha} f(w_{dn} \mid \Psi_k) & \text{k,j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{m_j}{\gamma+m_\cdot} f(w_{dn} \mid \Psi_j) & \text{k new, j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{\gamma}{\gamma+m_\cdot} f(w_{dn} \mid H_0) & \text{k, j new} \end{cases} \tag{11}$$

Gibbs Sampling

$$p(z_{dn} = k, t_{dn} = j \,|\, \boldsymbol{z}^{-ji}, \boldsymbol{t}^{-ji}) \propto \begin{cases} \frac{n_{d,k}}{n_{d,\cdot}+\alpha} f(w_{dn} \,|\, \Psi_k) & \text{k,j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{m_j}{\gamma+m_\cdot} f(w_{dn} \,|\, \Psi_j) & \text{k new, j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{\gamma}{\gamma+m_\cdot} f(w_{dn} \,|\, H_0) & \text{k, j new} \end{cases} \tag{11}$$

Number of tokens seated in lower-level table

Gibbs Sampling

$$p(z_{dn} = k, t_{dn} = j \,|\, \boldsymbol{z}^{-ji}, \boldsymbol{t}^{-ji}) \propto \begin{cases} \frac{n_{d,k}}{n_{d,\cdot}+\alpha} f(w_{dn} \,|\, \Psi_k) & \text{k,j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{m_j}{\gamma+m_\cdot} f(w_{dn} \,|\, \Psi_j) & \text{k new, j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{\gamma}{\gamma+m_\cdot} f(w_{dn} \,|\, H_0) & \text{k, j new} \end{cases} \tag{11}$$

Number of tokens seated at higher-level table

Gibbs Sampling

$$p(z_{dn} = k, t_{dn} = j \,|\, \boldsymbol{z}^{-ji}, \boldsymbol{t}^{-ji}) \propto \begin{cases} \frac{n_{d,k}}{n_{d,\cdot} + \alpha} f(w_{dn} \,|\, \Psi_k) & \text{k,j existing} \\ \frac{\alpha}{\alpha + n_{d,\cdot}} \frac{m_j}{\gamma + m_\cdot} f(w_{dn} \,|\, \Psi_j) & \text{k new, j existing} \\ \frac{\alpha}{\alpha + n_{d,\cdot}} \frac{\gamma}{\gamma + m_\cdot} f(w_{dn} \,|\, H_0) & \text{k, j new} \end{cases} \tag{11}$$

Lower-level concentration

Gibbs Sampling

$$p(z_{dn} = k, t_{dn} = j \,|\, \boldsymbol{z}^{-ji}, \boldsymbol{t}^{-ji}) \propto \begin{cases} \frac{n_{d,k}}{n_{d,\cdot}+\alpha} f(w_{dn} \,|\, \Psi_k) & \text{k,j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{m_j}{\gamma+m_\cdot} f(w_{dn} \,|\, \Psi_j) & \text{k new, j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{\gamma}{\gamma+m_\cdot} f(w_{dn} \,|\, H_0) & \text{k, j new} \end{cases} \tag{11}$$

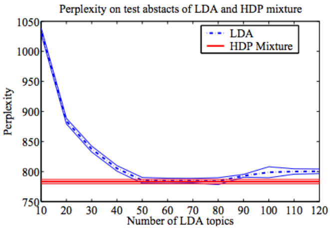
Higher-level concentration

Gibbs Sampling

$$p(z_{dn} = k, t_{dn} = j \,|\, \boldsymbol{z}^{-ji}, \boldsymbol{t}^{-ji}) \propto \begin{cases} \frac{n_{d,k}}{n_{d,\cdot}+\alpha} f(w_{dn} \,|\, \Psi_k) & \text{k,j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{m_j}{\gamma+m_\cdot} f(w_{dn} \,|\, \Psi_j) & \text{k new, j existing} \\ \frac{\alpha}{\alpha+n_{d,\cdot}} \frac{\gamma}{\gamma+m_\cdot} f(w_{dn} \,|\, H_0) & \text{k, j new} \end{cases} \tag{11}$$

Multinomial (or whatever base distribution)

Discovers Dimensionality



- Discovers dimensionality
- Additional layers can capture different aspects of data
- But only unsupervised objective

Inference

- Very similar to LDA
- Need to worry about truncation
- Can be slower