Graph Clustering Using Hierarchical Dirichlet Process and Variational Inference

Milad Kharratzadeh

Mcgill University
Department of Electrical and Computer Engineering

14/Mar/2012

Goals

- Proposing a model for graph clustering based on **Hierarchical Dirichlet Process**
- Inferring the parameters of the model using **variational inference**

Outline

- Introducing Hierarchical Dirichlet Process (HDP) topic model
- Introducing variational inference
- Proposing the model for graph clustering

Hierarchical Dirichlet Process - Definition

HDP: A **distribution** over a set of **probability measures**

- A global measure: $G_0 \sim \mathrm{DP}(\gamma, H)$
- A set of individual measures: $G_j \sim \mathrm{DP}(\alpha_0, G_0)$
- Hyperparameters: α_0, γ (concentration), H (baseline measure)
- Measures drawn from a Dirichlet Process are discrete with probability one[1]:

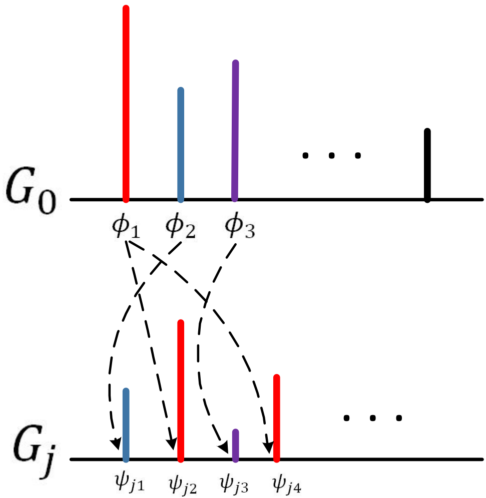
$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\
G_j &= \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}
\end{aligned}
$$

- **Observations:**

$$
\begin{aligned}
\theta_{ji} &\sim G_j \\
x_{ji} &\sim F(\theta_{ji})
\end{aligned}
$$

[1] Ferguson, T. S., "A Bayesian Analysis of Some Nonparametric Problems", *The Annals of Statistics*, 1973

HDP - Stick-breaking Construction [2]



[2] Sethuraman, J., "A Constructive Definition of Dirichlet Priors", *Statistica Sincia*, 1994

HDP - Stick-breaking Construction

Corpus Level

$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\
\beta &\sim \mathrm{GEM}(\gamma) \\
\phi_k &\sim H
\end{aligned}
$$

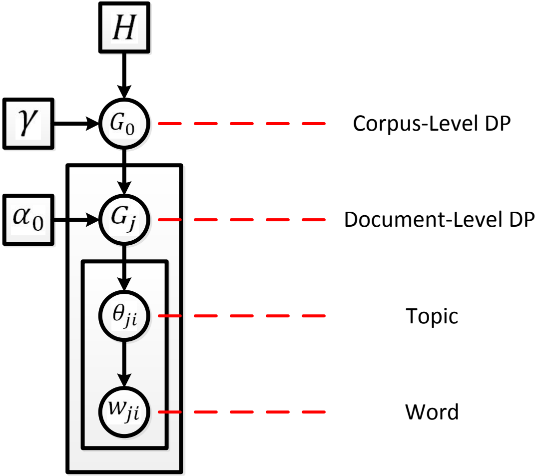
Document Level

$$
\begin{aligned}
G_j &= \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}} \\
\pi_{jt} &\sim \mathrm{GEM}(\alpha_0) \\
\psi_{jt} &\sim G_0
\end{aligned}
$$

Observations (Words)

$$
\begin{aligned}
\theta_{ji} &\sim G_j \\
w_{ji} &\sim \mathrm{Mult}(\theta_{ji})
\end{aligned}
$$

HDP - Topic Model



$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\
G_j &= \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}} \\
c_{jt} &\sim \mathrm{Mult}(\beta) \\
\psi_{jt} &\sim \phi_{c_{jt}} \\
z_{ji} &\sim \mathrm{Mult}(\pi_j) \\
\theta_{ji} &\sim \phi_{c_{jz_{ji}}} \\
w_{ji} &\sim \mathrm{Mult}(\theta_{ji})
\end{aligned}
$$

Variational Inference[3]

Set up

- Observations: $x_1, \ldots, x_n$
- Hidden variables: $z_1, \ldots, z_m$

Motivation

- Posterior distribution: $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{\int_{\mathbf{z}} p(\mathbf{z},\mathbf{x})}$.
- Problem: sometimes intractable denominator

Main idea

- Picking a family of distributions over the latent variables: $q(\mathbf{z}|\nu)$
- Variational parameters: ν
- **Idea:** Find the parameters, ν, that makes q close to the posterior

---

[3] Jordan et al., "An Introduction to Variational Methods for Graphical Models", *Journal of Machine Learning*, vol 37, 1999

Variational Inference

Making q close to posterior

- Jensen's inequality:

$$\log p(\mathbf{x}) \quad \geq \quad E_q\left[\log p(\mathbf{x}, \mathbf{z})\right] - E_q\left[\log q(\mathbf{z})\right]$$

evidence lower bound (ELBO)

- Maximizing ELBO $\equiv$ Minimizing the KL divergence of q and posterior

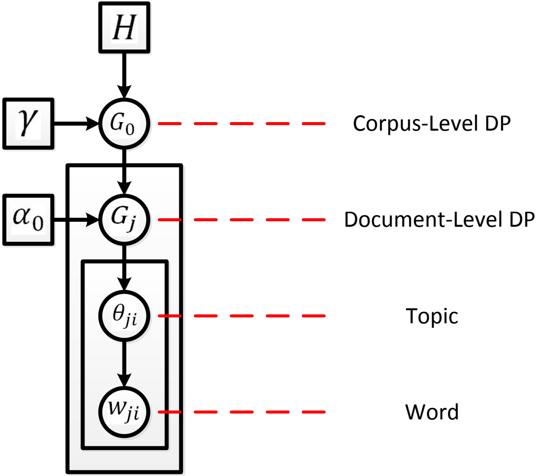
Mean field variational inference

- Assumption: $q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j)$.
- $\rightarrow$ **coordinate ascent inference**

Optimization solution

$$q(z_k) \propto \exp\left\{E_{-k}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right\}$$

HDP - Topic Model (Review!)



$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\
G_j &= \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}} \\
c_{jt} &\sim \mathrm{Mult}(\beta) \\
\psi_{jt} &\sim \phi_{c_{jt}} \\
z_{ji} &\sim \mathrm{Mult}(\pi_j) \\
\theta_{ji} &\sim \phi_{c_{jz_{ji}}} \\
w_{ji} &\sim \mathrm{Mult}(\theta_{ji})
\end{aligned}
$$

Variational inference for HDP topic model

- variational distributions

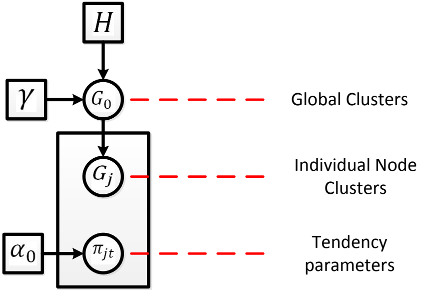
$$q(\beta, \pi, \mathbf{c}, \mathbf{z}, \phi) = q(\beta)q(\pi)q(\mathbf{c})q(\mathbf{z})q(\phi)$$

$$\begin{aligned} q(\mathbf{c}) &= \prod_j \prod_t q(c_{jt}|\varphi_{jt}) && (\textit{multinomial}) \\ q(\mathbf{z}) &= \prod_j \prod_n q(z_{jn}|\zeta_{jn}) && (\textit{multinomial}) \\ q(\phi) &= \prod_k (\phi_k|\lambda_k) && (\textit{Dirichlet}) \end{aligned}$$

- variational inference:
  - Document-level: $\varphi_{jtk} \propto \exp\left(\sum_n \zeta_{jnt} E\left[\log p(w_{jn}|\phi_k)\right] + E_q[\log \beta_k]\right)$
  - Corpus-level: $\lambda_{kw} \propto \eta + \sum_j \sum_{t=1}^{T} \varphi_{jtk} \left(\sum_n \zeta_{jnt} I[w_{jn} = w]\right)$

HDP - Graph Clustering

Model



$$
\begin{aligned}
G_0 &= \sum_{k=1}^{K} \beta_k \delta_{\phi_k} \\
G_j &= \sum_{t=1}^{T} \pi_{jt} \delta_{\psi_{jt}} \\
c_{jt} &\sim \mathrm{Mult}(\beta) \\
\psi_{jt} &\sim \phi_{c_{jt}} \\
\pi_{jt} &\sim \mathrm{GEM}(\alpha_0)
\end{aligned}
$$

- Number of edges between nodes i and j belonging to cluster k: Poisson RV [4] with expected value $\pi_{jt} \cdot \pi_{it'}$ where $c_{jt} = c_{it'} = k$

[4] Ball et al., "An Efficient and Principled Method for Detecting Communities in Networks", *arXiv:1104.3590v1*

Graph Clustering - Variational Inference

- ELBO:

$$\begin{aligned}\log p(G) &\geq E\left[\log p(G, \beta, \pi, c)\right] + E[\log q(G, \beta, \pi, c)] \\ &= E\left[\log p(G|\pi, c)p(c|\beta)p(\pi|\alpha)p(\beta)\right] + E[\log q(G, \beta, \pi, c)]\end{aligned}$$

- Inference (in general):

$$\begin{aligned}q(z_k) &\propto \exp\left\{E_{-k}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right\} \\ \Rightarrow E_k[\log q(z_k)] &= E\left[\log p(\mathbf{z}, \mathbf{x})\right] + const.\end{aligned}$$

- Inference (here):

$$\begin{aligned}E_k[\log q(c)] &= E\left[\log p(G, \beta, \pi, c)\right] + const. \\ &= E\left[\log\left\{p(G|\pi, c)p(c|\beta)p(\pi|\alpha)p(\beta)\right\}\right] + const.\end{aligned}$$

Inferring Parameter "c"

$$
\begin{aligned}
E_k[\log q(c)] &= E\left[\log p(G|\pi, c)\right] + E\left[\log p(c|\beta)\right] + const. \\
q(c) &= \prod_j \prod_t q(c_{jt}|\varphi_{jt}) \qquad \text{(multinomial)} \\
&= \prod_j \prod_t \prod_i \varphi_{jti}^{\mathbf{1}_{[c_{jt}=i]}} \\
E[\log q(c)] &= \sum_j \sum_t \sum_i q(c_{jt} = i) \log(\varphi_{jti}) \\
p(c|\beta) &= \prod_j \prod_t p(c_{jt}|\beta) \qquad \text{(multinomial)} \\
E[\log p(c|\beta)] &= \sum_j \sum_t \sum_i q(c_{jt} = i) \log(\beta_i) \\
E\left[\log p(G|\pi, c)\right] &= ???
\end{aligned}
$$

Graph Likelihood - Pending Work

$$\log p(G|c,\pi) = \sum_{i,j} A_{ij} \log S_{ij} - S_{ij} + const.$$
$$S_{ij} = \sum_{c_{jk}=c_{il}} \pi_{jk}\pi_{il}$$

- We need to compute $E[\log p(G|c,\pi)]$ parametrically — as a function of $q(c_{jt} = i)$.
- Hard to compute parametrically!

Summary

- HDP: A distribution over distributions
- HDP - Topic Model
- Variational inference
- Graph Clustering

Thank You!