Structured Deep Generative Models

Adji Bousso Dieng

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

IPAM Workshop on Interpretable Learning in Physical Sciences
University of California Los Angeles, Los Angeles, CA
October, 2019
Structure
Bayesian Hierarchical Models

- Model is a joint over data \( x \), global latent variables \( \theta \), and local latent variables \( z \)

\[
p(x, z, \theta) = p(\theta) \prod_{n=1}^{N} p(x_n | z_n, \theta)p(z_n | \theta)
\]

- Find structure via posterior inference

\[
p(z, \theta | x) = \frac{p(x, z, \theta)}{p(x)}
\]

- Interpretable probabilistic structure
Define \( K \) shared latent topics \( \beta_k \in \text{Dir}(\eta) \)

For each document \( d \)

1. Draw proportions \( \theta_d \sim \text{Dir}(\alpha) \)

2. For each word \( n \) in the document:
   - Draw assignment \( z_{nd} \sim \text{Cat}(\theta_d) \)
   - Draw word \( w_{nd} \sim \beta_{z_{nd}} \)
Example: Latent Dirichlet Allocation

- Enjoys conjugacy
- Can be fit using coordinate ascent variational inference
- Potential problem in high dimensions (very large vocabularies)
“Example”: Deep Latent Gaussian Model

- Replace prior over $\theta$ with a neural network.
- Model is a joint over data $x$ and local latent variables $z$.

$$p(x, z \mid \theta) = \prod_{n=1}^{N} p(x_n \mid z_n, \theta)p(z_n \mid \theta)$$

- Often $p(z_n \mid \theta) = p(z_n) = \mathcal{N}(0, I)$.
- The conditional $p(x_n \mid z_n, \theta)$ is a neural network with parameters $\theta$ that takes $z$ as input.
- We lost the probabilistic structure over shared global variables $\theta$. 
Variational Inference

Minimizing the KL divergence

\[ KL(q(z \mid \phi) \| p(z \mid x, \theta)) = \log p(x \mid \theta) - \text{ELBO} \]

is equivalent to maximizing the ELBO,

\[ \text{ELBO} = \mathbb{E}_{q(z \mid \phi)} [\log p(x, z \mid \theta) - \log q(z \mid \phi)] \]
Amortized Variational Inference

Parameterize $q$ with a neural network that takes data $x$ as input and maximize ELBO as before:

$$\text{ELBO} = \mathbb{E}_{q(z|x,\phi)} [\log p(x, z | \theta) - \log q(z | x, \phi)]$$
Hierarchical Bayes + Neural Networks

- to learn models with interpretable probabilistic structure
- to deal well with the high dimensionality of the data
- for efficient inference
I will next describe three instances of a family of models for high dimensional data that (1) have an interpretable probabilistic structure and (2) are fit efficiently using amortized variational inference. The key underlying idea is to work in the *meaning space*. 
Embedded Topic Model

- Define a deterministic per-word embedding $\rho \in \mathbb{R}^{V \times E}$
- Define a deterministic shared global embedding matrix $\alpha \in \mathbb{R}^{K \times E}$
- For each document $d$
  1. Draw proportions $\theta_d \sim \mathcal{L}\mathcal{N}(0, I)$
  2. For each word $n$ in the document:
     - Draw assignment $z_{nd} \sim \text{Cat}(\theta_d)$
     - Draw word $w_{nd} \sim \text{Cat}(\text{softmax}(\rho^T \alpha z_{nd}))$
- Deal with high dimensions by working on the embedding space
- Same interpretable probabilistic structure as LDA
- Fit model using amortized variational inference
Embedded Topic Model

- Comparing ETM to LDA on the 20Newsgroup corpus.
- LDA’s performance degrades as the dimensionality increases
- ETM deals well with high dimensionality

![Graph showing Coherence-Normalized Perplexity vs Vocabulary Size for LDA and ETM]
Embedded Topic Model

- Comparing ETM to several document models on the New York Times corpus as the vocabulary (V) increases.
- A good document model is on the top right; interpretable with high predictive power.
Embedded Topic Model

- ETM topic embeddings found in the New York Times corpus.
Embedded Topic Model

- ETM word embeddings found in the New York Times corpus compared to Skipgram word embeddings.

<table>
<thead>
<tr>
<th>Skip-gram embeddings</th>
<th>ETM embeddings</th>
</tr>
</thead>
<tbody>
<tr>
<td>love</td>
<td>love</td>
</tr>
<tr>
<td>loved</td>
<td>joy</td>
</tr>
<tr>
<td>passion</td>
<td>loves</td>
</tr>
<tr>
<td>loves</td>
<td>loved</td>
</tr>
<tr>
<td>affection</td>
<td>passion</td>
</tr>
<tr>
<td>adore</td>
<td>wonderful</td>
</tr>
<tr>
<td>families</td>
<td>children</td>
</tr>
<tr>
<td>grandparents</td>
<td>son</td>
</tr>
<tr>
<td>mother</td>
<td>mother</td>
</tr>
<tr>
<td>friends</td>
<td>father</td>
</tr>
<tr>
<td>relatives</td>
<td>daughter</td>
</tr>
<tr>
<td>man</td>
<td>girl</td>
</tr>
<tr>
<td>girl</td>
<td>boy</td>
</tr>
<tr>
<td>boy</td>
<td>mother</td>
</tr>
<tr>
<td>teenager</td>
<td>daughter</td>
</tr>
<tr>
<td>ideology</td>
<td>ideology</td>
</tr>
<tr>
<td>politicking</td>
<td>speeches</td>
</tr>
<tr>
<td>political</td>
<td>political</td>
</tr>
<tr>
<td>religion</td>
<td>politician</td>
</tr>
<tr>
<td>politic</td>
<td>ideology</td>
</tr>
<tr>
<td>partisanship</td>
<td>ideological</td>
</tr>
</tbody>
</table>
Dynamic Embedded Topic Model

- Define a deterministic per-word embedding $\rho \in \mathbb{R}^{V \times E}$
- Define a latent per-time-step shared global embedding matrix

$$\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \gamma^2 I) \quad \text{where} \quad \alpha_t \in \mathbb{R}^{K \times E}$$

- For each document $d$
  1. Draw $\theta_d \sim \mathcal{L}\mathcal{N}(\eta_{t_d}, \sigma^2 I)$ where $\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 I) \ \forall t$
  2. For each word $n$ in the document:
     - Draw assignment $z_{nd} \sim \text{Cat}(\theta_d)$
     - Draw word $w_{nd} \sim \text{Cat}(\text{softmax}(\rho^T \alpha z_{nd}, t_d))$

- Same interpretable probabilistic structure as Dynamic LDA
- Fit model using structured amortized variational inference w/ LSTM
Dynamic Embedded Topic Model

Trajectory of a topic about climate change found by the model on the United Nations Debates corpus.
Dynamic Embedded Topic Model

- Evolution of word probability across time for eight different topics learned by the model.
Latent Implicit Model Allocator

- Define a deterministic per-observation embedding \( \rho \in \mathbb{R}^{V \times E} \)
- Define a shared set of \( K \) neural networks; each with parameters \( \gamma_k \)
- For each observation \( d \)
  1. Draw proportions \( \theta_d \sim \mathcal{L}\mathcal{N}(0, I) \)
  2. For each element \( n \) in \( d \):
     - Draw assignment \( z_{nd} \sim \text{Cat}(\theta_d) \)
     - Draw noise \( \epsilon_{nd} \sim \mathcal{N}(0, I) \)
     - Compute landmark \( \alpha_{nd} \sim \text{NN}(\epsilon_{nd}; \gamma_{z_{nd}}) \)
     - Draw element \( x_{nd} \sim \text{ExpFam}(g(\rho^T \alpha_{nd})) \)
- Fit model using amortized variational inference
Latent Implicit Model Allocator

- LIMA vs VAE and other structured deep generative models
- All models have same complexity (#parameters)
- Generalization performance as measured by log-likelihood on three benchmark image datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>MNIST</th>
<th>CIFAR-10</th>
<th>CELEBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE</td>
<td>-85.05</td>
<td>-2121</td>
<td>-6518</td>
</tr>
<tr>
<td>SVAE (Johnson et al., 2016)</td>
<td>-85.58</td>
<td>-2144</td>
<td>-6441</td>
</tr>
<tr>
<td>kVAE (Locatello et al., 2018)</td>
<td>-87.61</td>
<td>-2217</td>
<td>-7007</td>
</tr>
<tr>
<td>LIMA + pretraining</td>
<td>-84.92</td>
<td>-2107</td>
<td>-6178</td>
</tr>
<tr>
<td>LIMA w/o pretraining</td>
<td>-85.20</td>
<td>-2108</td>
<td>-6073</td>
</tr>
</tbody>
</table>