Augment and Reduce: Stochastic Inference for Large Categorical Distributions

Adji Bousso Dieng
Collaborators

+ Francisco J. R. Ruiz
+ Michalis Titsias
+ David M. Blei

Augment and Reduce: Stochastic Inference for Large Categorical Distributions
Under review at ICML, 2018.
Categorical Distributions: Applications

Categorical distributions are ubiquitous in Statistics and Machine Learning

→ discrete choice models
→ language models
→ recommendation systems
→ reinforcement learning
One widely applied parameterization of a categorical is the softmax,

\[ p(y = k \mid \psi) = \text{softmax}(\psi)|_k = \frac{e^{\psi_k}}{\sum_{k'} e^{\psi_{k'}}} \]

Transforms reals into probabilities

Can be costly because of normalization ... \( O(K) \)

A computational burden when learning with categorical distributions
A Closer Look at Softmax

→ Draw random standard Gumbel errors i.i.d.,

\[ \varepsilon_k \sim \text{Gumbel}(\varepsilon \mid 0, 1) \]

→ Define a utility for each outcome \( k \),

\[ \psi_k + \varepsilon_k \]

→ Choose the outcome with the largest utility,

\[ y = \arg \max_k (\psi_k + \varepsilon_k) \]

→ Integrate out the error terms (\( \varepsilon_k \)'s) to find the marginal \( p(y \mid \psi) \)

Softmax is the marginal!!
The Augmented Model

→ The augmented model is

\[ p(y = k, \varepsilon \mid \psi) = \phi(\varepsilon) \prod_{k' \neq k} \Phi(\varepsilon + \psi_k - \psi_{k'}) \]

→ Nice property: The log-joint has a summation over the categories,

\[ \log p(y = k, \varepsilon \mid \psi) = \log \phi(\varepsilon) + \sum_{k' \neq k} \log \Phi(\varepsilon + \psi_k - \psi_{k'}) \]

→ This enables fast unbiased estimates,

- Sample a subset of outcomes \( S \subseteq \{1, \ldots, K\} \setminus \{k\} \)
- Compute an estimate of the log-joint

\[ \log \phi(\varepsilon) + \frac{K - 1}{|S|} \sum_{k' \in S} \log \Phi(\varepsilon + \psi_k - \psi_{k'}) \]

→ This has \( \mathcal{O}(|S|) \) complexity
The Inference Algorithm: Variational EM

→ We are not interested in the log-joint, but in the log-marginal

→ Variational inference relates both quantities,

\[ \log p(y \mid \psi) \geq \mathbb{E}_{q(\varepsilon)} \left[ \log p(y, \varepsilon \mid \psi) - \log q(\varepsilon) \right] \]

→ Maximize the bound using variational EM
  
  - E step: Optimize w.r.t. the distribution \( q(\varepsilon) \)
  - M step: Take a gradient step w.r.t. \( \psi \)

→ The complexity is controlled by the user (via \(|S|\))
Things are Prettier with Softmax

→ We can compute the optimal $q(\varepsilon)$ distribution,

$$q^*(\varepsilon) = \text{Gumbel}(\log \eta^*, 1), \quad \eta^* = 1 + \sum_{k' \neq k} e^{\psi_{k'} - \psi_k}$$

→ This is $O(K)$. Instead, set

$$q(\varepsilon) = \text{Gumbel}(\log \eta, 1)$$

→ Estimate the optimal natural parameter in $O(|S|)$,

$$\tilde{\eta} = 1 + \frac{K - 1}{|S|} \sum_{k' \in S} e^{\psi_{k'} - \psi_k}$$

(to update $\eta$, take a step in the direction of the natural gradient)
Scale All Categorical Distributions!

→ Choose other distributions for \( \varepsilon \) to get other models,
  
  – Gaussian for multinomial probit
  – Logistic for multinomial logistic

→ Form Monte Carlo gradient estimators using reparameterization

→ Useful for both E and M steps
Empirical Evidence

→ Baselines:
  - Exact Softmax for MNIST and Bibtext
  - OVE – Also a lower bound but only applicable to softmax

→ Time complexity (top) and Predictive performance (bottom)

<table>
<thead>
<tr>
<th>dataset</th>
<th>OVE (Titsias, 2016)</th>
<th>A&amp;R [this paper]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>softmax</td>
<td>multi. probit</td>
</tr>
<tr>
<td></td>
<td>log lik</td>
<td>acc</td>
</tr>
<tr>
<td>MNIST</td>
<td>0.336 s</td>
<td>0.337 s</td>
</tr>
<tr>
<td>Bibtext</td>
<td>0.181 s</td>
<td>0.188 s</td>
</tr>
<tr>
<td>Omniglot</td>
<td>4.47 s</td>
<td>4.65 s</td>
</tr>
<tr>
<td>EURLex-4K</td>
<td>5.54 s</td>
<td>5.65 s</td>
</tr>
<tr>
<td>AmazonCat-13K</td>
<td>2.80 h</td>
<td>2.80 h</td>
</tr>
</tbody>
</table>

|                  | multi. logistic      |                  |
|                  | log lik              | acc              |
|                  |                      |                  |
| MNIST             |                      | -0.271           |
| Bibtext           |                      | -3.036           |
| Omniglot          |                      | -5.171           |
| EURLex-4K         |                      | -4.593           |
| AmazonCat-13K     |                      | -3.795           |

Table 2.

Table 3 &

Table 4.

Table 5.

Table 6.
Empirical Evidence

→ Quality of the bound
Take Home: The A&R Recipe

→ Choose a distribution for $\varepsilon$

→ Augment your model with $\varepsilon$ to get an augmented model—

$$\mathcal{L} = \log p(y = k, \varepsilon | \psi) = \log \phi(\varepsilon) + \sum_{k' \neq k} \log \Phi(\varepsilon + \psi_k - \psi_{k'})$$

→ Reduce cost to $\mathcal{O}(|S|)$ with an estimate of the log-joint,

$$\tilde{\mathcal{L}} \approx \tilde{\mathcal{L}} = \log \phi(\varepsilon) + \frac{K - 1}{|S|} \sum_{k' \in S} \log \Phi(\varepsilon + \psi_k - \psi_{k'})$$

→ Use stochastic variational EM with the bound

$$\log p(y | \psi) \geq \mathbb{E}_{q(\varepsilon)} [\mathcal{L} - \log q(\varepsilon)]$$

A&R is a principled method that scales up training for models involving large categorical distributions using latent variable augmentation and stochastic variational inference.