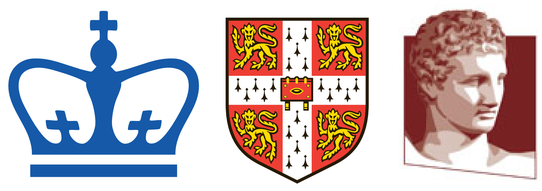
Scalable Large-Scale Classification with Latent Variable Augmentation

Francisco J. R. Ruiz[†‡] Michalis K. Titsias[*] David M. Blei[†]

[†]Columbia University [‡]University of Cambridge [*]Athens University of Economics and Business



Summary

- **Goal:** Scalable inference method for categorical distributions with a large number of outcomes
- We develop an inference algorithm based on stochastic variational inference
- We use the utility-based perspective
- Valid for distributions such as softmax and multinomial probit
- Key ideas: Latent variable augmentation, stochasticity via subsampling classes

Introduction

- Categorical distributions are widely applied in many areas of machine learning
  - Classification
  - Neural language models [Bengio et al.]
  - Discrete choice models [McFadden; McFadden & Train]
- Many approaches to parameterize a categorical distribution
- One approach: Define the utilities ψ_k and pick choice

$$y = \arg\max_k (\psi_k + \varepsilon_k)$$

  for some errors ε_k and $k = 1, \dots, K$.
- Scalability problem: Evaluating $p(y = k \,|\, \psi)$ is $\mathcal{O}(K)$, which is problematic when the number of classes (choices) K is large
- Our approach:
  1. Augment the model with an auxiliary latent variable
  2. Perform variational inference on the augmented model
  3. Subsample classes to obtain unbiased gradient estimators with cost lower than K
- Advantages:
  - Rigorous lower bound (no heuristics or approximations)
  - Speed controlled by the class subsampling procedure
  - Maintains the utility-based perspective (no need to change the model)
  - Valid for a general class of methods (softmax, multinomial probit, multinomial logistic)

Related Work

- Approximate distributed approaches [Grave et al.]
- Methods that attempt to perform exact computations [Gopal & Yang; Vijayanarasimhan et al.]
- Sampling-based methods [Bengio & Sénécal, Mikolov et al.; Devlin et al., Ji et al.]
- Noise contrastive estimation [Smith & Jason; Gutmann & Hyvärinen]
- Random nearest neighbor search [Mussmann et al.]
- Hierarchical or stick-breaking models [Kurzynski; Morin & Bengio; Tsoumakas et al.; Beygelzimer et al.; Dembczyński et al.; Khan et al.]
- Softmax variational lower bound [Titsias]

$$\log p(y = k \,|\, \psi) \geq \sum_{k' \neq k} \log \sigma(\psi_k - \psi_{k'})$$

Utility Model

- Utilities:

$$y = \arg\max_k (\psi_k + \varepsilon_k)$$

- Place an iid prior over the errors, $\varepsilon_k \sim \phi(\varepsilon)$
- Probability of the kth outcome:

$$\begin{aligned} p(y = k \,|\, \psi) &= \text{Prob}\left(\psi_k + \varepsilon_k \geq \psi_{k'} + \varepsilon_{k'} \;\; \forall k' \neq k\right) \\ &= \int_{-\infty}^{+\infty} \phi(\varepsilon_k) \left(\prod_{k' \neq k} \int_{-\infty}^{\varepsilon_k + \psi_k - \psi_{k'}} \phi(\varepsilon_{k'}) d\varepsilon_{k'} \right) d\varepsilon_k \\ &= \int_{-\infty}^{+\infty} \phi(\varepsilon) \left(\prod_{k' \neq k} \Phi(\varepsilon + \psi_k - \psi_{k'}) \right) d\varepsilon \end{aligned}$$

  where $\Phi(\varepsilon) = \int_{-\infty}^{\varepsilon} \phi(\tau) d\tau$
- This expression is analogous to the expression for the multinomial probit model [Girolami & Rogers]
- We do not necessarily assume a Gaussian distribution $\phi(\varepsilon)$ over the errors
- For a Gumbel distribution $\phi(\varepsilon)$, this simplifies to the well-known softmax

Latent Variable Augmentation

General idea

- We consider the augmented model

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

- The log-joint involves a summation over the classes
- This enables stochastic variational inference via class subsampling [Hoffman et al.]

Variational inference

- Variational inference on the augmented model $p(y, \varepsilon \,|\, \psi)$
- The ELBO (evidence lower bound) is

$$\mathcal{L} = \mathbb{E}_{q(\varepsilon)} \left[\log \phi(\varepsilon) + \sum_{k' \neq k} \log \Phi(\varepsilon + \psi_k - \psi_{k'}) - \log q(\varepsilon) \right]$$

- Optimal variational distribution:

$$q^\star(\varepsilon) = p(\varepsilon \,|\, y = k, \psi) \propto \phi(\varepsilon) \prod_{k' \neq k} \Phi(\varepsilon + \psi_k - \psi_{k'})$$

- The optimal $q^\star(\varepsilon)$ achieves the exact marginal log-likelihood, but it is generally not available in closed form

1. Softmax augmentation

- Consider a softmax model,

$$\phi_{\text{softmax}}(\varepsilon) = \exp\{-\varepsilon - e^{-\varepsilon}\}, \quad \Phi_{\text{softmax}}(\varepsilon) = \exp\{-e^{-\varepsilon}\}$$

- The optimal variational distribution has closed form,

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

- But it is hard to compute because it involves a $\mathcal{O}(K)$ summation
- Instead, we set

$$q_{\text{softmax}}(\varepsilon \,;\, \eta) = \text{Gumbel}(\varepsilon \,;\, \log \eta, 1)$$

  This is an exponential family distribution with natural parameter η
- This choice of $q_{\text{softmax}}(\varepsilon)$ gives the ELBO

$$\mathcal{L}_{\text{softmax}} = 1 - \log(\eta) - \frac{1}{\eta} \left(1 + \sum_{k' \neq k} e^{\psi_{k'} - \psi_k} \right)$$

This coincides with the log-concavity bound [Bouchard; Blei & Lafferty]

Algorithm (See Algorithm 1 below)

1. Local step (E step)
   - Subsample a set of classes $\mathcal{S} \subseteq \{1, \dots, K\} \setminus \{y\}$
   - Estimate the natural parameter η from these classes
2. Global step (M step)
   - Take a gradient step w.r.t. (the parameters of) ψ

Classification

- Datapoints are (x_n, y_n) pairs, with $y_n \in \{1, \dots, K\}$
- Typically, the utility is $\psi_{nk} = w_k^\top x_n + w_k^{(0)}$
- Gaussian prior on the weights, $p(w) = \mathcal{N}(w \,|\, 0, \sigma_w^2 \mathbf{I})$
- Find the MAP solution for the weights w (also valid for posterior inference on w)

Numerical stability

- The $\exp\{\cdot\}$ function to obtain η may lead to numerical instabilities
- Solution: Use the same set of samples $\mathcal{S}_n$ in the local and global steps
- This leads to a numerically stable algorithm

2. Multinomial Probit Augmentation

- Consider a multinomial probit model,

$$\phi_{\text{probit}}(\varepsilon) = \mathcal{N}(\varepsilon \,;\, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\varepsilon^2},$$

$$\Phi_{\text{probit}}(\varepsilon) = \int_{-\infty}^{\varepsilon} \mathcal{N}(\tau \,;\, 0, 1) d\tau$$

- The expectations in the ELBO are now intractable
- We set $q_{\text{probit}}(\varepsilon) = \mathcal{N}(\varepsilon \,;\, \mu, v^2)$
- We do inference using reparameterization gradients [Kingma & Welling; Titsias & Lázaro-Gredilla; Rezende et al.]
  - Subsample classes to evaluate the log-joint
- For classification, we use amortized inference, with $\mu_n = \mu(x_n)$ and $v_n = v(x_n)$ [Dayan et al.; Gershman & Goodman; Mnih & Gregor; Kingma & Welling; Rezende et al.]

3. Multinomial Logistic Augmentation

- Consider a multinomial logistic model,

$$\phi_{\text{logistic}}(\varepsilon) = \sigma(\varepsilon)\sigma(-\varepsilon), \quad \Phi_{\text{logistic}}(\varepsilon) = \sigma(\varepsilon)$$

- The resulting ELBO resembles the one-vs-each bound [Titsias]
- We set $q_{\text{logistic}}(\varepsilon \,;\, \mu, v) = \frac{1}{v}\sigma\left(\frac{\varepsilon - \mu}{v}\right)\sigma\left(-\frac{\varepsilon - \mu}{v}\right)$ and use amortized inference

Algorithm 1: Variational inference on the augmented softmax

for $t = 1, 2, \dots,$ **do**
  Sample a minibatch of data, $\mathcal{B} \subseteq \{1, \dots, N\}$
  `# Local step (E step):`
  for $n \in \mathcal{B}$ **do**
    Sample a set of labels, $\mathcal{S}_n \subseteq \{1, \dots, K\} \setminus \{y_n\}$
    Update the natural parameter, $\eta_n \leftarrow 1 + \frac{K-1}{|\mathcal{S}_n|} \sum_{k \in \mathcal{S}_n} e^{\psi_{nk} - \psi_{ny_n}}$
  end
  `# Global step (M step):`
  Sample a set of labels, $\mathcal{S}_n \subseteq \{1, \dots, K\} \setminus \{y_n\}$ for each $n \in \mathcal{B}$
  Set $g^{(t)} \leftarrow -\frac{w}{\sigma_w^2} - \frac{N}{|\mathcal{B}|} \frac{K-1}{|\mathcal{S}_n|} \sum_{n \in \mathcal{B}} \frac{1}{\eta_n} \sum_{k \in \mathcal{S}_n} \nabla_w e^{\psi_{nk} - \psi_{ny_n}}$
  Update $w \leftarrow w + \rho^{(t)} g^{(t)}$
end

Experiments

- Synthetic data extending MNIST up to $K = 100$ classes
- Methods: exact softmax, one-vs-each [Titsias], and the three latent variable augmentation schemes
- Parameters: $|\mathcal{S}| = 10$ samples, $|\mathcal{B}| = 200$ datapoints, $\sigma_w^2 = 1$

model	log-likelihood train	log-likelihood test	accuracy train	accuracy test
exact softmax	−0.1835	−0.3594	0.956	0.903
one-vs-each (Titsias, 2016)	−0.2635	**−0.3618**	0.923	**0.900**
softmax augmentation	−0.2925	**−0.3662**	0.917	**0.901**
multinomial probit augmentation	−0.2656	−0.4122	0.922	0.895
multinomial logistic augmentation	−0.2730	−0.3805	0.918	0.898

Table 1: Performance on the extended MNIST data with 100 classes.