Doubly Stochastic Variational Inference

Dmitry Molchanov

Lomonosov Moscow State University

March 18, 2016

Introduction

X – training data
$\Theta \in \mathbb{R}^D$ – latent variables and model parameters
Joint density:

$$p(X, \Theta) = p(X \mid \Theta)p(\Theta) = g(\Theta)$$

How to obtain posterior distribution?

$$p(\Theta \mid X) = ?$$

- MCMC sampling – too slow
- Variational Bayes – usually intractable, doesn't scale well

Variational approximation

$$p(\Theta \mid X) \approx q(\Theta)$$

$$\mathcal{D}_{KL}(q(\Theta) \,||\, p(\Theta \mid X)) \to \min_q$$

$$\Updownarrow$$

$$\mathcal{L}(q) = \int q(\Theta) \log \frac{p(X, \Theta)}{q(\Theta)} d\Theta \to \max_q$$

- Ideal solution: $q(\Theta) = p(\Theta \mid X)$
- Variational Bayes: $q(\Theta) = \prod_{d=1}^{D} q_i(\theta_i) = \arg\max \mathcal{L}(q)$
- Stochastic Variational Inference:
  $q(\Theta) = q(\Theta|\alpha);\ \mathcal{L}(q) = \mathcal{L}(\alpha) \to \max_\alpha$

Approximating family in DSVI

$z \in \mathbb{R}^D \sim \phi(z)$ – arbitry distribution
C is a lower triangular positive definite matrix
$\mu \in \mathbb{R}^D$

$$\Theta = Cz + \mu$$

$$z = C^{-1}(\Theta - \mu)$$

$$q(\Theta) = \frac{1}{\det C}\phi(C^{-1}(\Theta - \mu)) = q(\Theta \,|\, \mu, C)$$

Variational lower bound

$$\mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)}\left[\log g(Cz + \mu)\right] + \sum_{d=1}^{D} \log C_{dd} + \mathcal{H}(\phi)$$

- $\mathcal{L}(\mu, C)$ is concave wrt μ and C if $g(\Theta)$ is concave
- $\mathcal{H}(\phi)$ is just a constant
- We can now easily diffirentiate the lower bound

DSVI gradients

$$\mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)}\left[\log g(Cz + \mu)\right] + \sum_{d=1}^{D} \log C_{dd} + \mathcal{H}(\phi)$$

$$\nabla_\mu \mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)}\left[\nabla_\mu \log g(Cz + \mu)\right]$$

$$\nabla_C \mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)}\left[\nabla_C \log g(Cz + \mu)\right] + \Delta_C$$

$$\Delta_C = [C_{11}^{-1}, C_{22}^{-1}, \ldots, C_{DD}^{-1}]^T$$

$$\nabla_\mu \mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)}\left[\nabla_\Theta \log g(\Theta)|_{\Theta=Cz+\mu}\right]$$

$$\nabla_C \mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)}\left[\nabla_\Theta \log g(\Theta) z^T\Big|_{\Theta=Cz+\mu}\right] + \Delta_C$$

DSVI algorithm

- Initialize μ, C, t
- Repeat
- $t = t + 1$
- $z \sim \phi(z)$
- $\Theta^{(t-1)} = C^{(t-1)} z + \mu^{(t-1)}$
- $\mu^{(t)} = \mu^{(t-1)} + \rho_t \nabla_\Theta \log g(\Theta^{(t-1)})$
- $C^{(t)} = C^{(t-1)} + \rho_t \left(\nabla_\Theta \log g(\Theta^{(t-1)}) z^T + \Delta_{C^{(t-1)}} \right)$

Data subsampling: instead of $\log g(\Theta)$, use $\log g_t(\Theta)$:

$$\log g_t(\Theta) = \frac{N}{M} \log p(X^M, \Theta),$$

where X^M is a minibatch of size M and N is the total number of objects.

Monitoring convergence

- Stabilization of μ and C
- Rolling-window average (i.e. of size 200) of the instantaneous lower bound

The instantaneous value of the lower bound at the tth iteration:

$$\mathcal{L}^{(t)} = \log g(\Theta^{(t)}) + \log \det C^{(t)} + \mathcal{H}(\phi), \qquad \Theta^{(t)} \sim q(\Theta \mid \mu, C)$$

Some tips

- Choose a sophisticated $\phi(z)$, i.e. $\phi(z) = \prod_{d=1}^{D} \phi_d(z_d)$ with different ϕ_d
- Use a simpler μ and C. C can be a block diagonal matrix / diagonal matrix and some components of μ can be fixed at zero
- Simplify $\mathbb{E}_{\phi(z)}\left[\log g(Cz + \mu)\right]$ as much as possible

ARD logistic regression

Model definition:

$$g(\Theta) = \tilde{g}(\Theta)\mathcal{N}(\Theta \mid 0, \Lambda)$$

$$\tilde{g}(\Theta) = p(X \mid \Theta) = \prod_{n=1}^{N} \sigma(y_n x_n^T \Theta)$$

$$\phi(z) = \mathcal{N}(z \mid 0, I), \quad C = \mathrm{diag}(c)$$

Lower bound:

$$\mathcal{L}(\mu, C, \Lambda) = \mathbb{E}_{\phi(z)} \left[\log \tilde{g}(c \circ z + \mu)\right] + \frac{1}{2} \sum_{d=1}^{D} \log c_d^2 + \mathcal{H}(\phi) +$$

$$+ \mathbb{E}_{\phi(z)} \left[\log \mathcal{N}(c \circ z + \mu \mid 0, \Lambda)\right]$$

ARD logistic regression

$$\mathcal{L}(\mu, C, \Lambda) \propto \mathbb{E}_{\phi(z)} \left[\log \tilde{g}(c \circ z + \mu)\right] +$$

$$+\frac{1}{2}\sum_{d=1}^{D}\left(\log c_d^2 - \log \lambda_d^2 - \frac{c_d^2 + \mu_d^2}{\lambda_d^2}\right) \to \max_{C,\mu,\Lambda}$$

$$\Updownarrow$$

$$\mathcal{L}(\mu, C) = \mathbb{E}_{\phi(z)} \left[\log \tilde{g}(c \circ z + \mu)\right] + \frac{1}{2}\sum_{d=1}^{D} \log \frac{c_d^2}{c_d^2 + \mu_d^2} \to \max_{C,\mu}$$

$$\lambda_d^2 = c_d^2 + \mu_d^2$$

ARD logistic regression

Datasets:

Data set	#Train	#Test	D	#Nonzeros
a9a	32,561	16,281	123	451,592
rcv1	20,242	677,399	47,236	49,556,258
Epsilon	400,000	100,000	2000	800,000,000

Test error rates for DSVI-ARD and l_1-logistic regression:

Data set	DSVI ARD	L1-LR
a9a	0.1507	0.1500
rcv1	0.0414	0.0420
Epsilon	0.1014	0.1011

Gaussian Process regression

$$y_n = f(x_n) + \varepsilon_n$$

$$\varepsilon_n \sim \mathcal{N}(\varepsilon_n \,|\, 0, \sigma^2), \qquad f(x) \sim \mathcal{GP}(0, k(x, x'; \theta))$$

$$k(x, x'; \theta) = \sigma_f^2 \exp\left\{ -\frac{1}{2} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\lambda_d^2} \right\}$$

$$g(\Theta) = \mathcal{N}(y \,|\, 0, K + \sigma^2 I)\mathcal{N}(\Theta \,|\, 0, 10I)$$

$$\Theta = \left[\log \lambda_1^2, \ldots, \log \lambda_D^2, \log \sigma_f^2, \log \sigma^2\right]^T$$

Full scale C, 20000 iterations of DSVI, no data subsampling.

Gaussian Process regression

Negative log-predictive densities (nlpd) and standartised mean square errors (smse) in test data:

Data set		ML-II	DSVI	MCMC
`Boston`	(smse)	0.0743	0.0709	0.0699
	(nlpd)	0.1783	0.1425	0.1317
`Bodyfat`	(smse)	0.1992	0.0726	0.0726
	(nlpd)	-0.1284	-2.0750	-2.0746
`Pendulum`	(smse)	0.2727	0.2807	0.2801
	(nlpd)	0.4537	0.4465	0.4462
Average training time		40 sec	30 min	20 h

Conclusion

Doubly Stochastic Variational Inference algorithm...

- ... can be applied to a huge variety of models
- ... is simple and efficient
- ... has theoretical convergence guarantees

Related methods

Black Box Variational Inference:

- Arbitry variational distribution $q(\Theta|\alpha)$
- Very high variance of gradients (variance reduction tricks are needed)
- Usually very slow

Rajesh Ranganath, Sean Gerrish, and David M Blei.
Black Box Variational Inference.
Aistats, 33:814–822, 2014.

Related methods

Autoencoding Variational Bayes:

- A wide approximation family due to the reparametrization trick in a general case
- No convergence guarantees for arbitry reparametrization
- Iterations are slightly more difficult

Rajesh Ranganath, Sean Gerrish, and David M Blei.
Black Box Variational Inference.
Aistats, 33:814–822, 2014.

Related methods

Variational Inference with Normalizing Flows:

- A very interesting variational distribution that can approximate virtually any function
- Iterations are slower (quadratic complexity), but the precision is usually much higher

Rajesh Ranganath, Sean Gerrish, and David M Blei.
Black Box Variational Inference.
Aistats, 33:814–822, 2014.

Related methods

Any other modern paper with words "Variational Inverence" or "Variational Bayes" in title

- The main difference is in the approximating family
- Most of these methods use reparametrization to sample from variational distribution