Variational Approximation

Or Being a Bayesian in a world with too much data

Jacob Carey

December 3, 2017

Motivation


GIVE ME ALL THE DATA!
imgflip.com


GIVE ME ALL THE DATA?
imgflip.com

Bayesians vs Frequentists

- Frequentist methods are typically faster than their Bayesian counterparts (e.g. logistic regression)

Bayesians vs Frequentists

- Frequentist methods are typically faster than their Bayesian counterparts (e.g. logistic regression)
- Computationally - why is this?

Bayesians vs Frequentists

- Frequentist methods are typically faster than their Bayesian counterparts (e.g. logistic regression)
- Computationally - why is this?
- Frequentist MLEs are typically solved using (fast) optimization

Bayesians vs Frequentists

- Frequentist methods are typically faster than their Bayesian counterparts (e.g. logistic regression)
- Computationally - why is this?
- Frequentist MLEs are typically solved using (fast) optimization
- Bayesian posteriors are typically approximated using (slow) integration

Speeding up integration

- Alternatives to the Gibbs/MH samplers
- Examples
  - Hamiltonian and extensions
  - Collapsed Gibbs Sampling
  - Others
- Still slow in comparison

Theory of Variational Approximation

Approximate the posterior using optimization

- Idea: find a density $q^*(z)$ such that

$$q^*(z) = \arg\min_{q(z)} H(q(z)||p(z|x))$$

- H is some "distance" between a density $q(z)$ and the true posterior $p(z|x)$.
- Typical distance used is the *Kullback-Leibler Divergence*

$$\mathrm{KL}(q(z)||p(z|x)) = \int q(z) \log\left\{\frac{q(z)}{p(z|x)}\right\} dz$$

- D_{KL} is *asymmetric*, greater than or equal to 0 for all densities q and equal iff $q(z) = p(z|x)$ almost everywhere.

Derivation

$$
\begin{aligned}
\log p(x) &= \int q(z) \log p(x) dz \\
&= \int q(z) \log \left\{ \frac{p(z,x)/q(z)}{p(z|x)/q(z)} \right\} dz \\
&= \int q(z) \log \left\{ \frac{p(z,x)}{q(z)} \right\} dz + \int q(z) \log \left\{ \frac{q(z)}{p(z|x)} \right\} dz \\
&\geq \int q(z) \log \left\{ \frac{p(z,x)}{q(z)} \right\} dz
\end{aligned}
$$

ELBO

- We call $\int q(z) \log \left\{ \frac{p(z,x)}{q(z)} \right\} = \mathbb{E}[\log p(z,x)] - \mathbb{E}[\log q(z)]$ the *evidence lower bound* or *ELBO*.
- From the derivation in the previous slide, it is apparent that maximizing the ELBO is equivalent to minimizing the $\mathrm{D_{KL}}$ between $q(z)$ and the posterior.

Variational Approximation

- Finding the density $q(z)$ which maximizes the ELBO is called *Variational Approximation*.
- Typically, we limit the candidate densities $q(z)$ to a family $\mathscr{Q}$ to make this optimization more analytically tractable.
- The common assumption made is that $q(z)$ factorizes into $\prod_{i=1}^{M} q_i(z_i)$ for some partition of z.
- This assumption is called the *mean field approximation*.
- Using Lagrange multipliers (in conjuction with the independence assumption), we can derive an algorithm

Coordinate Ascent Variational Inference

- The derived algorithm - CAVI - is as follows

Coordinate Ascent Variational Inference

- The derived algorithm - CAVI - is as follows
- Iterate over each variable/partition

Coordinate Ascent Variational Inference

- The derived algorithm - CAVI - is as follows
- Iterate over each variable/partition
- Update j-th variational density as follows

Coordinate Ascent Variational Inference

- The derived algorithm - CAVI - is as follows
- Iterate over each variable/partition
- Update j-th variational density as follows
- $q^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j, z_{-j}, x)\right]\right\}$

Coordinate Ascent Variational Inference

- The derived algorithm - CAVI - is as follows
- Iterate over each variable/partition
- Update j-th variational density as follows
- $q^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j, z_{-j}, x)\right]\right\}$
- Stop when the change in the ELBO is "negligible"

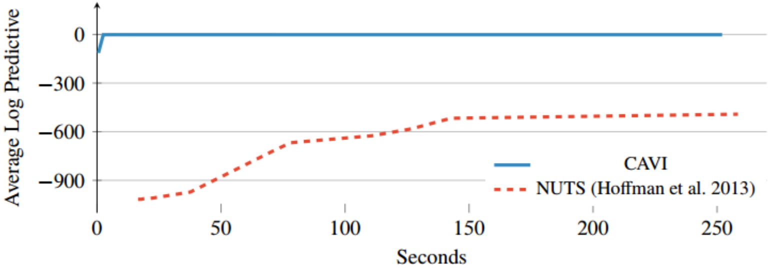
Pros

- Fast: some complicated models (even on small-moderate data) may converge prohibitively slowly with MCMC. MCMC does not perform well on large data
- Convergence is much easier to diagnose than MCMC

Cons

- Traditional VA underestimates the posterior variance (i.e. overestimates our "confidence" in the posterior point estimate)
- Only guaranteed to find a local optimum
- Much more difficult to derive updates than for many MCMC methods
- Independence assumption may not be a good one in the case of MFA

Comparison



Gaussian mixture models fit to ten thousand images

Blei et al 2016

Examples

Normal with conjugate priors

Model:

$$X_i|\mu, \tau \sim \text{Normal}(\mu, \tau)$$

Priors:

$$\mu \sim \text{Normal}(\mu_0, \tau_\mu)$$

$$\tau \sim \text{Gamma}(A_0, B_0)$$

Variational Densities:

$$q(\tau; A_1, B_1)$$

$$q(\mu; m, s^2)$$

Update for Precision

$$
\begin{aligned}
q^*_\tau(\tau) &\propto \exp \mathbb{E}[\log p(\mu) + \log p(\tau) + \log \prod p(x_i|\mu,\tau); m, s^2] \\
&\propto \exp\{(A_0 - 1)\log\tau - B_0\tau + \\
&\quad \sum \mathbb{E}[\frac{1}{2}\log\tau - \tau(x_i-\mu)^2/2; m, s^2]\} \\
&\propto \exp\{(A_0 + \frac{n}{2} - 1)\log\tau - B_0\tau + \\
&\quad \sum \mathbb{E}[-\tau(x_i^2 - 2\mu x_i + \mu^2)/2; m, s^2]\} \\
&\propto \exp\{(A_0 + \frac{n}{2} - 1)\log\tau - B_0\tau + \\
&\quad \sum(-\tau x_i^2/2 - \tau m x_i + \tau(m^2+s^2)/2))\} \\
&\implies A_1 = A_0 + \frac{n}{2}; B_1 = B_0 + \frac{1}{2}\sum x_i^2 - n\bar{x}m + \frac{n}{2}(s^2+m^2)
\end{aligned}
$$

Update for Mean

$$
\begin{aligned}
q^*_\mu(\mu) &\propto \exp\{\mathbb{E}\log p(\mu) + \sum \mathbb{E}[\log p(x_i|\mu,\tau); A_1, B_1]\} \\
&= \exp\{\frac{1}{2}\log\tau_\mu - \frac{\tau_\mu}{2}(\mu-\mu_\mu)^2 + \\
&\quad \sum \mathbb{E}[\frac{1}{2}\log\tau - \frac{\tau}{2}(x_i-\mu)^2; A_1, B_1]\} \\
&\propto \exp\{(\mu^2 - 2\mu\mu_\mu)\tau_\mu/2 + \sum \mathbb{E}[-(\mu^2 - 2\mu x_i)\frac{\tau}{2}; A_1, B_1]\} \\
&= \exp\{-\frac{1}{2}\tau_\mu\mu^2 + \tau_\mu\mu_\mu\mu + \sum(-\frac{A}{2B}\mu^2 + \frac{A}{B}\mu x_i)\} \\
&= \exp\{-\frac{1}{2}((n\frac{A}{B} + \tau_\mu)\mu^2 - 2(n\bar{x}\frac{A}{B} + \tau_\mu\mu_\mu)\mu)\} \\
&\implies m = \frac{n\bar{x}\frac{A}{B} + \tau_m u\mu_\mu}{n\frac{A}{B} + \tau_\mu}; s^2 = \frac{1}{n\frac{A}{B} + \tau_\mu}
\end{aligned}
$$

ELBO

$$
\begin{aligned}
\mathrm{ELBO}(m, s^2, A_1, B_1) &= \mathbb{E}[\log p(\mu, \tau, x_1, ..., x_n)] - \mathbb{E}[\log q(\mu, \tau)] \\
&= \mathbb{E}[\log p(\mu)] + \mathbb{E}[\log p(\tau)] + \\
&\quad \sum \mathbb{E}[\log p(x_i|\mu, \tau)] - \\
&\quad \mathbb{E}[\log q(\mu)] - \mathbb{E}[\log q(\tau)]
\end{aligned}
$$

ELBO (Cont)

$$
\begin{aligned}
\mathbb{E}[\log p(\mu)] &= \mathbb{E}[\frac{1}{2}\log\frac{\tau_\mu}{2\pi} - \frac{\tau}{2}(\mu^2 - 2\mu\mu_\mu + \mu_\mu^2)] \\
&= \frac{1}{2}\log\frac{\tau_\mu}{2\pi} - \frac{\tau}{2}((m^2 + s^2) - 2m\mu_\mu + \mu_\mu^2)] \\
\mathbb{E}[\log q(\mu)] &= \mathbb{E}[-\frac{1}{2}\log 2s^2\pi - \frac{1}{2s^2}(\mu^2 - 2\mu m + m^2)] \\
&= -\frac{1}{2}\log 2s^2\pi - \frac{1}{2s^2}((m^2 + s^2) - 2m^2 + m^2) \\
&= -\frac{1}{2} - \frac{1}{2}\log 2s^2\pi
\end{aligned}
$$

ELBO (Cont)

$$
\begin{aligned}
\log p(x_i|\mu,\tau) &= \frac{1}{2}\log\frac{\tau}{2\pi} - \frac{\tau}{2}(x_i^2 - 2x_i\mu + \mu^2) \\
\mathbb{E}[\log p(x_i|\mu,\tau)] &= \mathbb{E}[\frac{1}{2}\log\frac{\tau}{2\pi} - \frac{\tau}{2}x_i^2 - \tau x_i\mu + \frac{\tau}{2}\mu^2] \\
&= \frac{1}{2}\,\mathbb{E}[\log\tau] - \frac{1}{2}\log 2\pi - \frac{A_1}{2B_1}(x_i^2 - 2x_i m_m^2 + s^2)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{ELBO}(m, s^2, A_1, B_1) = \frac{1}{2} - \frac{n}{2}\log 2\pi + \frac{1}{2}\log s^2\tau_\mu - \\
\frac{\tau_\mu}{2}(s^2 + m^2 - 2m\mu_\mu + \mu_\mu^2)
\end{aligned}
$$

Black Box Variational Inference

Foundation

- Gradient Descent
- Automatic Differentiation

Gradient Descent

- Method of minimizing a function
- $\theta_{n+1} = \theta_n + \eta \nabla_\theta J(\theta)$

Questions?