Variational Inference

References

These notes are based on the following papers:

- Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. Journal of the American Statistical Association, 112:518,859-877.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. Machine Learning, 37, 183-233.

Introduction

<u>MCMC</u>
- Theoretical guarantees (asymptotically) of sampling rom the target density
- Computationally intensive but conceptually simple
- Handles multi-modal posterior distributions

<u>Variational Inference</u>
- No theoretical guarantees
- Good for big data and complex models
- Faster than MCMC but requires derivation of variational updates
- Can have problems with multi-modal posteriors

3

Introduction

- Variational methods based on calculus of variations
- Complex problem turned to a simpler one by decoupling degrees of freedom in the original problem
- Decoupling done by extending the original problem with additional variational parameters

4

Intuition

We first develop some intuition about variational methods using a simple example.

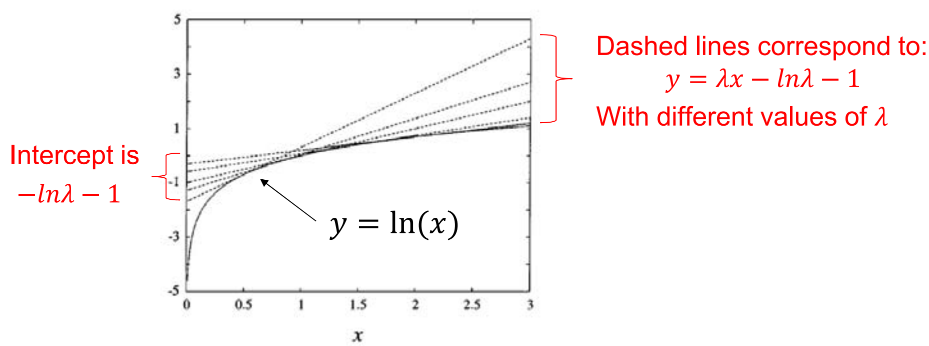
Write the log function as:

$$\ln(x) = \min_{\lambda}\{\lambda x - ln\lambda - 1\}$$

Note:

- λ is the variational parameter
- For each value of x, we need to compute the minimization of λ.

Intuition



- Varying λ produces a series of upper bounds:
  $$\ln(x) \le \lambda x - ln\lambda - 1$$
- Minimizing λ produces the exact value for $\ln(x)$
- Note: $\ln(x)$ is a concave function

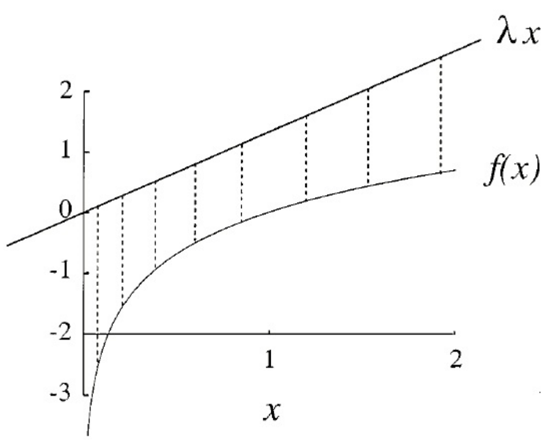
Intuition

Why did we write $\ln(x) = \min_{\lambda}\{\lambda x - ln\lambda - 1\}$?

- Comes from convex duality: a concave function $f(x)$ can be represented by a dual function as
$$f(x) = \min_{\lambda}\{\lambda^T x - f^*(\lambda)\}$$
  Where
$$f^*(\lambda) = \min_{x}\{\lambda^T x - f(x)\}$$
- Applies to convex functions as well but you get a lower bound

7

Intuition



$f(x) = \min_{\lambda}\{\lambda^T x - f^*(\lambda)\}$

8

Variational Inference

- Let $\boldsymbol{X}$ be a set of observed variables (e.g. evidence variables)
- Let $\boldsymbol{Z}$ be a set of latent variables
- Given inference query $P(\boldsymbol{Z}|\boldsymbol{X}) = \frac{P(\boldsymbol{X},\boldsymbol{Z})}{P(\boldsymbol{X})}$
- Need to compute $P(\boldsymbol{X}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{X}, \boldsymbol{Z})$ if $\boldsymbol{Z}$ is discrete or $\int P(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}$ if continuous
- The denominator is typically very expensive to compute

9

The ELBO

- Goal: choose a density $q(\boldsymbol{z}) \epsilon Q$ which is the closest approximation to $p(\boldsymbol{z}|\boldsymbol{x})$
- Here Q is a family of densities over the latent variables
- Need to solve the following optimization problem:

$$q^*(\boldsymbol{z}) = \underset{q(\boldsymbol{z})\epsilon Q}{\operatorname{argmin}} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

10

The ELBO

$KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$

$= \int q(\mathbf{z}) log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z} = \int q(\mathbf{z}) log \frac{q(\mathbf{z})p(\mathbf{x})}{p(\mathbf{x},\mathbf{z})} d\mathbf{z}$

$= \int q(\mathbf{z})[logq(\mathbf{z}) + logp(\mathbf{x}) - logp(\mathbf{x},\mathbf{z})]d\mathbf{z}$

$= \int [q(\mathbf{z})logq(\mathbf{z}) + q(\mathbf{z})logp(\mathbf{x}) - q(\mathbf{z})logp(\mathbf{x},\mathbf{z})]d\mathbf{z}$

$= \int q(\mathbf{z}) \log q(\mathbf{z})\, d\mathbf{z} + \int q(\mathbf{z})logp(\mathbf{x})d\mathbf{z} - \int q(\mathbf{z})logp(\mathbf{x},\mathbf{z})d\mathbf{z}$

$= E_{q(\mathbf{z})}[logq(\mathbf{z})] + E_{q(\mathbf{z})}\underbrace{[logp(\mathbf{x})]} - E_{q(\mathbf{z})}[logp(\mathbf{x},\mathbf{z})]$

Doesn't involve $q(\mathbf{z})$ so it can be taken out of the expectation

11

The ELBO

$KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$

$= E_{q(\mathbf{z})}[logq(\mathbf{z})] - E_{q(\mathbf{z})}[logp(\mathbf{x},\mathbf{z})] + logp(\mathbf{x})$

Remember that this is very hard to compute because $p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{z})d\mathbf{z}$

Rewrite as:

$logp(\mathbf{x}) = KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) + E_{q(\mathbf{z})}[logp(\mathbf{x},\mathbf{z})] - E_{q(\mathbf{z})}[logq(\mathbf{z})]$

$= KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) + ELBO(q)$

Where

$$ELBO(q) = E_{q(\mathbf{z})}[logp(\mathbf{x},\mathbf{z})] - E_{q(\mathbf{z})}[logq(\mathbf{z})]$$

12

The ELBO

- Because KL divergence is ≥ 0

$$logp(\boldsymbol{x}) = KL(q(\boldsymbol{z})||p(\boldsymbol{z})) + ELBO(q)$$

$$\Rightarrow logp(\boldsymbol{x}) \geq ELBO(q)$$

- $p(\boldsymbol{x})$ is the probability of the evidence, hence this is an evidence lower bound (ELBO)
- Instead of minimizing the KL divergence, we maximize the ELBO

$$ELBO(q) = E_{q(\boldsymbol{z})}[logp(\boldsymbol{x}, \boldsymbol{z})] - E_{q(\boldsymbol{z})}[logq(\boldsymbol{z})]$$

13

The ELBO

Another point about $ELBO(q)$

$$= E_{q(\boldsymbol{z})}[logp(\boldsymbol{x}, \boldsymbol{z})] - E_{q(\boldsymbol{z})}[logq(\boldsymbol{z})]$$

$$= E_{q(\boldsymbol{z})}[logp(\boldsymbol{x}|\boldsymbol{z})] + E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[logq(\boldsymbol{z})]$$

$$= E_{q(\boldsymbol{z})}[logp(\boldsymbol{x}|\boldsymbol{z})] + \int q(\boldsymbol{z})logp(\boldsymbol{z})d\boldsymbol{z} - \int q(\boldsymbol{z}) \log q(\boldsymbol{z})\, d\boldsymbol{z}$$

$$= E_{q(\boldsymbol{z})}[logp(\boldsymbol{x}|\boldsymbol{z})] + \int q(\boldsymbol{z})log\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}d\boldsymbol{z}$$

$$= E_{q(\boldsymbol{z})}[\underbrace{logp(\boldsymbol{x}|\boldsymbol{z})}] + \underbrace{KL(q(\boldsymbol{z})||p(\boldsymbol{z}))}$$

This is an expected likelihood. Places mass of $q(\boldsymbol{z})$ on configurations of the latent variables $\boldsymbol{z}$ that explain the observed data $\boldsymbol{x}$.

This makes $q(z)$ resemble the prior $p(z)$

14

Mean Field

In the mean-field variational family, the latent variables are:

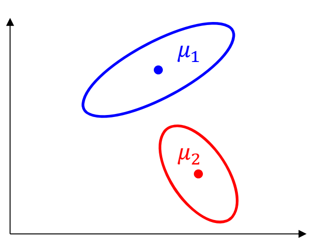
- Mutually independent
- Each has its own factor (and parameters) in the variational family

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j)$$

15

Bayesian Mixture of Gaussians

- A Gaussian mixture model assumes there are K Gaussians that generate the data, each with its own mean μ_k and variance σ_k



- We can make this a Bayesian model by putting a prior on the means of the K Gaussians

16

Bayesian Mixture of Gaussians

The Generative model:

- $\boldsymbol{\mu}_k \sim N(\mathbf{0}, \boldsymbol{\sigma}^2)$ for $k = 1, \ldots, K$
- $c_i \sim Categorical(\frac{1}{K}, \ldots, \frac{1}{k})$ for $i = 1, \ldots, n$
- $x_i | c_i, \boldsymbol{\mu} \sim N(c_i^T \boldsymbol{\mu}, 1)$ for $i = 1, \ldots, n$

The joint density is:

$$p(\boldsymbol{\mu}, c, x) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

17

Bayesian Mixture of Gaussians

Prior for the mean of each mixture component

σ is a given hyperparameter

The Generative model:

- $\mu_k \sim N(\mathbf{0}, \boldsymbol{\sigma}^2)$ for $k = 1, \ldots, K$
- $c_i \sim Categorical(\frac{1}{K}, \ldots, \frac{1}{k})$ for $i = 1, \ldots, n$
- $x_i | c_i, \boldsymbol{\mu} \sim N(c_i^T \boldsymbol{\mu}, 1)$ for $i = 1, \ldots, n$

Cluster assignment that takes values $1, \ldots, K$. Encoded as an indicator K-vector, with 0s everywhere except for a 1 in the position corresponding to the cluster c_i belongs to.

$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}$

The joint density is:

$$p(\boldsymbol{\mu}, c, x) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

18

Bayesian Mixture of Gaussians

- Computing the evidence requires:

$$p(x) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i | c_i, \boldsymbol{\mu})\, d\boldsymbol{\mu}$$

- This K-dimensional integral takes $O(K^n)$ time to compute.
- Variational Inference to the rescue!

Bayesian Mixture of Gaussians

Joint density:

$$p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \boldsymbol{\mu})$$

Mean-field variational family:

$$q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i)$$

ELBO:

$$E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logp(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{x})] - E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logq(\boldsymbol{\mu}, \boldsymbol{c})]$$

Bayesian Mixture of Gaussians

$$
\begin{aligned}
&ELBO(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\varphi}) \\
&= E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logp(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{x})] - E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logq(\boldsymbol{\mu}, \boldsymbol{c})] \\
&= E_{q(\boldsymbol{\mu},\boldsymbol{c})}\left[log\left(p(\boldsymbol{\mu})\prod_{i=1}^{n} p(c_i)p(x_i|c_i, \boldsymbol{\mu})\right)\right] \\
&\quad -E_{q(\boldsymbol{\mu},\boldsymbol{c})}\left[log\left(\textstyle\prod_{k=1}^{K} q(\mu_k; m_k, s_k^2)\prod_{i=1}^{n} q(c_i; \varphi_i)\right)\right] \\
&= \sum_{k=1}^{K} E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logp(\mu_k); m_k, s_k^2] \\
&+ \sum_{i=1}^{n}\left(E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logp(c_i); \varphi_i] + E_{q(\boldsymbol{\mu},\boldsymbol{c})}[\log p(x_i; c_i, \boldsymbol{\mu}) ; \varphi_i, \boldsymbol{m}, \boldsymbol{s}^2]\right) \\
&- \sum_{i=1}^{n} E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logq(c_i; \varphi_i)] - \sum_{k=1}^{K} E_{q(\boldsymbol{\mu},\boldsymbol{c})}[logq(\mu_k; m_k, s_k^2)]
\end{aligned}
$$

21

Bayesian Mixture of Gaussians

- With the ELBO, we now need to optimize the variational parameters
- One way to do this is coordinate ascent variational inference (CAVI) (Bishop 2006)
- Works by optimizing each parameter while keeping the others fixed
- Need to come up with updates for φ_{ik}, m_k, s_k
- Done iteratively until ELBO converges

22

Bayesian Mixture of Gaussians

For CAVI:

- Uses the complete conditional of z_j i.e. $p(z_j|\mathbf{z}_{-j},\boldsymbol{x})$
- Optimization uses the following:

$$q_j^*(z_j) \propto \exp\{E_{-j}[log p(z_j,\mathbf{z}_{-j},\boldsymbol{x})]\}$$

This expectation is over all the other variational factors being fixed i.e. $\prod_{l\neq j} q_l(z_l)$

- We won't go through the derivation. See (Bishop 2006) for details

23

Bayesian Mixture of Gaussians

For Bayesian Mixture of Gaussians:

1. Compute update for mixture assignments.
2. Compute update for mixture component means and variances.

24

Bayesian Mixture of Gaussians

1. Computing update for φ_{ik}

$$q^*(c_i; \varphi_i) \propto \exp\{\underbrace{logp(c_i)} + \underbrace{E[logP(x_i|c_i, \boldsymbol{\mu}); \boldsymbol{m}, \boldsymbol{s}^2]}\}$$

$logp(c_i) = -1/K$

$\varphi_{ik} \propto \exp\{E[\mu_k; m_k, s_k^2]x_i - E[\mu_k^2; m_k, s_k^2]/2\}$
(derivation left as an exercise)

Note: Expectation will be over $\prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{j \neq i} q(c_j; \varphi_j)$

25

Bayesian Mixture of Gaussians

2. Computing update for m_k, s_k

$$q(\mu_k; m_k, s_k^2) \propto exp\left\{logp(\mu_k) + \sum_{i=1}^{n} E[logp(x_i|c_i, \boldsymbol{\mu}); \varphi_i, \boldsymbol{m}_{-k}, \boldsymbol{s}^2_{-k}]\right\}$$

Note: Expectation will be over $\prod_{l \neq k} q(\mu_l; m_l, s_l^2) \prod_{i=1}^{n} q(c_i; \varphi_i)$

This leads to update equations: (derivation left as an exercise)

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}$$

$$s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}$$

26

Concluding remarks

- It takes some work to derive variational inference equations
- Generic variational updates have been derived for special cases e.g. when complete conditional is in the exponential family

27