Variational Inference

Di Zhao

June 2018

1 Introduction

Variational inference approximates probability distributions through optimization. Usually, it finds wide applicability in approximating difficult-to-compute probability distributions, a problem which is especially important in Beyesian Inference to estimate posteriror distributions.

The idea behind variational inference is to posit a family of distributions and then to find the member of that family which is close to the target, the closeness is measured by KL divergence.
Suppose a posterior $p(z|x)$:

$$p(z|x) = \frac{p(x,z)}{p(x)}$$

Since we know the evidence lower bound $\mathcal{L}$:

$$\mathcal{L} = ln(p(x)) - KL$$

$$\mathcal{L} = E_q[log(p(z,x))] - E_q[log(q(z)]$$

$$\mathcal{L} = E_q[log(\frac{p(z,x)}{q(z)})] = E_q[log(\frac{p(x|z)p(z)}{q(z)})] = E_q[log(p(x|z))] + E_q[log(\frac{p(z)}{q(z)})]$$

$$\mathcal{L} = E_{q(z)}[(log(p(x|z)))] - KL(q(z)||p(z))$$

2 Mean Field Approximation

The variational objective function is specified with ELBO $\mathcal{L}$, then we need to specify the variational family of distributions from which we pick the approximate variational distribution. AS might be intuited that the complexity of the family of distributions from which we pick the approximate distribution determines the complexity of the optimization. The more flexibility of distributions, the closer the approximation and the harder the optimizations.

A common family of distributions to pick is the Mean-field variational family, in which the latent variables are mutually independent and each governed by a

distinct factor in the variational distribution. A generic member of the mean-field variational family is given by the below equation:

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

3 Optimization Algorithm: coordinate ascent mean-field variational inference

Algorithm 1 coordinate ascent mean-field variational inference

procedure
  Inputs $\leftarrow p(x, z)$ *with data set* x
  Outputs $\leftarrow q(z) = \prod_j q_j(z_j)$
  Initialize $\leftarrow q_j(z_j)$
  while *ELBO has not converged (or z have not converged)* **do**
    for all j **do**:
      $q_j \propto exp(E_{-j}[log(p(z_j|z_{-j}, x))])$
    Compute ELBO

4 Example:Variational Bayes for a univariate Gaussian

This example shows how to use variational bayes to infer the posterior over the parameters for a 1d Gaussian,$p(\mu, \lambda|\mathcal{D})$, where $\lambda = \frac{1}{\sigma^2}$ is the precision. The conjugate prior of the form

$$p(\mu, \lambda) = N(\mu|\mu_0, (k_0\lambda)^{-1})Ga(\lambda|a_0, b_0)$$

And we will use an approimate factored posterior of the form

$$q(\mu, \lambda) = q_\mu(\mu)q_\lambda(\lambda)$$

We do not need to specify the forms for the distributions $q_{[}\mu]$ and q_λ;the optimal forms will "fall out" automatically during the derivation (they turn out to be Gaussian and Gamma respectievly).

4.1 Target distribution

The unnormalized log posterior has the form

$$log(\hat{p}(\mu, \lambda)) = log(p(\mu, \lambda, D)) = log(p(D|\mu, \lambda)) + log(p(\mu|\lambda)) + log(p(\lambda)) =$$

$$\frac{N}{2}log(\lambda) - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{k_0\lambda}{2}(\mu - \mu_0)^2 + \frac{1}{2}log(k_0\lambda) + (a_0 - 1)log(\lambda) - b_0\lambda + const$$

4.2 Updating $q_\mu(\mu)$

The optimal form for $q_\mu(\mu)$ is obtained by averaging over λ

$$log(q_\mu(\mu)) = E_{q_\lambda}[log(p(D|\mu,\lambda)) + log(p(\mu|\lambda))] + const$$

$$= -\frac{E_{q_\lambda}[\lambda]}{2}[k_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2] + const$$

By completing the square one can show that $q_\mu = N(\mu|\mu_N, {}_{k_N^{-1}})$, where

$$\mu_N = \frac{k_0\mu_0 + N\bar{x}}{k_0+N}, k_N = (k_0+N)E_{q_\lambda}[\lambda]$$

4.3 Updating $\mathbf{q}_\lambda(\lambda)$

The optimal form for $q_\lambda(\lambda)$ is given by

$$log(q_\lambda(\lambda)) = E_{q_\mu}[log(p(D|\mu,\lambda)) + log(p(\mu|\lambda)) + log(p(\lambda))] + const$$

$$= (a_0-1)log(\lambda) - b_0\lambda + + \frac{1}{2}log(\lambda) + \frac{N}{2}log(\lambda)$$

$$-\frac{\lambda}{2}E_{q_\mu}[k_0(\mu - mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2] + const$$

We recognize this as the log of a Gamma distribution, hence $q_\lambda(\lambda) = Ga(\lambda|a_N, b_N)$, where

$$a_N = a_0 + \frac{N+1}{2}$$

$$b_N = b_0 + \frac{1}{2}E_{q_\mu}[k_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2]$$

4.4 Computing the expectations

To implement the updates, we have to specify how to compute the various expectations. Since $q(\mu) = N(\mu|\mu_N, k_N^{-1})$, we have

$$E_{q(\mu)}[\mu] = \mu_N$$

$$E_{q(\mu)}[\mu^2] = \frac{1}{k_N} + \mu_N^2$$

Since $q(\lambda) = Ga(\lambda|a_N, b_N)$, we have

$$E_{q(\lambda)}[\lambda] = \frac{a_N}{b_N}$$

We can now give explicit forms for the update equations. For $q(\mu)$ we have

$$\mu_N = \frac{k_0\mu_0 + N\bar{x}}{k_0 + N}$$

$$k_N = (k_0 + N)\frac{a_N}{b_N}$$

and for $q(\lambda)$ we have

$$a_N = a_0 + \frac{N+1}{2}$$

$$b_N = b_0 + k_0(E[\mu^2] + \mu_0^2 - 2E[\mu]\mu_0) + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + E[\mu^2] - 2E[\mu]x_i)$$