Variational inference [1]

Lesson No. 10

Gustavo de J. Merli - 262948

1 Introduction

This chapter approaches a more general class of deterministic approximate inference algorithms based on variational inference. The basic idea is to pick an approximation $q(x)$ to the distribution from some tractable family, and then to try to make this approximation as close as possible to the true posterior, $p^*(x) = p(x|D)$. This reduces inference to an optimization problem. By relaxing the constraints and/or approximating the objective, it is possible to trade accuracy for speed. The bottom line is that variational inference often gives us the speed benefits of MAP estimation but the statistical benefits of the Bayesian approach.

2 Variational inference

Suppose $p^*(x)$ is a true but intractable distribution and $q(x)$ is some approximation, chosen from some tractable family, such as a multivariate Gaussian or a factored distribution. Assume q has some free parameters which can be optimized so as to make q "similar to" p^*.

Defining an objective function as follows:

$$J(q) = \sum_x q(x) log \frac{q(x)}{p^*(x)} - log Z \tag{1}$$

$$J(q) = \mathbb{KL}(q||p^*) - log Z \tag{2}$$

where $Z = p(D)$ is a constant.

2.1 Forward or reverse KL?

Since the KL divergence is not symmetric in its arguments, minimizing $\mathbb{KL}(q||p)$ wrt q will give different behavior than minimizing $\mathbb{KL}(p||q)$.

Consider the reverse KL, $\mathbb{KL}(q||p)$. By definition

$$\mathbb{KL}(q||p) = \sum_x q(x) ln \frac{q(x)}{p(x)} \tag{3}$$

This is infinite if $p(x) = 0$ and $q(x) > 0$. Thus if $p(x) = 0$ it needs to ensure $q(x) = 0$. This way, the reverse KL is **zero forcing** for q. Hence q will tipically under-estimate the support for p.

Consider the forward KL:

$$\mathbb{KL}(p||q) = \sum_x p(x) ln \frac{p(x)}{q(x)} \tag{4}$$

This is infinite if $q(x) = 0$ and $p(x) > 0$. So if $p(x) > 0$, it needs to ensure $q(x) > 0$. This way, the forwards KL is **zero avoiding** for q. Hence q will tipically over-estimate the support of p.

3 The mean field method

One of the most popular forms of variational inference is called the mean field approximation. Assume the posterior is a fully factorized approximation of the form

$$q(x) = \prod_i q_i(x_i) \tag{5}$$

The goal is to solve this optimization problem

$$\min_{q_1,\ldots,q_D} \mathbb{KL}(q||p) \tag{6}$$

3.1 Derivation of the mean field update equations

Recall that the goal of variational inference is to minimize the upper bound $J(q) \geq -log\, p(D)$. Equivalently, it is possible to try to maximize the lower bound

$$L(q) = -J(q) = \sum_i q(x) log \frac{\widetilde{p}(x)}{q(x)} \leq log\, p(D) \tag{7}$$

Where $\widetilde{p}(x) = p(x, D) = p^*(x)p(D)$. Doing this one term at a time having.

$$L(q_j) = -\mathbb{KL}(q_j||f_j) \tag{8}$$

where

$$log\, f_j(x_j) = \sum_{x_{-j}} \prod_{i \neq j} q_i(x_i) log\, \widetilde{p}(x) = \mathbb{E}_{-q_j}[log\, \widetilde{p}(x)] \tag{9}$$

It is possible to maximize L by minimizing KL, which is done by setting $q_j = f_j$, as follows

$$log\, q_j(x_j) = \mathbb{E}_{q_j}[log\, \widetilde{p}(x)] + const \tag{10}$$

4 Structured mean field

Assuming that all the variables are independent in the posterior is a very strong assumption that can lead to poor results. Sometimes it is possible to exploit tractable substructure in our problemsuch as the **structured mean field** approach. The approach is the same as before, except this one groups sets of variables together, to update them simultaneously. As long as it is possible to perform efficient inference in each q_i, the method is tractable overall.

5 Variational Bayes

So far been concentrating on inferring latent variables z_i assuming the parameters θ of the model are known. Now suppose wanting to infer the parameters themselves. If made a fully factorized (i.e., mean field) approximation, $p(\theta|D) \approx \prod_k q(\theta_k)$, this is a method known as **variational Bayes** or **VB**.

5.1 VB for a univariate Gaussian

Consider how to apply VB to infer the posterior over the parameters for a 1d Gaussian, $p(\mu, \lambda|D)$, where $\lambda = 1/\sigma^2$ is the precision. For convenience, use a conjugate prior of the form.

$$p(\mu, \lambda) = \mathcal{N}(\mu, \mu_0, (\kappa_0\lambda)^{-1}) Ga(\lambda|a_0, b_0) \tag{11}$$

However, use an approximate factored posterior of the form

$$q(\mu, \lambda) = q_\mu(\mu) q_\lambda(\lambda) \tag{12}$$

The unnormalized log posterior has the form

$$log\tilde{p}(\mu,\lambda) = \frac{N}{2}log\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2 + \frac{1}{2}log(\kappa_0\lambda) + (a_0-1)log\lambda - b_0\lambda + const \quad (13)$$

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

$$log\,q_\mu(\mu) = -\frac{\mathbb{E}_{q\lambda}[\lambda]}{2}\left\{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2 + const\right\} \quad (14)$$

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

$$\mu_N = \frac{\kappa_0\mu_0 + N\bar{x}}{\kappa_0+N}, \kappa_N = (\kappa_0+N)\mathbb{E}_{q\lambda}[\lambda] \quad (15)$$

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

$$log\,q_\lambda(\lambda) = (a_0-1)log\lambda - b_0\lambda + \frac{1}{2}log\lambda + \frac{N}{2}log\lambda - \frac{\lambda}{2}\mathbb{E}_{q\mu}\left[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right] + const \quad (16)$$

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

$$a_N = a_0 + \frac{N+1}{2} \quad (17)$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_{q\mu}\left[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right] \quad (18)$$

Giving explicit forms for the update equations. For $q(\mu)$

$$\mu_N = \frac{\kappa_0\mu_0 + N\bar{x}}{\kappa_0+N} \quad (19)$$

$$\kappa_N = (\kappa_0+N)\frac{a_N}{b_N} \quad (20)$$

and for $q(\lambda)$

$$a_N = a_0 + \frac{N+1}{2} \quad (21)$$

$$b_N = b_0 + \kappa_0(\mathbb{E}[\mu^2] + \mu_0^2 - 2\mathbb{E}[\mu]\mu_0) + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + \mathbb{E}[\mu^2] - 2\mathbb{E}[\mu]x_i) \quad (22)$$

Noticing that μ_N and a_N are fixed constatns, and only κ_N and b_N are needed to be updated iteratively.

5.2 VB for linear regression

Deriving a VB algorithm for this model. Initially use the following prior:

$$p(w,\lambda,\alpha) = \mathcal{N}(w|0,(\lambda\alpha)^{-1}I)Ga(\lambda|a_0^\lambda,b_0^\lambda)Ga(\alpha|a_0^\alpha,b_0^\alpha) \quad (23)$$

The following factorized approximation to the posterior will be:

$$q(w,\alpha,\lambda) = q(w,\lambda)q(\alpha) \quad (24)$$

The optimal form for the posterior is

$$q(w,\alpha,\lambda) = \mathcal{N}(w|w_N, \lambda^{-1}V_N)Ga(\lambda|a_N^\lambda, b_N^\lambda)Ga(\alpha|a_N^\alpha, b_N^\alpha) \tag{25}$$

where

$$V_N^{-1} = \bar{A} + X^X \tag{26}$$

$$w_N = V_N X^T y \tag{27}$$

$$a_N^\lambda = a_0^\lambda + \frac{N}{2} \tag{28}$$

$$b_N^\lambda = b_0^\lambda + \frac{1}{2}(||y - Xw||^2 + w_N^T \bar{A} w_N) \tag{29}$$

$$a_N^\alpha = a_0^\alpha + \frac{D}{2} \tag{30}$$

$$b_N^\alpha = b_0^\alpha + \frac{1}{2}\left(\frac{a_N^\lambda}{b_N^\lambda} w_N^T w_N + tr(V_N)\right) \tag{31}$$

$$\bar{A} = \frac{a_N^\alpha}{b_N^\alpha} I \tag{32}$$

6 Variational Bayes EM

VB provides a way to be "more Bayesian", by modeling uncertainty in the parameters θ as well in the latent variables z_i, at a computational cost that is essentially the same as EM. This method is known as variational Bayes EM or VBEM. The basic idea is to use mean field, where the approximate posterior has the form

$$p(\theta, z_{1:N}|D) \approx q(\theta)q(z) = q(\theta)\prod_i q(z_i) \tag{33}$$

6.1 VBEM for mixtures of Gaussians

The likelihood function is the usual one for Gaussian mixture models:

$$p(z, X|\theta) = \prod_i \prod_k \pi_k^{z_{ik}} \mathcal{N}(x_i|\mu_k, \Lambda^{-1})^{z_{ik}} \tag{34}$$

where $z_{ik} = 1$ if data point i belongs to cluster k, and $z_{ik} = 0$ otherwise.
Assume the following factored conjugate prior

$$p(\theta) = Dir(\pi|\alpha_0)\prod_k \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})Wi(\Lambda_k|L_0), \nu_0) \tag{35}$$

where Λ_k is the precision matrix for cluster k. Using the standard VB approximation to the posterior:

$$p(\theta, z_{1:N}|D) \approx q(\theta)\prod_i q(z_i) \tag{36}$$

It is possible to reach the optimal form as follows:

$$q(z,\theta) = q(z|\theta) = \left[\prod_i Cat(z_i|r_i)\right]\left[Dir(\pi|\alpha)\prod_k \mathcal{N}(\mu_k|m_k, (\beta_k\Lambda_k)^{-1})Wi(\Lambda_k|L_k, \nu_k)\right] \tag{37}$$

where

$$\beta_k = \beta_0 + N_k \tag{38}$$

$$m_k = (\beta_0 m_0 + N_k \bar{x}_k)/\beta_k \tag{39}$$

$$L_k^{-1} = L_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \tag{40}$$

$$\nu_k = \nu_0 + N_k + 1 \tag{41}$$

$$\bar{x}_k = \frac{1}{N_k}\sum_i r_{ik} x_i \tag{42}$$

$$S_k = \frac{1}{N_k}\sum_i r_{ik}(x_i - \bar{k})(x_i - \bar{x}_k)^T \tag{43}$$

7 Local variational bounds

There is another kind of variational inference, where a simpler function replaces a specific term in the joint distribution, to simplify computation of the posterior. Such an approach is sometimes called a local variational approximation.

7.1 Motivating applications

Some examples of where this is useful.

7.1.1 Variational logistic regression

Consider the problem of how to approximate the parameter posterior for multiclass logistic regression model under a Gaussian prior. With a variational approach it can produce an accurate approximation to the posterior, since it has tunable parameters.

The likelihood can be written as follows:

$$p(y|X,w) = \prod_{i=1}^{N} exp[y_i^T \eta_i - lse(\eta_i)] \tag{44}$$

where $\eta_i = X_i w_i = [x_i^T w_1, ..., x_i^T w_M]$, where $M = C - 1$ (since $w_C = 0$ is set for identifiability), and where the log-sum-exp or lse functions is as fllows:

$$lse(\eta_i) = log\left(1 + \sum_{m=1}^{M} e^{\eta_{im}}\right) \tag{45}$$

7.1.2 Multi-task learning

One important application of Bayesian inference for logistic regression is where there are multiple related classifiers to fit. In this case, they need to share information between the parameters for each classifier.

7.1.3 Discrete factor analysis

A topic model is a latent variable model for text documents and other forms of discrete data. Often is assumed that the distribution over topics has a Dirichlet prior, but a more powerful model, known as the correlated topic model, uses a Gaussian prior, which can model correlations more easily.

7.2 Bohning's quadratic bound to the log-sum exp function

All of the above examples require dealing with multiplying a Gaussian prior by a multinomial likelihood; this is difficult because of the log-sum-exp (lse) term. The "Gaussianized" version of the observation model:

$$p(y_i|x_i,w) \geq f(x_i,\psi_i)\mathcal{N}(\widetilde{y}_i|X_i,w,A_i^{-1}) \tag{46}$$

where $f(x_i,\psi_i$ is some function that does not depend on w.

$$g(\psi_i) = exp[\psi_i - lse(\psi_i)] \tag{47}$$

$$A_i = \frac{1}{2}\left[I_M - \frac{1}{M+1}1_M 1_M^T\right] \tag{48}$$

$$b_i = A_i\psi_i - g(\psi) \tag{49}$$

$$\widetilde{y}_i = A_i^{-1}(b_i + y_i) \tag{50}$$

Given this, it is possible to compute the posterior $q(w) = \mathcal{N}(m_N, V_N)$, using Bayes rule for Gaussians.

7.3 Bounds for the sigmoid function

With $y_i \in \{0,1\}$, $M = 1$ and $\eta_i = w^T x_i$ where $w \in \mathbb{R}^D$ is a weight vector. The Bohning bound becomes

$$log(1+e^{\eta}) \leq \frac{1}{2}a\eta^2 - b\eta + c \tag{51}$$

$$a = \frac{1}{4} \tag{52}$$

$$b = A\psi - (1+e^{-\psi})^{-1} \tag{53}$$

$$c = \frac{1}{2}A\psi^2 - (1+e^{-\psi})^{-1}\psi + log(1+e^{\psi}) \tag{54}$$

Another bound, called JJ bound has the following form

$$log(1+e^{\eta}) \leq \lambda(\xi)(\eta^2 - \xi^2) = \frac{1}{2}(\eta - \xi) + log(1+e^{\xi}) \tag{55}$$

$$\lambda(\xi) = \frac{1}{4\xi}tanh(\xi/2) = \frac{1}{2\xi}\left[sigm(\xi) - \frac{1}{2}\right] \tag{56}$$

7.4 Other bounds and approximations to the log-sum-exp function

There are several other bounds and approximations to the multicalss lse function which can be used.

7.4.1 Product of sigmoids

$$log\left(\sum_{k=1}^{K} e^{\eta k}\right) \leq \alpha + \sum_{k=1}^{K} log(1+e^{\eta k - \alpha}) \tag{57}$$

Apllying the JJ bound to the term on the right.

7.4.2 Jensen's inequality

$$\mathbb{E}[lse(\eta_i)] \leq log\left(1 + \sum_{c=1}^{M} exp(x_i^T m_{N,c} + \frac{1}{2}x_i^T V_{N,cc} x_i)\right) \tag{58}$$

where the last term follows from the mean of a log-nromal distribution, which is $e^{\mu+\sigma^2/2}$

7.4.3 Multivariate delta method

The approach uses the multivariate delta method, which is a way to approximate moments of a function using a Taylor series expansion. In more detail, let $f(w)$ be the function of interest. Using a second-order approximation around m comes

$$\mathbb{E}_q[lse(X_i w)] \approx lse(X_i m) + \frac{1}{2} tr[X_i H X_i^T V] \tag{59}$$

where H is the Hessian for the lse function.

References

[1] K. P. Murphy, *Machine Learning : A Probabilistic Perspective*, 1st. Cambridge, Mass. [u.a.]: MIT Press, 2013.