Variational Inference

Lesson No. 10

João Victor da Silva Guerra

1 Introduction

Along the course, we have learned several algorithms for computing a posterior distribution, including discrete graphical models and Gaussian graphical models. However, some distribution does have a intractable form; hence, we will discuss a more general class of deterministic approximate inference algorithms based on **variational inference** ([1]).

2 Variational inference

Suppose we have an intractable distribution ($p^*(x)$) and an approximation of it ($q(x)$), with some tractable form, such as multivariate Guassian distribution.

The q distribution has some free parameters that have to be optimized to make it close to p^*. The cost funtion to be minimized is the reverse KL divergence:

$$\mathbb{KL}(q||p^*) = \sum_x q(x) \log \frac{q(x)}{p^*(x)} \tag{1}$$

However, the Eq.(1) is intractable, because the pointwise evalutation of $p^*(x) = p(x|D)$ is difficult and the normalization constant ($Z = p(D)$) is intractable. Hence, the unnormalized distribution $\tilde{p}(x) = p(x, D) = p^*(x)Z$ is tractable. We rewrite our cost function as follows:

$$\begin{aligned} J(q) &= \mathbb{KL}(q||\tilde{p}) \\ &= \sum_x q(x) \log \frac{q(x)}{p^*(x)Z} \\ &= \sum_x q(x) \log \frac{q(x)}{p^*(x)} - \log Z \\ &= \mathbb{KL}(q||p^*) - \log Z \end{aligned} \tag{2}$$

As Z is constant, by minimizing the cost function, we will make $q(x)$ close to $p(x)$.

As the KL divergence is always greater than zero, then $J(q)$ is an upper bound on the negative log likelihood as:

$$\begin{aligned} J(q) &= \mathbb{KL}(q||p^*) - \log Z \geq -\log Z \\ &= \mathbb{KL}(q||p^*) - \log Z \geq -\log p(D) \end{aligned} \tag{3}$$

An alternative interpretation of the variational objective is as follows:

$$\begin{aligned} J(q) &= \mathbb{KL}(q||\tilde{p}) \\ &= \mathbb{E}_q[\log q(x)] - \mathbb{E}_q[\log p^*(x)] \\ &= -\mathbb{H}(q(x)) + \mathbb{E}_q[-\log p^*(x)] \end{aligned} \tag{4}$$

Further, another alternative interpretation of the objective is:

$$\begin{aligned} J(q) &= \mathbb{KL}(q||\tilde{p}) \\ &= \mathbb{E}_q[\log q(x) - \log p(x)p(D|x)] \\ &= \mathbb{E}_q[\log q(x) - \log p(x) - \log p(D|x)] \\ &= \mathbb{KL}(q||p) - \mathbb{E}_q[\log p(D|x)] \end{aligned} \tag{5}$$

3 Mean field approximation

One of the forms of the variational inference is the **mean field approximation**([1]), in which we assume that the posterior is a factorized approximation as follows:

$$q(x) = \prod_{i=1}^{D} q_i(x_i) \tag{6}$$

The goal is to solve a minimization problem in which we minimize the KL divergence for each q_i as follows:

$$\min_{q_1,\ldots,q_D} \mathbb{KL}(q||p) \tag{7}$$

where the parameters of each q_i is optimized by a coordinate descent method, where at each step we make an update as follows:

$$\log q_i(x_i) = \mathbb{E}_{q_{-i}}[\log \tilde{p}(x)] + constant \tag{8}$$

The goal of the variational inference is to minimize the upper bound or, equivalently, to maximimize the lower bound. Then, we maxime the lower bound

$$L(q) = -J(q) = -\mathbb{KL}(q||\tilde{p}) = \sum_x q(x) \log \frac{\tilde{p}(x)}{q(x)} \leq \log p(D) \tag{9}$$

Writing only the terms that involve q_i and considering all other terms constant, we have

$$\begin{aligned}
L(q_i) &= \sum_x \prod_i q_i(x_i) \left[\log \tilde{p}(x) - \sum_k \log q_k(x_k) \right] \\
&= \sum_{x_i} \sum_{x_{-i}} q_i(x_i) \prod_{i \neq j} q_j(x_j) \left[\log \tilde{p}(x) - \sum_{k \neq j} \log q_k(x_k) + \log q_j(x_j) \right] \\
&= \sum_{x_i} q_i(x_i) \sum_{x_{-i}} \prod_{i \neq j} q_j(x_j) \log \tilde{p}(x) - \sum_{x_i} q_i(x_i) \sum_{x_{-i}} \prod_i q_i(x_i) \left[\sum_{k \neq j} \log q_k(x_k) + \log q_i(x_i) \right] \\
&= \sum_{x_i} q_i(x_i) \log f_i(x_i) - \sum_{x_i} q_i(x_i) \sum x_{-i} \prod_{i \neq j} q_j(x_j) \log q_i(x_i) + \text{constant} \\
&= -\mathbb{KL}(q_i || f_i) + \text{constant}
\end{aligned} \tag{10}$$

where $\log f_i(x_i) = \sum_{-j} \prod_{i \neq j} q_i(x_i) \log \tilde{p} = \mathbb{E}_{q_{-i}}[\log \tilde{p}(x)]$

Then, we maximize $L(q_i)$ with the minimization of the KL divergence above, which is solved by defining $q_i = f_i$, as follows:

$$q_i(x_i) = \frac{1}{Z_i} \exp\left(\mathbb{E}_{q_{-i}}[\log \tilde{p}(x)] \right) \tag{11}$$

Since $\log q_i = \log f_i$, we also work with the following form:

$$\log q_i(x_i) = \mathbb{E}_{q_{-i}}[\log \tilde{p}(x)] + \text{constant} \tag{12}$$

4 Variational Bayes

Until now in the course, we focused on infering latent variables z_i assuming the parameters θ of the model are known. Here, we want to infer the parameters. We can make a fully factorized approximation, such as $p(\theta|D) \approx \prod_k q(\theta_k)$, then we have a variational Bayes (VB) method. Also, we can infer both latent variables and parameters, making a approximation of the form $p(\theta, z_{i:N}|D) \approx q(\theta) \prod_i q_i(z_i)$, then we have a variational Bayes EM method ([1]).

4.1 Varitional Bayes for a 1D Gaussian

A VB can be applied to infer the posterior on the parameters of a 1D Gaussian ($p(\mu,\lambda|D)$), where $\lambda = 1/\sigma^2$ is the precision.

We have the following form of the conjugate prior:

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})Ga(\lambda|a_0,b_0) \tag{13}$$

So, we want to approximate the parameter given the data, using an approximate factored posterior as follows:

$$q(\mu,\lambda) = q_\mu(\mu)q_\lambda(\lambda) \tag{14}$$

4.1.1 Log posterior distribution

Here, we have the unnormalized log posterior in the following form:

$$\begin{aligned}\log\tilde{p}(\mu,\lambda) &= \log p(\mu,\lambda,D) = \log p(D|\mu,\lambda) + \log p(\mu,\lambda) + \log p(\lambda)\\ &= \frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2 + \frac{1}{2}\log(\kappa_0\lambda) + (a_0-1)\log\lambda - b_0\lambda + \text{constant}\end{aligned} \tag{15}$$

4.1.2 Optimal $q_\mu(\mu)$

From Eq.(12), we can derive the optimal form for $q_\mu(\mu)$ by averaging over λ:

$$\begin{aligned}\log q_\mu(\mu) &= \mathbb{E}_{q_\lambda}[\log p(D|\mu,\lambda) + \log p(\mu|\lambda)] + \text{constant}\\ &= -\frac{\mathbb{E}_{q_\lambda}[\lambda]}{2}\left\{\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right\} + \text{constant}\\ &\equiv \mathcal{N}\left(\mu|\frac{\kappa_0\mu_0+N\bar{x}}{\kappa_0+N},\left((\kappa_0+N)\mathbb{E}_{q_\mu}[\lambda]\right)^{-1}\right)\\ &\equiv \mathcal{N}\left(\mu|\mu_N,\kappa_N^{-1}\right)\end{aligned} \tag{16}$$

4.1.3 Optimal $q_\lambda(\lambda)$

Again, from Eq.(12), we can derive the optimal form for $q_\lambda(\lambda)$:

$$\begin{aligned}\log q_\lambda(\lambda) &= \mathbb{E}_{q_\mu}[\log p(D|\mu,\lambda) + \log p(\mu|\lambda)] + \text{constant}\\ &= (a_0-1)\log\lambda - b_0\lambda + \frac{1}{2}\log\lambda + \frac{N}{2}\log\lambda - \frac{\lambda}{2}\mathbb{E}_{q_\mu}\left[\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right] + \text{constant}\\ &\equiv Ga\left(\lambda|\left(a_0+\frac{N+1}{2}\right),\left(b_0+\frac{1}{2}\mathbb{E}_{q_\mu}\left[\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(x_i-\mu)^2\right]\right)\right)\\ &\equiv Ga(\lambda|a_N,b_N)\end{aligned} \tag{17}$$

4.1.4 Expectations

To be able to find the optimals q_μ and q_λ, we have to compute the several expectations from above.

Since we derived a form for $q(\mu)$, we can compute the expectations $\mathbb{E}_{q(\mu)}[\mu]$ and $\mathbb{E}_{q(\mu)}[\mu^2]$ as follows:

$$\mathbb{E}_{q(\mu)}[\mu] = \mu_N \tag{18}$$

$$\mathbb{E}_{q(\mu)}[\mu^2] = \frac{1}{\kappa_N} + \mu_N^2 \tag{19}$$

Since we know that $q(\lambda) = Ga(\lambda|a_N,b_N)$, we can compute the $\mathbb{E}_{q_\lambda}[\lambda]$ as:

$$\mathbb{E}_{q_\lambda}[\lambda] = \frac{a_N}{b_N} \tag{20}$$

Now, with the Eqs. (18), (19) and (20), we have a explict for the optimal equations q_μ and q_λ.

For our optimal $q_\mu(\mu)$, we have the following explicit parameters:

$$\mu_N = \frac{\kappa_0 \mu_0 + N\bar{x}}{\kappa_0 + N} \tag{21}$$

$$\kappa_N = (\kappa_0 + N)\frac{a_N}{b_N} \tag{22}$$

For our optimal $q_\lambda(\lambda)$, we have the following explicit parameters:

$$a_N = a_0 + \frac{N+1}{2} \tag{23}$$

$$\begin{aligned} b_N &= b_0 + \kappa_0(\mathbb{E}[\mu^2] + \mu_0^2 - 2\mathbb{E}[\mu]\mu_0) + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + \mathbb{E}[\mu^2] - 2\mathbb{E}[\mu]x_i) \\ &= b_0 + \kappa_0(\frac{1}{\kappa_N} + \mu_N^2 + \mu_0^2 - 2\mu_N\mu_0) + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + \frac{1}{\kappa_N} + \mu_N^2 - 2\mu_N x_i) \\ &= b_0 + \frac{\kappa_0}{2\kappa_N} + \frac{\kappa_0}{2}(\mu_N - \mu_0)^2 + \frac{N}{2\kappa_N}\sum_{i=1}^{N}(x_i - \mu_0)^2 \end{aligned} \tag{24}$$

From the Eqs. (21), (22), (23) and (24), we notice that the μ_N and a_N are constants as they depend only on fixed variables, and κ_N and b_N dependent on changing variables, then they need to be updated iteractively.

5 Variational Bayes Expectation Maximization

Let us consider a latent variable models of the form $z_i \to x_i \leftarrow \theta$. The parameters θ and the latent variables z_i are unknown. A useful approach is to fit such models using Expectation Maximization (EM) algorithm, in which we infer the posterior on the latent variables ($p(z_i|x_i,\theta)$) in the Expectation (E) step, and compute a point estimate of the parameters (θ) in the Maximization step (M) ([1]).

Here, the basic ideia is to use a field mean approximation on the posterior as follows:

$$p(\theta, z_{i:N}|D) \approx q(\theta)q(z) = q(\theta)\prod_i q(z_i) \tag{25}$$

In the variational Bayes EM (VBEM), we alternate between the update of $p(z_i|D)$ in the E step and the update of $p(\theta|D)$ in the M step.

References

[1] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, pp. 27–33.