CSCI 590: Machine Learning

Lecture 22: Variational Bayes

Instructor: Murat Dundar

Acknowledgement:

1. PRML by Chris Bishop

Central Task (1)

Evaluate the posterior distribution of the latent variables $P(Z|X)$ or evaluate expectations with respect to $P(Z|X)$

For many models of practical interest, it is infeasible to evaluate this distribution or compute expectations

The dimensionality of the latent space may be too high to work with directly

The posterior may have a highly complex form for which expectations are not analytically tractable

Need approximations!

Central Task (2)

Approximations broadly fall into two categories: Deterministic and Stochastic

Deterministic: Variational inference or variational Bayes (Chapter 10)

Stochastic: Markov chain monte carlo (MCMC) (Chapter 11)

Variational Inference (1)

Standard Calculus: Deals with finding derivatives of functions.

A function is a mapping that takes the values of a variable as the input and returns the value of the function as the output.

The derivative of the function describes how the output value varies as we make infinitesimal changes to the input value

A functional is a mapping that takes a function as the input and that returns the value of the functional as the output.

An example: Entropy H[p]. Takes a probability distribution p(x) as the input and returns the quantity as the output.

$$\mathrm{H}[p] = \int p(x) \ln p(x) \, \mathrm{d}x$$

Variational Inference (2)

A functional derivative expresses how the value of the functional changes in response to infinitesimal changes to the input function.

The field that deals with functionals is called calculus of variations or variational calculus.

Many problems can be expressed in terms of an optimization problem in which the quantity being optimized is a functional

Suppose we have a fully Bayesian model in which all parameters are given prior distributions.

Our probabilistic model specifies the joint distribution $p(X,Z)$, and our goal is to find an approximation for the posterior distribution $p(Z|X)$ as well as for the model evidence $p(X)$

Variational Inference (3)

$$\log p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) = \log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) + \log p_{\boldsymbol{X}}(\boldsymbol{x})$$

$$\log p_{\boldsymbol{X}}(\boldsymbol{x}) = \log p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z}) - \log p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})$$

$$\log a - \log b = \log \frac{a}{c} - \log \frac{b}{c}$$

$$\log p_{\boldsymbol{X}}(\boldsymbol{x}) = \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} - \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})},$$

and multiplying both sides by $q_{\boldsymbol{Z}}(\boldsymbol{z})$ we obtain

$$q_{\boldsymbol{Z}}(\boldsymbol{z}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) = q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} - q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})}$$

Variational Inference (4)

By integrating with respect to $\boldsymbol{z}$, we have

$$\int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) \, d\boldsymbol{z} = \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} \, d\boldsymbol{z} - \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} \, d\boldsymbol{z}$$

$$\begin{aligned}
\mathcal{L}(q_{\boldsymbol{Z}}) &:= \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} \, d\boldsymbol{z} \\
&= \mathbb{E}_{q_{\boldsymbol{Z}}} \left[\log \frac{p_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{X},\boldsymbol{Z})}{q_{\boldsymbol{Z}}(\boldsymbol{Z})} \right] \\
\mathrm{KL}\left(q_{\boldsymbol{Z}} || p_{\boldsymbol{Z}|\boldsymbol{X}}\right) &:= -\int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{Z}}(\boldsymbol{z})} \, d\boldsymbol{z} \\
&= \int_{\mathbb{R}^m} q_{\boldsymbol{Z}}(\boldsymbol{z}) \log \frac{q_{\boldsymbol{Z}}(\boldsymbol{z})}{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})} \, d\boldsymbol{z} \\
&= \mathbb{E}_{q_{\boldsymbol{Z}}} \left[\log \frac{q_{\boldsymbol{Z}}(\boldsymbol{Z})}{p_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{X})} \right],
\end{aligned}$$

Variational Inference (5)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{10.2}$$

where we have defined

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \tag{10.3}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}. \tag{10.4}$$

We can maximize the lower bound L(q) by optimization with respect to q(Z), which is equivalent to minimizing the KL divergence.

If we allow any possible choice for q(Z), then the maximum of the lower bound occurs when the KL divergence vanishes, which occurs when q(Z)=p(Z|X)

However we assume p(Z|X) is intractable and instead consider a restricted family of distributions q(Z) and seek the member of this family for which KL divergence is minimized.

Variational Inference (6)

One way to restrict q(Z) is to use a parametric distribution q(Z|w) governed by a set of parameters w.

The lower bound L(q) then becomes a function of w, and we can use standard nonlinear optimization techniques to determine the optimal values for the parameters.

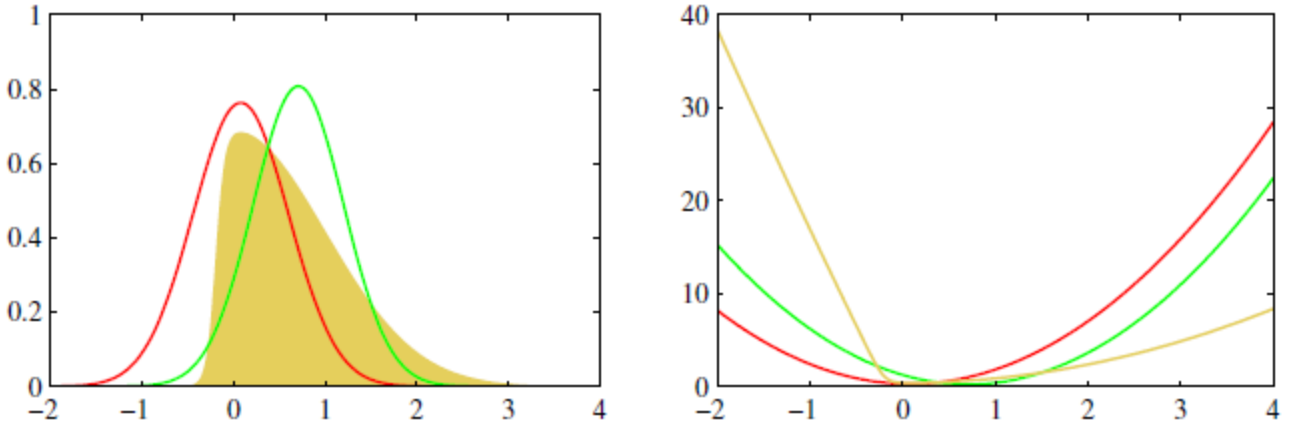

Figure 10.1 Illustration of the variational approximation for the example considered earlier in Figure 4.14. The left-hand plot shows the original distribution (yellow) along with the Laplace (red) and variational (green) approximations, and the right-hand plot shows the negative logarithms of the corresponding curves.

Factorized Distributions (1)

An alternative way to restrict q(Z) is to partition Z into disjoint groups Z_i, i=1,…,M.

We assume that the distribution factorizes with respect to these groups so that

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i).$$

This factorized form of variational inference corresponds to an approximation framework developed in Physics called mean field theory.

Goal: Amongst all distributions q(Z) having the above form, we seek that distribution for which the lower bound L(q) is largest.

We make a free form optimization of L(q) with respect to all of the distributions $q_i(Z_i)$, which we do by optimizing with respect to each of the factors in turn.

Factorized Distributions (2)

We first substitute the factorized form of q(Z) into L(q) and then dissect out the dependence on one of the factors $q_j(Z_j)$ simply defined by q_j

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \qquad (10.6)
\end{aligned}$$

where we have defined a new distribution $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.} \qquad (10.7)$$

Here the notation $\mathbb{E}_{i \neq j}[\cdots]$ denotes an expectation with respect to the q distributions over all variables $\mathbf{z}_i$ for $i \neq j$, so that

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i. \qquad (10.8)$$

Factorized Distributions (3)

Suppose we keep the $\{q_{i\neq j}\}$ fixed and maximize L(q) with respect to all possible forms for the distribution $q_j(Z_j)$.

Note that L(q)= $\int q_j \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j)\, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}$ is the

Negative KL divergence between $q_j(Z_j)$ and $\tilde{p}(X, Z_j)$. Thus maximizing L(q) is equivalent to minimizing the KL divergence and the minimum occurs when $q_j(Z_j) = \tilde{p}(X, Z_j)$.

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const.}$$

Solution: Initialize all of the factors $q_i(Z_i)$ appropriately and then cycle through the factors and replace each in turn with a revised estimate given by the above equation. Convergence is quaranteed because L(q) is convex with respect to each factor.

Properties of Factorized Approximations (1)

Consider approximating a Gaussian distribution using a factorized Gaussian.

Consider a Gaussian distribution $p(z) = N(z|\mu, \Lambda^{-1})$ over two correlated variables $z = (z_1, z_2)$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

Now suppose we wish to approximate this distribution using a factorized Gaussian of the form $q(z) = q_1(z_1)q_2(z_2)$

$$\begin{aligned} \ln q_1^\star(z_1) &= \mathbb{E}_{z_2}[\ln p(\mathbf{z})] + \text{const} \\ &= \mathbb{E}_{z_2}\left[-\frac{1}{2}(z_1-\mu_1)^2\Lambda_{11} - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + \text{const} \\ &= -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_1\Lambda_{11} - z_1\Lambda_{12}\left(\mathbb{E}[z_2]-\mu_2\right) + \text{const.} \qquad (10.11) \end{aligned}$$

Properties of Factorized Approximations (2)

Using the technique of completing the square, we can identify the mean and precision of this Gaussian, giving

$$q^\star(z_1) = \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) \tag{10.12}$$

where

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mathbb{E}[z_2] - \mu_2\right). \tag{10.13}$$

By symmetry, $q_2^\star(z_2)$ is also Gaussian and can be written as

$$q_2^\star(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) \tag{10.14}$$

in which

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mathbb{E}[z_1] - \mu_1\right). \tag{10.15}$$

Minimizing the Reverse KL divergence (1)

Suppose instead of maximizing L(q), we minimize the reverse KL divergence, i.e. KL(p||q).

When q(Z) is a factorized approximation, the KL divergence can then be written in the form

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] \mathrm{d}\mathbf{Z} + \text{const}$$

We can now optimize with respect to each of the factors $q_j(Z_j)$ which is easily done using a Lagrange multiplier to give

$$q_j^\star(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} \mathrm{d}\mathbf{Z}_i = p(\mathbf{Z}_j).$$

Note that this is a closed-form solution and does not require iteration.

Minimizing the Reverse KL divergence (2)

What is the difference between the two results?

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

There is a large positive contribution to the KL divergence from regions of Z space in which p(Z) is near zero unless q(Z) is also close to zero. Thus minimizing this form of KL divergence leads to q(Z) that avoid regions in which p(Z) is small.

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z}) \left[\sum_{i=1}^{M} \ln q_i(\mathbf{Z}_i) \right] \mathrm{d}\mathbf{Z} + \mathrm{const}$$

Conversely, KL(p||q) is minimized by q(Z) that are nonzero in reqions where p(Z) is nonzero.

Minimizing the Reverse KL divergence (3)

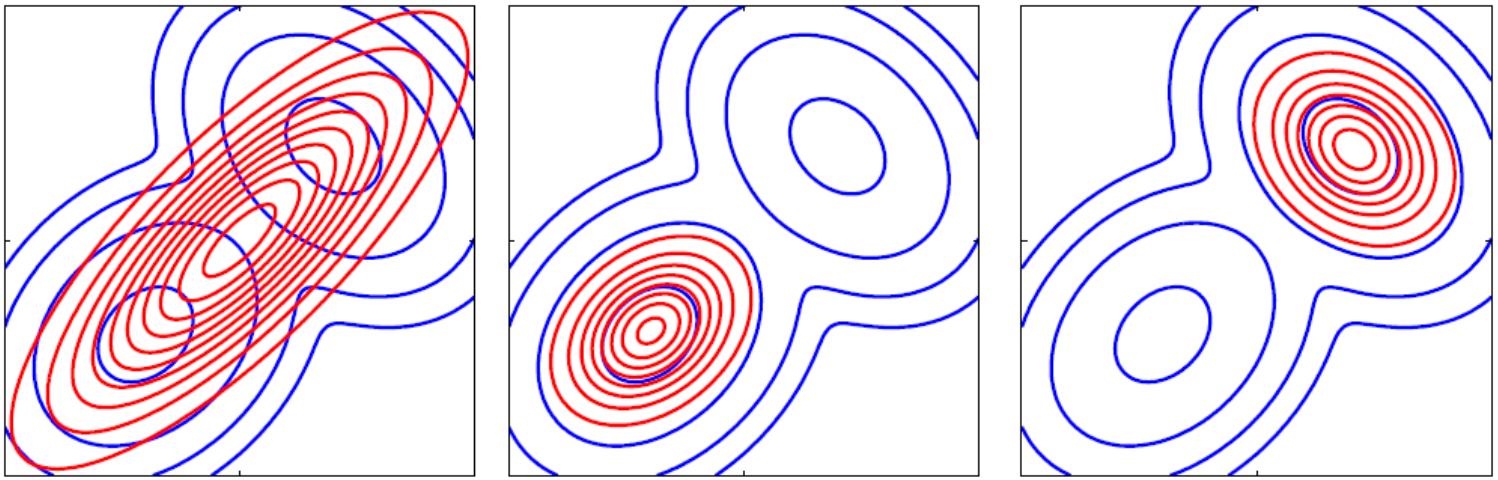

Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.