CS 8501 Advanced Topics in Machine Learning

Lecture 07: Variational Inference

Yangfeng Ji

Information and Language Processing Lab

Department of Computer Science

University of Virginia

https://yangfengji.net/

Posterior Inference

A Simple Problem Setup

Consider the following graphical model

$$x \to \mathcal{D}$$

where $\mathcal{D}$ is the observation (a set of training examples) and x is the latent variable.

Problem Setup

$$x \to \mathcal{D}$$

This is actually a generic setup of generative modeling.

Depending how we interpret x and $\mathcal{D}$, it can be mapped to many problems

- Clustering: x is the cluster index variable
- Dimension reduction: x is the low-dimensional representation
- Supervised learning: x represents the model parameter of a supervised model

Inference

Assume we know the parameter of the following two components

- $p(x; \theta_{prior})$
- $p(\mathcal{D} \mid x) = \prod_{i=1}^{n} p(d_i \mid x; \theta_{lik})$

The inference problem is to estimate the posterior distribution of x given $\mathcal{D}$ and its prior distribution $p(x)$, $p(x \mid \mathcal{D})$

Recall that

$$p(x \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid x)p(x)}{p(\mathcal{D})}$$

Conjugate Family

If we pick the special pair of prior $p(x)$ and likelihood function $p(\mathcal{D} \mid x)$, we can compute the analytical form

- $p(\mathcal{D} \mid x)$: Bernoulli; $p(x)$: Beta
- $p(\mathcal{D} \mid x)$: Gaussian; $p(x)$: Gaussian (assume the covariance matrix is I)

In many cases, computing an analytical posterior (mostly, because of $p(\mathcal{D})$) is impossible

Variational Inference

Basic Idea

Assume $p(x \mid \mathcal{D})$ is intractable, **variational inference** proposes to approximate $p(x \mid \mathcal{D})$ with another distribution $q(x)$.

Following our discussion in the previous lecture, we have two options to measure the distribution difference with *KL divergence*:

1. $\mathrm{KL}[p(x|\mathcal{D})\|q(x)] = \int_x p(x|\mathcal{D}) \log \frac{p(x|\mathcal{D})}{q(x)} dx$
2. $\mathrm{KL}[q(x)\|p(x|\mathcal{D})] = \int_x q(x) \log \frac{q(x)}{p(x|\mathcal{D})} dx$

Basic Idea (II)

- As KL divergence is not symmetric,

$$\mathrm{KL}[p\|q] \neq \mathrm{KL}[q\|p]$$

- Using $\mathrm{KL}[q(x)\|p(x \mid \mathcal{D})]$ is mostly due to some practical reasons

$$\mathrm{KL}[q(x)\|p(x|\mathcal{D})] = -\int q(x) \log p(x|\mathcal{D}) - H(q)$$

  as most of the computation is about $q(x)$, this formula gives more weight on picking a suitable approximation distribution

Remaining Question

With

$$\mathrm{KL}[q(x)||p(x|\mathcal{D})] = -\int q(x)\log p(x|\mathcal{D}) - H(q)$$

the underlying assumption is that we know $p(x|\mathcal{D})$.

However, most of the time, computing $q(x|\mathcal{D})$ itself is the main challenge

Evidence Lower Bound

Recall $p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})}$, we have

$$\int q(x) \log p(x|\mathcal{D}) = \int q(x) \log p(\mathcal{D}|x) + \int q(x) \log p(x) - \int q(x) \log p(\mathcal{D})$$

Therefore,

$$\mathrm{KL}[q(x)\|p(x|\mathcal{D})] = -\int q(x) \log p(\mathcal{D} \mid x) + \mathrm{KL}[q(x)\|p(x)] + \log p(\mathcal{D})$$

Evidence Lower Bound (ELBo)

Note that

$$\mathrm{KL}[q\|p] = -\int q(x)\log p(\mathcal{D}\mid x) + \mathrm{KL}[q(x)\|p(x)] + \log p(\mathcal{D})$$

Since $\mathrm{KL}[q\|p] \geq 0$, we have

$$\log p(\mathcal{D}) \geq \int q(x)\log p(\mathcal{D}\mid x) - \mathrm{KL}[q(x)\|p(x)]$$

In other words, RHS is the lower bound of the (log-) evidence $\log p(\mathcal{D})$

Optimization

Bring back the parameters of these distributions, variational inference can be reduced to the following optimization problem

$$\min_{\theta,\phi} - \int q(x;\phi) \log p(\mathcal{D} \mid x;\theta) + \mathrm{KL}[q(x;\phi) \| p(x;\theta)]$$

With θ and ϕ explicitly written in the above equation to represent the parameters for original data distribution and the variational distribution.

An Alternative Derivation

We can get the same objective function by starting from $\log p(\mathcal{D})$

$$\log p(\mathcal{D};\theta) = \log \int_x p(x|\mathcal{D};\theta)dx = \log \int_x q(x;\phi)\frac{p(x|\mathcal{D};\theta)}{q(x;\phi)}$$

With Jensen's inequality, we have

$$\log p(\mathcal{D};\theta) \geq \int_x q(x;\phi)\log\frac{p(x|\mathcal{D};\theta)}{q(x;\phi)}$$

An Alternative Derivation (II)

$$\int_x q(x;\phi) \log \frac{p(x|\mathcal{D};\theta)}{q(x;\phi)} =$$

Example: Gaussian Distributions

Consider the following Gaussian distribution $p(x) = \mathcal{N}(\mu, \Lambda)$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix}$$

The variational distribution $q(x)$ is defined as the product of two 1-D Gaussian distributions

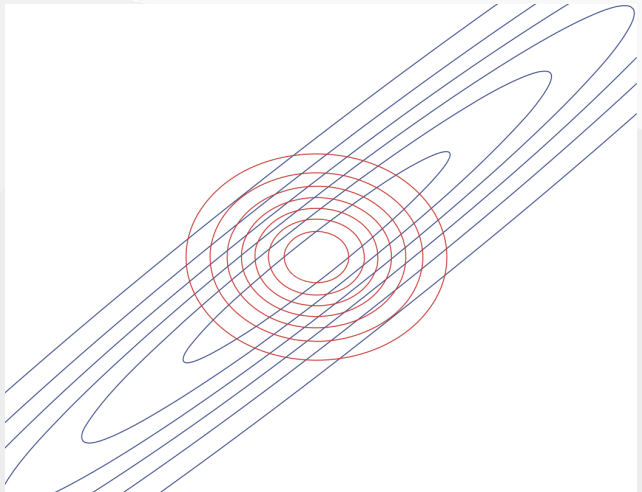
$$q(x) = \mathcal{N}(x_1; m_1, \sigma_1^2) \cdot \mathcal{N}(x_2; m_2, \sigma_2^2)$$

Example: Gaussian Distributions (II)

With both of them are Gaussian distributions, we can calculate the closed-form solution

$$q(x) = \mathcal{N}(x_1; m_1, \sigma_1^2) \cdot \mathcal{N}(x_2; m_2, \sigma_2^2)$$

- $\sigma_1^2 = \lambda_{11}^{-1}; \sigma_2^2 = \lambda_{22}^{-1}$
- $m_1 = \mu_1 - \lambda_{11}^{-1}\lambda_{12}(m_2 - \mu_2); m_2 = \mu_2 - \lambda_{22}^{-1}\lambda_{21}(m_1 - \mu_1)$



Further Comments

Given $\mathcal{D} = \{d_1, \ldots, d_n\}$

- In probabilistic modeling, $p(d_i \mid x)$ and $q(x)$ are often formulated with traditional probability distribution
- In the context of deep learning, each of them can be represented with a neural network

$$p(d_i \mid x) = \text{a neural network model}$$

$$q(x) = \text{another neural network model}$$

Forward vs. Reverse KL

- Forward:

$$\mathrm{KL}[p\|q] = \int p(x) \log \frac{p(x)}{q(x)}$$

- Reverse:

$$\mathrm{KL}[q\|p] = \int q(x) \log \frac{q(x)}{p(x)}$$

The key to understand the difference is to imagine a case where $p(x)$ or $q(x) \approx 0$

Example

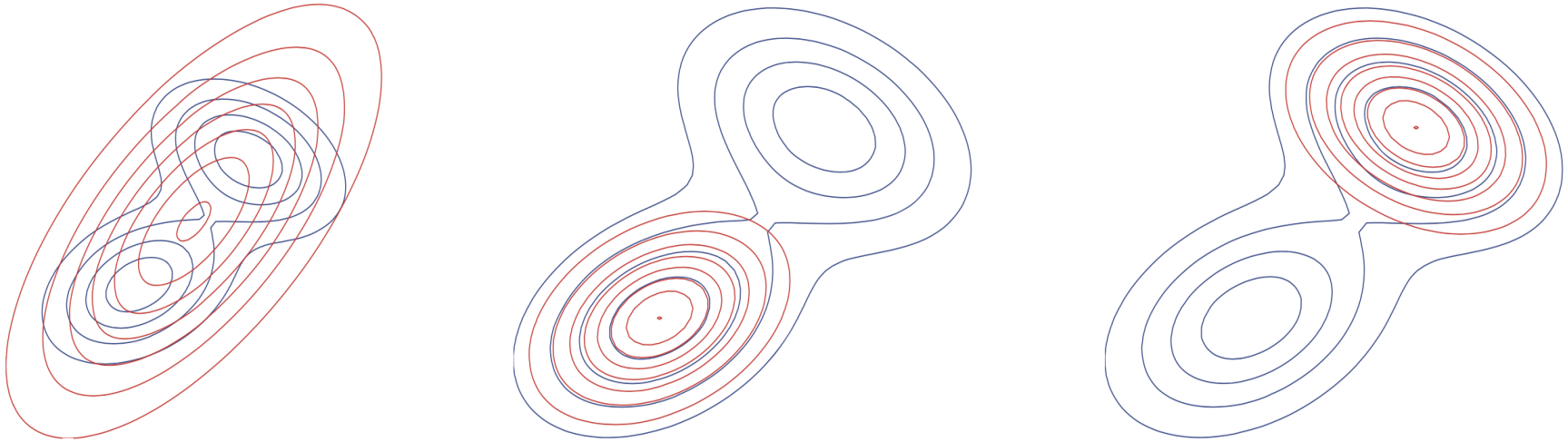


Figure 21.1 Illustrating forwards vs reverse KL on a bimodal distribution. The blue curves are the contours of the true distribution p. The red curves are the contours of the unimodal approximation q. (a) Minimizing forwards KL: q tends to "cover" p. (b-c) Minimizing reverse KL: q locks on to one of the two modes. Based on Figure 10.3 of (Bishop 2006b). Figure generated by `KLfwdReverseMixGauss`.

Example (II)

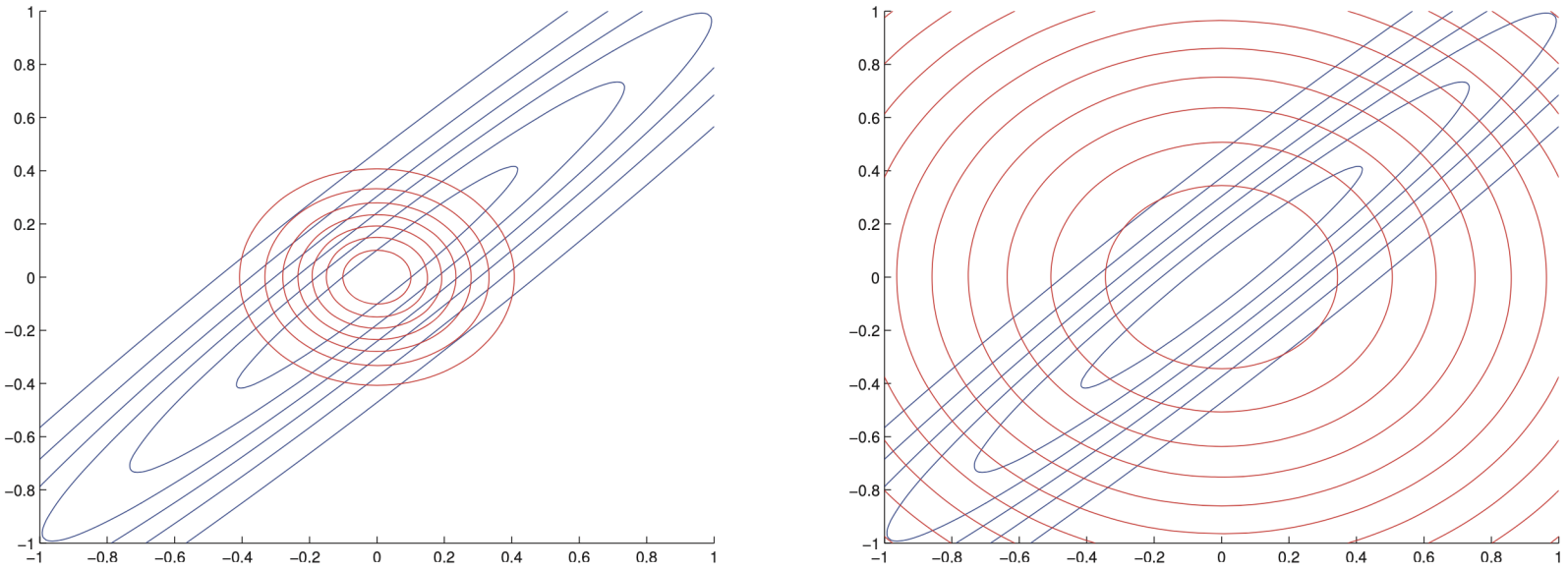


Figure 21.2 Illustrating forwards vs reverse KL on a symmetric Gaussian. The blue curves are the contours of the true distribution p. The red curves are the contours of a factorized approximation q. (a) Minimizing $\mathbb{KL}(q||p)$. (b) Minimizing $\mathbb{KL}(p||q)$. Based on Figure 10.2 of (Bishop 2006b). Figure generated by `KLpqGauss`.

The Mean Field Method

Example: Gaussian Distribution

Recall the previous example:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix}$$

The variational distribution $q(x)$ is defined as the product of two 1-D Gaussian distributions

$$q(x) = \mathcal{N}(x_1; m_1, \sigma_1^2) \cdot \mathcal{N}(x_2; m_2, \sigma_2^2)$$

General Form

In general, the mean field method consider $q(x)$ as a fully factored distribution. If x is the multi-variate random vector $x = (x_1, \ldots, x_n)$, the $q(x)$ is defined as

$$q(x) = \prod_n q_n(x_n)$$

Ising Model: Definition

The definition of Ising models with $x \in \{-1, +1\}^N$

$$p(x; \beta, J) = \frac{1}{Z(\beta, J)} \exp(-\beta E(x; J))$$

- the energy function is defined as

$$E(x; J) = -\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n$$

  where $J = \{J_{mn}, h_n\}$
- In this example, let's assume we know the parameters J -- we will remove this assumption in the next section

Ising Model: Variational Distribution

We define the variational distribution with parameter $a = \{a_n\}$as

$$q(x; a) = \frac{1}{Z(a)} \exp(\sum_n a_n x_n)$$

- Fully factorized: $q(x; a) = \prod_n q_n(x_n; a_n)$
- Probability

$$q(x_n = +1; a_n) \propto \exp(a_n); \quad q(x_n = -1; a_n) = \propto \exp(-a_n)$$

- Expectation

$$\bar{x}_n = \sum_{x_n} x_n q(x_n) = \frac{e^{a_n} - e^{-a_n}}{e^{a_n} + e^{-a_n}} = \tanh(a_n)$$

Objective Function

We follow the notation in statistical physics and use $\langle \cdot \rangle_q$ to represent the expectation under distribution q

With the variational distribution, we have

$$\mathrm{KL}[q(x;a) \| p(x;\beta,J)] = -\langle \log p(x;\beta,J) \rangle_q - H(q)$$

Minimizing the KL divergence will give us the $q(x;a)$ involves two terms

- the expectation term
- the entropy term

The Entropy Term

As $q(x; a)$ can be fully factorized, each entropy of $q(x_n; a_n)$ can be computed independently

$$H(q_n) = q_n(x_n = +1; a_n) \log \frac{1}{q_n(x_n = +1; a_n)} + q_n(x_n = -1; a_n) \log \frac{1}{q_n(x_n = -1; a_n)}$$

both $q_n(x_n)$ is a function of the variational parameter a_n

The Expectation Term

Now consider the expectation term:

$$\langle \log p(x;\beta,J)\rangle_q = \langle -\log Z(\beta,J) - \beta E(x;J)\rangle_q = -\log Z(\beta,J) - \beta\langle E(x;J)\rangle_q$$

Because of the independence defined in $q\big(x;a\big)$, we have

$$\langle E(x;J)\rangle_q = -\frac{1}{2}\sum_{m,n} J_{mn}\bar{x}_m\bar{x}_n - \sum_n h_n\bar{x}_n$$

where $\bar{x}_n$ is the expectation x_n under the distribution $q_n\big(x_n;a_n\big)$. In other words, $\bar{x}_n$ is the a function of a_n.

VI as Optimization

Given

$$\mathrm{KL}[q(x;a)\|p(x;\beta,J)] = -\langle \log p(x;\beta,J)\rangle_q - H(q)$$

as a function of a.

Take the derivative of $\mathrm{KL}[q(x;a)\|p(x;\beta,J)]$ with respect to a_n, we have

$$a_n = \beta(\sum_m J_{mn}\bar{x}_m + h_n)$$

With a_n, we can decode x_n with by taking the mode or the average

Thank You!