Variational Inference: An Introduction

Elvis Cui

MS in Biostatistics, UCLA

November 18, 2019

Overview

References

- Maximum Likelihood from Incomplete Data via the EM Algorithm, Dempster et al., JRSSB, 1977. [paper]
- Main reference: **Pattern recognition and machine learning, C. Bishop, Springer, 2006. [Chapter 9, 10]**
- Elements of statistical learning, R. Tibshirani et al, Springer, 2009. [Chapter 8]
- Modern multivariate statistical techniques, A. Izeman, Springer, 2013. [Chapter 12]
- Numerical Analysis for Statisticians, K. Lange, Springer, 2010. [Chapter 12]

Model set-up

Suppose the following:

- $\mathbf{X} \in \mathbb{R}^{n \times p}$: observed variables.
- $\mathbf{Z} \in \mathbb{R}^{n \times k}$: latent variables.
- $\boldsymbol{\beta}$: parameters of interest.
- $\log p(\mathbf{X}|\boldsymbol{\beta})$: observed log-likelihood.

Lower bound of log-likelihood

Denote $Q(\mathbf{Z})$ as any positive measurable pmf/pdf of $\mathbf{Z}$, we have

$$
\begin{aligned}
\log p(\mathbf{X}|\beta) &= \log \sum_{\mathbf{z}} p(\mathbf{X}, \mathbf{z}|\beta) \\
&= \log \left[\sum_{\mathbf{z}} \frac{p(\mathbf{X}, \mathbf{z}|\beta)}{Q(\mathbf{z})} Q(\mathbf{z}) \right] \\
&= \log \mathbb{E}_{\mathbf{Z}}[\frac{p(\mathbf{X}, \mathbf{Z}|\beta)}{Q(\mathbf{Z})}] \\
&\geq \mathbb{E}_{\mathbf{Z}}[\log \frac{p(\mathbf{X}, \mathbf{Z}|\beta)}{Q(\mathbf{Z})}]
\end{aligned}
$$

By Jensen's Inequality.

Note that the lower bound is attained iff $\frac{p(\mathbf{X},\mathbf{Z}|\beta)}{Q(\mathbf{Z})}$ is a constant w.r.t. $\mathbf{Z}$ (it can depend on $\mathbf{X}$).

Derivation of Q(Z)

Assume $k = \frac{p(\mathbf{X},\mathbf{Z}|\beta)}{Q(\mathbf{Z})}$, then

$$\begin{aligned} k &= \int_\Omega kQ(\mathbf{Z})d\mathbf{z} \\ &= \int_\Omega p(\mathbf{X},\mathbf{Z}|\beta)d\mathbf{z} \\ &= p(\mathbf{X}|\beta) \end{aligned}$$

Therefore, the lower bound of $\log p(\mathbf{X}|\beta)$ is attained if we set $Q(\mathbf{Z})$ to be the posterior distribution of $\mathbf{Z}$, i.e.

$$Q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\beta)$$

MM Algorithm

- E-step: At t^{th} iteration, compute the lower bound of $\log p(\mathbf{X}|\boldsymbol{\beta})$, which is

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\beta}^{(t)}}[\log \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\beta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\beta}^{(t)})}]$$

$$= \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\beta}^{(t)}}[\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\beta})] + H(p(\mathbf{Z}|\mathbf{X},\boldsymbol{\beta}^{(t)}))$$

  Where $H(\cdot)$ denotes the entropy of a distribution.

- M-step: Note that $H(\cdot)$ does not involve $\boldsymbol{\beta}$, thus maximizing $\log p(\mathbf{X}|\boldsymbol{\beta})$ is equivalent to maximize the first conditional expectation:

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\beta}^{(t)}}[\log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\beta})]$$

EM as variational inference

We can rewrite

$$\log p(\mathbf{X}) = \mathcal{L}(q) + KL(q \| p)$$

where we have defined

$$\mathcal{L}(q) = \int_{\Omega} q(\mathbf{Z}) \log\{\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\} d\mathbf{Z}$$

$$KL(q \| p) = -\int_{\Omega} q(\mathbf{Z}) \log\{\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\} d\mathbf{Z}$$

Note that we absorb β into $\mathbf{Z}$. Also, $\mathcal{L}(q)$ is a lower bound of $\log p(\mathbf{X})$ and such bound is attained iff

$$KL(q \| p) = 0 \text{ iff } q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}) \text{ a.e.}$$

Intractable posterior distribution

What can we do if $p(\mathbf{Z}|\mathbf{X})$ is intractable ?

- Sampling methods via Markov Chain Monte Carlo (MCMC)
- Consider a restricted family of distributions $q(\mathbf{Z})$ and then seek the member of this family for which the KL divergence is minimized.

Example 1 (parametric models as restricted families)

We assume $q(\mathbf{Z})$ can be specified by a set of parameters $\boldsymbol{\omega}$. Then the lower bound $\mathcal{L}(q)$ becomes a function of $\boldsymbol{\omega}$.

Factorized distributions

- **Mean field theory**: we can partition $\mathbf{Z}$ into M disjoint groups, that is,

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

- **Rewrite $\mathcal{L}(q)$ as a function of $q_i(\mathbf{Z_i})$:**

$$\begin{aligned}\mathcal{L}(q) &= \int_{\Omega} q(\mathbf{z}) \log\{\frac{p(\mathbf{X}, \mathbf{z})}{q(\mathbf{z})}\} d\mathbf{z} \\ &= \int_{\Omega} \prod_{i=1}^{M} q_i [\log \mathbb{P}(\mathbf{X}, \mathbf{z}) - \sum_{i=1}^{M} \log q_i] d\mathbf{z}\end{aligned}$$

Where $q_i = q_i(\mathbf{Z_i})$ in short.

Factorized distributions cont'd

$$\mathcal{L}(q) = \int_{\Omega} \prod_{i=1}^{M} q_i [\log \mathbb{P}(\mathbf{X}, \mathbf{Z}) - \sum_{i=1}^{M} \log q_i] d\mathbf{Z}$$

$$= \int q_j \left(\int \prod_{i \neq j} q_j \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{z}_{\{-\mathbf{j}\}} \right) d\mathbf{z_j}$$

$$- \int \prod_{i=1}^{M} q_i (\sum_{i=1}^{M} q_i) d\mathbf{z}$$

Key step: the first can be rewritten as a probability density plus a constant and the second term can be considered as the entropy of $p(\mathbf{z_j})$ plus a constant (see next slide for detail).

Factorized distributions: Minimizing KL divergence

- First term can be written as (left as an exercise):

$$\int q_j \mathbb{E}_q(\log p(\mathbf{X}, \mathbf{Z}) | \mathbf{Z}_{\{-j\}}) d\mathbf{z}_j = \int q_j \log \widetilde{p}(\mathbf{X}, \mathbf{z_j}) d\mathbf{z_j} + C_1$$

Where $\widetilde{p}(\mathbf{X}, \mathbf{z_j})$ is a new pdf w.r.t. $\mathbf{z}_j$ and C_1 is a constant.

- Second term can be written as (left as an exercise):

$$\int q_j \log q_j d\mathbf{z}_j + C_2$$

Where C_2 is also a constant.

Minimizing KL divergence

Therefore, taking $\mathcal{L}(q)$ as a function of q_j, we have

$$\mathcal{L}(q) = -KL(q_j \| \widetilde{p}(\mathbf{X}, \mathbf{z_j})) + C$$

Where C is a constant. Thus, maximizing $\mathcal{L}(q)$ is equivalent to minimizing KL divergence which occurs when:

$$q_j(\mathbf{Z}_j) = \widetilde{p}(\mathbf{X}, \mathbf{z_j}))$$

In fact, we have more (but this is not an explicit solution):

$$q_j^*(\mathbf{z}_j) = \widetilde{p}(\mathbf{X}, \mathbf{z_j})) = \frac{\exp\{\mathbb{E}_q(\log p(\mathbf{X}, \mathbf{Z})|\mathbf{Z}_{\{-j\}})\}}{\int \exp\{\mathbb{E}_q(\log p(\mathbf{X}, \mathbf{Z})|\mathbf{Z}_{\{-j\}})\} d\mathbf{z}_j}$$

2 components factorized Gaussian

Example 2

- Suppose $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda}^{-1})$ where

$$\boldsymbol{\mu} = (\mu_1^T, \mu_2^T), \mathbf{\Lambda} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

- Goal: find $q_1^*(Z_1)$ where

$$\log q_1^*(Z_1) = \mathbb{E}_{q_2}(\log \mathbb{P}(\mathbf{X}, \mathbf{Z}) | Z_1) + C$$

- Solution: the log likelihood is

$$-\frac{1}{2} Z_2^T \Lambda_{11} Z_1 + \mu_1^T \Lambda_{11} Z_1 - (\mathbb{E}_{q_2}(Z_2) - \mu_2)^T \Lambda_{21} Z_1 + C$$

2 components factorized Gaussian

Example 3 (cont'd)

- Recall solution:

$$-\frac{1}{2}Z_2^T\Lambda_{11}Z_1 + \mu_1^T\Lambda_{11}Z_1 - (\mathbb{E}_{q_2}(Z_2) - \mu_2)^T\Lambda_{21}Z_1 + C$$

- This is a quadratic form $\Rightarrow q_1^*(Z_1)$ is Gaussian

$$\mathcal{N}(Z_1|m_1, \Lambda_{11}^{-1})$$

where

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}_{q_2}(Z_2) - \mu_2)$$

2 components Gaussian

Example 4 (cont'd)

- Besides, knowing that $\mathbb{E}_{q_2}(Z_2) = m_2$ and $\mathbb{E}_{q_1}(Z_1) = m_1$:

$$m_1 = \mu_1, m_2 = \mu_2$$

- Solution:

$$q_1^*(Z_1) = \mathcal{N}(Z_1 | \mu_1, \Lambda_{11}^{-1})$$
$$q_2^*(Z_2) = \mathcal{N}(Z_2 | \mu_2, \Lambda_{22}^{-1})$$

Minimizing the reverse KL divergence

Previously, we are minimizing $KL(q_j \| \widetilde{p}(\mathbf{X}, Z_j))$. Now suppose we want to minimize

$$KL(\widetilde{p}(\mathbf{X}, Z_j) \| q_j)$$

Which is

$$KL(p \| q) = -\int p(\mathbf{z})(\sum_{j=1}^{M} \log q_j(Z_j))d\mathbf{z} + C$$

This gives us

$$q_1^*(Z_1) = p(Z_1) = \mathcal{N}(\mu_1, \Sigma_{11})$$
$$q_2^*(Z_2) = p(Z_2) = \mathcal{N}(\mu_2, \Sigma_{22})$$

Sidenotes

- Factorized VI leads to estimation that is too compact (i.e. variance is too concentrated).
- Minimizing reversed KL leads to too broad estimation (i.e. variance is too high).

Sidenotes

Example 5 ($\alpha-$family of divergence)

$$D_\alpha(p\|q) = \frac{4}{1-\alpha^2}(1 - \int p(x)^{\frac{1+\alpha}{2}} q(x)^{\frac{1-\alpha}{2}} dx)$$

- $\alpha = 0$: Hellinger distance: $D_H(p\|q) = \int (p^{\frac{1}{2}} - q^{\frac{1}{2}})^2 dx)$
- $\alpha \to 1$: $KL(p\|q)$: zero-forcing
- $\alpha \to -1$: $KL(q\|p)$: zero-avoiding