CS 7140: Advanced Machine Learning

Lecture 9: Variational EM (12 Feb 2018)

Instructor

Jan-Willem van de Meent (j.vandemeent@northeastern.edu)

Scribes

Yaoshen Yuan (yuan.yaos@husky.neu.edu)
Elizabeth Baranovic (baranovic.e@husky.neu.edu)
Ruiyang Xu (xu.r@husky.neu.edu)

1 Variational Bayesian Expectation Maximization

Variational inference is the general name used to refer to inference strategies that approximate a posterior $p(z, \theta \mid y)$, which is generally intractable, with a variational distribution $q(z, \theta; \phi)$, which is chosen from some more tractable family of distributions. Variational inference maximizes an objective known as the evidence lower bound (ELBO),

$$\begin{aligned}\mathscr{L}(\phi) &= E_{q(z,\theta;\phi)}\left[\log \frac{p(y, z, \theta; \lambda)}{q(z, \theta; \phi)}\right], \\ &= E_{q(z,\theta;\phi)}\left[\log \frac{p(y; \lambda)p(z, \theta \mid y; \lambda)}{q(z, \theta; \phi)}\right], \\ &= \log p(y; \lambda) - \mathrm{KL}(q(z, \theta; \phi) \mid\mid (p(z, \theta \mid y; \lambda)), \leqslant \log p(y; \lambda),\end{aligned} \tag{1}$$

In this lecture, we discuss a particular algorithm known as variational Bayesian expectation maximization (VBEM). In this form of variational inference, we assume a variational distribution that factorizes into a distribution over latent variables z and a distribution over model parameters θ

$$q(z, \theta; \phi) = q(z; \phi^z)q(\theta; \phi^\theta). \tag{2}$$

We then maximize the ELBO with an extension of the expectation maximization algorithm:

- Repeat until $\mathcal{L}(\phi^\theta, \phi^z)$ has converged
  1. E-Step: $\phi^z = \mathrm{argmax}_{\phi^z}\, \mathcal{L}(\phi^\theta, \phi^z)$
  2. M-Step: $\phi^\theta = \mathrm{argmax}_{\phi^\theta}\, \mathcal{L}(\phi^\theta, \phi^z)$

This algorithm is closely related to the EM algorithm for maximum likelihood estimation that we discussed in the last lecture. Both algorithms maximize a lower bound. In EM we optimize a lower bound on the log likelihood $\log p(y; \theta)$, which is defined as an expectation over latent variables z,

$$\begin{aligned}\mathscr{L}(\theta, \phi) &= E_{q(z;\phi)}\left[\log \frac{p(y, z; \theta)}{q(z; \phi)}\right], \\ &= E_{q(z;\phi)}\left[\log \frac{p(y; \theta)p(z \mid y; \theta)}{q(z; \phi)}\right], \\ &= \log p(y; \theta) - \mathrm{KL}(q(z; \phi) \mid\mid (p(z \mid y; \theta))) \leqslant \log p(y; \theta).\end{aligned} \tag{3}$$

This lower bound is defined in terms of an expectation over z. VBEM differs from EM in that the we extend the generative model by assuming a prior $p(\theta; \lambda)$ over parameters

$$p(y, z, \theta; \lambda) = p(y, z \mid \theta) p(\theta; \lambda). \tag{4}$$

The ELBO that we optimize in VBEM is an expectation over both θ and z, rather than an expectation over z alone, and defines a lower bound on the log *marginal* likelihood

$$\log p(y; \lambda) = \log \left(\int p(y, z, \theta \,;\, \lambda) \, dz \, d\theta \right). \tag{5}$$

The main difference between variational inference and maximum likelihood estmation is that we learn a distribution $q(\theta, z; \phi)$ over both parameters θ and other latent variables z, whereas maximum likelihood estimation learns a distribution $q(z; \phi)$ and a point estimate θ for the parameter values that are most likely to explain the data. Learning a distribution over parameter values helps guard against overfitting, but it also poses computational challenges, since we now have to perform an additional expectation with respect to the parameter values. However, it turns out that the factorization $q(\theta, z; \phi) = q(\theta; \phi^\theta) q(z; \phi^z)$ ensures that we can still compute the E-step and M-step updates using closed-form expressions that are in practice very similar to those that we would use in EM for maximum likelihood estimation.

We can either derive the E-step and M-step updates directly by solving for 0 partial derivatives, or by using a more general argument bases on calculus of variations. We will begin by showing the first form of derviation, and then continue to the more general argument.

1.1 Derivation using Partial Derivatives (shown in class)

We can derive the E-step and M-step by solving for 0 derivatives

$$\nabla_{\phi^z} \mathcal{L}(\phi^z, \phi^\theta) = 0, \qquad \nabla_{\phi^\theta} \mathcal{L}(\phi^z, \phi^\theta) = 0. \tag{6}$$

We can simplify these derivatives by separating out terms that depend only on both z and θ, terms that depend only on z and terms that depend only on θ. When solving for ϕ^z, we can ignore any terms that only depend on θ,

$$\begin{aligned}
\nabla_{\phi^z} \mathscr{L}(\phi^z, \phi^\theta) &= \nabla_{\phi^z} \mathbb{E}_{q(z;\phi^z) q(\theta;\phi^\theta)} \left[\log \frac{p(y, z \mid \theta) p(\theta; \lambda)}{q(z; \phi^z) q(\theta; \phi^\theta)} \right] \\
&= \nabla_{\phi^z} \mathbb{E}_{q(z;\phi^z) q(\theta;\phi^\theta)} \left[\log p(y, z \mid \theta) + \log p(\theta) - \log q(z; \phi^z) - \log q(\theta; \phi^\theta) \right] \\
&= \nabla_{\phi^z} \mathbb{E}_{q(z;\phi^z)} \left[\mathbb{E}_{q(\theta;\phi^\theta)} \left[\log p(y, z \mid \theta) \right] - \log q(z; \phi^z) \right] = 0.
\end{aligned} \tag{7}$$

We see that solving for 0 derivative with respect to ϕ^z yields the condition

$$\nabla_{\phi^z} \mathbb{E}_{q(z;\phi^z)} \left[\log q(z; \phi^z) \right] = \nabla_{\phi^z} \mathbb{E}_{q(z;\phi^z)} \left[\mathbb{E}_{q(\theta;\phi^\theta)} \left[\log p(y, z \mid \theta) \right] \right]. \tag{8}$$

We now observe that this condition is satisfied when we define

$$q(z; \phi^z) \propto \exp \left(\mathbb{E}_{q(\theta;\phi^\theta)} \left[\log p(y, z \mid \theta) \right] \right). \tag{9}$$

When solving for ϕ^θ, we can similarly ignore any terms that only depend on z,

$$\nabla_{\phi^\theta} \mathscr{L}(\phi^z, \phi^\theta) = \nabla_{\phi^\theta} \mathbb{E}_{q(\theta;\phi^\theta)} \left[\mathbb{E}_{q(z;\phi^z)} \left[\log p(y, z \mid \theta) \right] + \log p(\theta; \lambda) - \log q(\theta; \phi^\theta) \right] = 0. \tag{10}$$

Algorithm 1: Variational Bayesian EM algorithm

Initialize: ϕ_θ

Define: $\mathscr{L}(\lambda, \phi^z, \phi^\theta) = \mathbb{E}_{q(z,\theta;\phi)}[\log \frac{p(y,z,\theta;\lambda)}{q(z,\theta;\phi)}] \leqslant \log p(y;\lambda)$

Repeat until $\mathscr{L}(\lambda, \phi^z, \phi^\theta)$ converges:

1. E-Step: Update the variational distribution on z

$$q(z;\phi^z) \propto \exp\left(E_{q(\theta;\phi^\theta)}[\log p(y,z,\theta)]\right)$$

2. M-Step: Update the variational distribution on θ

$$q(\theta;\phi^\theta) \propto \exp\left(E_{q(z;\phi^z)}[\log p(y,z,\theta)]\right)$$

We now similarly observe that this condition is satisfied when we define

$$q(\theta;\phi^\theta) \propto \exp\left(\mathbb{E}_{q(z;\phi^z)}\left[\log p(y,z \mid \theta)\right] + \log p(\theta;\lambda)\right). \tag{11}$$

For reasons of notational simplicity we often write these two conditions in the form

$$q(z;\phi^z) \propto \exp\left(\mathbb{E}_{q(\theta;\phi^\theta)}\left[\log p(y,z,\theta;\lambda)\right]\right), \tag{12}$$

$$q(\theta;\phi^\theta) \propto \exp\left(\mathbb{E}_{q(z;\phi^z)}\left[\log p(y,z,\theta;\lambda)\right]\right). \tag{13}$$

Here we have multiplied the expression in Equation 9 by a factor $p(\theta;\lambda)$, which does not alter the functional form of the distribution, since this term does not depend on z.

1.2 Functional derivative explanation (calculus of variations)

A more general form of derivation can be obtained using calculus of variations. To do so, we define the ELBO as a Lagrangian functional of $q(z)$ and $q(\theta)$, where we ensure that $q(z)$ and $q(\theta)$ are normalized by introducing Lagrange multipliers

$$\mathscr{L}[q(z),q(\theta)] = E_{q(z)q(\theta)}\left[\log \frac{p(y,z,\theta)}{q(z)q(\theta)}\right] + \lambda_\theta\left(1 - \int d\theta\, q(\theta)\right) + \lambda_z\left(1 - \int dz\, q(z)\right). \tag{14}$$

We can now solve for the 0 functional derivative by writing

$$\begin{aligned}\frac{\delta\mathscr{L}[q(z),q(\theta)]}{\delta q(\theta)} &= \frac{\partial}{\partial q(\theta)}\left(\int q(z)q(\theta)\log\frac{p(y,z,\theta)}{q(z)q(\theta)}\,dz - \lambda_\theta q(\theta)\right), \\ &= \left(\int q(z)\left(\log\frac{p(y,z,\theta)}{q(z)q(\theta)} - 1\right)dz - \lambda_\theta\right) = 0.\end{aligned} \tag{15}$$

From this we can read off the identity

$$\log q(\theta) = \mathbb{E}_{q(z)}\left[\log p(y,z,\theta)\right] - \mathbb{E}_{q(z)}[\log q(z)] - 1 - \lambda_\theta. \tag{16}$$

If we now treat all terms that do not depend on θ as constants then we recover Equation 13

$$q(\theta) \propto \exp\left(\mathbb{E}_{q(z)}\left[\log p(y,z,\theta)\right]\right). \tag{17}$$

Conversely, since these equations are symmetric in θ and z we can also write

$$\log q(z) = \mathbb{E}_{q(\theta)}\left[\log p(y,z,\theta)\right] - \mathbb{E}_{q(\theta)}[\log q(\theta)] - 1 - \lambda_z. \tag{18}$$

This in turn recovers the identity from Equation 12

$$q(z) \propto \exp\left(\mathbb{E}_{q(\theta)}\left[\log p(y,z,\theta)\right]\right). \tag{19}$$

2 Exponential Families and Conjugate Priors

2.1 Definition of Exponential Family

A distribution is an exponential family if it can be parameterized in the form

$$p(x|\eta) = h(x) \exp\left(\eta^T t(x) - a(\eta)\right), \tag{20}$$

This parameterization has the following components

- $h(x)$ is a base measure such as the Lebesgue or counting measure.
- η is a vector of natural (canonical) parameters.
- $t(x)$ is a vector of the sufficient statistics.
- $a(\eta)$ is a log normalizer. Also called the log partition function or cumulant function.

2.2 Univariate Gaussians

The exponential family includes many commonly used distributions including the Gaussian, Exponential, Gamma, Beta, Dirichlet, Bernoulli, Discrete, and Poisson distributions. As an example we will consider the univariate Gaussian. To see that this distribution is indeed an exponential family, we rearrange the terms in the density function as follows

$$p(x|\eta) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{21}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{x^2 - 2x\mu + \mu^2}{\sigma^2}\right), \tag{22}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(\left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]\begin{bmatrix} x \\ x^2 \end{bmatrix} - \frac{\mu^2}{2\sigma^2}\right), \tag{23}$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \exp\left(\left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]\begin{bmatrix} x \\ x^2 \end{bmatrix} - \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma^2)\right)\right). \tag{24}$$

We can now read off the components of the exponential family parameterization, which are

$$h(x) = \frac{1}{\sqrt{2\pi}}, \qquad \eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}, \qquad t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \qquad a(\eta) = \frac{\mu^2}{2\sigma^2} + \log(\sigma^2). \tag{25}$$

2.3 Moments

One of the properties of the exponential family is that its expected values of the sufficient statistics can be computed from the derivatives of the log normalizer. To see this we will make use of the fact that the distribution is normalized, which means that

$$1 = \int dx p(x \mid \eta). \tag{26}$$

If we now take the gradient of the integral above with respect to η, then we obtain the identity

$$0 = \nabla_\eta \int dx\, p(x \mid \eta), \tag{27}$$

$$= \int dx\, h(x) \nabla_\eta \exp\left(\eta^T t(x) - a(\eta)\right), \tag{28}$$

$$= \int dx\, h(x) \exp\left(\eta^T t(x) - a(\eta)\right) \left(t(x) - \nabla_\eta a(\eta)\right), \tag{29}$$

$$= \mathbb{E}_{p(x|\eta)}[t(x)] - \nabla_\eta a(\eta). \tag{30}$$

From this we see that we can express the expected value of the sufficient statistics, which is is known as the first moment of an exponential family distribution, in terms of the gradient of the log normalizer

$$\mathbb{E}_{p(x|\eta)}[t(x)] = \nabla_\eta a(\eta). \tag{31}$$

More generally, the n-th moment of the i-th component of the sufficient statistics $t_i(x)$ can be computed from the n-th partial derivative of the log normalizer with respect to the i-th component of the natural parameters η_i,

$$\frac{\partial^n a(\eta)}{\partial \eta_i^n} = \mathbb{E}_{p(x|\eta)}[t_i(x)^n]. \tag{32}$$

2.4 Minimal Exponential Families

When the sufficient statistics has a $t_i(x)$ of an exponential family are linearly independent, then the we refer to this distribution as a *minimal* exponential family, which has the following properties:

- The function $a(\eta)$ is convex.
- Because of convexity, there is a 1-1 mapping between η and $\nabla_\eta a(\eta)$.
- From this it in turn follows that there is a 1-1 mapping between η and $\mathbb{E}[t(x)]$.

2.5 Conjugate priors

In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

Suppose that we have a likelihood prior which are both exponential families

$$p(y|\eta) = h(y) \exp(\eta^T t(y) - a(\eta)), \qquad p(\eta|\lambda) = h(\eta) \exp(\lambda^T t(\eta) - a(\lambda)). \tag{33}$$

The prior is conjugate to the likelihood when we define

$$\lambda = [\lambda_1, \lambda_2], \qquad t(\eta) = [\eta, -a(\eta)], \tag{34}$$

which means that the conjugate prior has the form

$$p(\eta \mid \lambda) = h(\eta) \exp(\eta^T \lambda_1 - a(\eta)\lambda_2 - a(\lambda)). \tag{35}$$

For a conjugate prior and likelihood we can express the joint probability as

$$p(y, \eta|\lambda) = h(y)h(\eta) \exp(\eta^T (t(y) + \lambda_1) - a(\eta)(1 + \lambda_2) - a(\lambda)). \tag{36}$$

We can rewrite this expression by defining the posterior parameters

$$\tilde{\lambda} = \left[\tilde{\lambda}_1, \tilde{\lambda}_2\right], \qquad \tilde{\lambda}_1 = t(y) + \lambda_1, \qquad \tilde{\lambda}_2 = 1 + \lambda_2. \tag{37}$$

When we do this, we see that we can now express the joint as

$$p(y, \eta) = h(y)\, h(\eta)\, \exp(\eta^T \tilde{\lambda}_1 - a(\eta)\tilde{\lambda}_2 - a(\tilde{\lambda}))\, \exp(a(\tilde{\lambda}) - a(\lambda)), \tag{38}$$

$$= h(y)\, p(\eta \mid \tilde{\lambda})\, \exp(a(\tilde{\lambda}) - a(\lambda)). \tag{39}$$

Based on the fact that $p(y, \eta) = p(\eta \mid y)p(y)$, we can now read off the identities:

$$p(\eta \mid y) = p(\eta \mid \tilde{\lambda}), \tag{40}$$

$$p(y) = h(y) \exp(a(\tilde{\lambda}) - a(\lambda)). \tag{41}$$

In other words, the posterior $p(\eta \mid y)$ is once again an exponential family of the same form as the prior, whose parameters can be computed from the sufficient statistics $t(y)$. Moreover, we can compute the marginal likelihood from the log normalizer.

Finally, conjugate priors also allow us to compute the predictive distribution in closed form. Suppose that we have a series of observations $(y_1, \ldots, y_{n-1})$ and we now wish to compute the predictive distribution for the next data point y_n. We can express this predictive distribution as a ratio of joint distributions

$$p(y_n \mid y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n)}{p(y_1, \ldots, y_{n-1})} = \frac{\exp\left(a(\tilde{\lambda}_n) - a(\lambda)\right)}{\exp\left(a(\tilde{\lambda}_{n-1}) - a(\lambda)\right)} = \frac{\exp\left(a(\tilde{\lambda}_n)\right)}{\exp\left(a(\tilde{\lambda}_{n-1})\right)}, \tag{42}$$

where we define the posterior parameters $\tilde{\lambda}_n$ as

$$\tilde{\lambda}_n = \left(\sum_{i=1}^{n} t(y_i) + \lambda_1, n + \lambda_2\right) = \left(t(y_n) + \tilde{\lambda}_{n-1}, 1 + \tilde{\lambda}_{n-1}\right). \tag{43}$$

2.6 Variational EM with Exponential Families

Suppose that we define a conjugate likelihood and prior of the form

$$p(y, z \mid \eta) = h(y, z) \exp(\eta^T t(y, z) - a(\eta)), \tag{44}$$

$$p(\eta\,;\, \lambda) = h(\eta) \exp(\eta^T \lambda_1 - a(\eta)\lambda_2 - a(\lambda)), \tag{45}$$

When performing Variational EM we want to perform the updates

$$q(\eta) \propto \exp(\mathbb{E}_{q(z)}[\log p(y, z, \eta\,;\, \lambda)]), \tag{46}$$

$$q(z) \propto \exp(\mathbb{E}_{q(\eta)}[\log p(y, z, \eta\,;\, \lambda)]). \tag{47}$$

We can derive the update for $q(\eta)$ by ignoring all terms in $\log p(y, z, \eta; \lambda)$ that do not depend on η, which can be treated as constants of proportionality for purposes of computing the distribution. This yields the simplified form

$$q(\eta) \propto h(\eta) \exp\left(\eta^\top \left(\mathbb{E}_{q(z)}[t(y, z)] + \lambda_1\right) - a(\eta)(1 + \lambda_2)\right). \tag{48}$$

From this we see that $q(\eta)$ has the same form as the prior $p(\eta; \phi^\eta)$ with parameters

$$q(\eta) = p(\eta; \phi^\eta), \qquad \phi^\eta = \left(\mathbb{E}_{q(z)}[t(y, z)] + \lambda_1, 1 + \lambda_2\right). \tag{49}$$

Note here that we have not made any assumptions about $q(\eta)$. The analytical form of the variational posterior is entirely determined by the update equation. In practice, we can compute this update as long as we choose $q(z)$ in a manner that allows us to compute the expected value of the sufficient statistics $\mathbb{E}_{q(z)}[t(y, z)]$, which will typically be the case when $q(z)$ is a distribution over discrete variables such as topic or cluster assignments.

We similarly can derive the update for $q(z)$ by noticing that we can ignore all terms in $\log p(y, z, \eta \mid \lambda)$ that do not depend on z,

$$q(z) \propto h(y, z) \exp\left[\mathbb{E}_{q(\eta)}[\eta]^\top t(y, z)\right]. \tag{50}$$

Again, in the case where $q(z)$ takes the form of a distribution over discrete variables, then we see that we can now compute $q(z)$ from the equation above as long as we are able to compute $\mathbb{E}_{q(\eta)}[\eta]$, which is usually the case for exponential family priors.