Variational Bayes

Consider a model with observed data x and hidden variables z.

$$p(x,z) = p(x|z)\,p(z)$$

And the marginal likelihood is

$$p(x) = \int_Z p(x,z)\,dz = E_{p(z)}\left[p(x|z) \cdot p(z) \right]$$

Goal: find posterior $p(z|x)$.

~~Background~~

Idea: - posit an approximation from some family Q.
- Find the member of the family that minimizes the KL divergence to the true posterior.

$$q^*(z) = \underset{q(z) \in Q}{\arg\min}\; KL\left[q(z) \,\|\, p(z|x) \right] = E_{q(z)}\left(\log \frac{q(z)}{p(z|x)} \right)$$

$$= E_q\left(\log q(z)\right) - E_q\left(\log p(z|x)\right)$$

Reminder: $\cancel{KL(p \,\|\, q)}$

$$KL\left(P_L \,\|\, P_R \right) = \int_\Theta P_L(\Theta) \cdot \log \frac{P_L(\Theta)}{P_R(\Theta)}\, d\Theta$$

= expected log ratio under P_L

- We find this $q^*(z)$ by optimization.
- It only ever approximates the true posterior... underestimates variance
- less precise; usually faster

In general, we have

$$p(z|x) = \frac{p(z)\,p(x|z)}{p(x)} = \frac{p(z)\,p(x|z)}{\int_Z p(x,z)\,dz} \leftarrow \text{hard}$$

z could include "parameters" as well as "latent variables"

<u>Example</u>

$$(x_i|c_i,\mu) \sim N(c_i^T\mu,\ 1)$$

$$c_i \sim \text{Multinomial}\left(M=1,\ w=\left(\tfrac{1}{K},\dots,\tfrac{1}{K}\right)\right)$$ Mixture of normals

$$\mu_k \sim N(0,\ \tau^2)$$

"one-hot" encoding of the mixture component

"Params": μ, σ^2

"Latents": c_i for $i=1,\dots,n$

The joint distn of (data, pars) is

$$p(\mu,\sigma^2,c,x) = p(\mu,\sigma^2)\cdot\prod_{i=1}^{N} p(c_i)\,p(x_i|c_i,\mu)$$

The marginal likelihood ("evidence") is

$$p(x) = \int p(\mu,\sigma^2)\prod_{i=1}^{N}\sum_{c_i} p(c_i)\,p(x|c_i,\mu)\,d\mu$$

The evidence lower bound (ELBO)

Family of possible approximations: Q.

Each $q(z) \in Q$ is a candidate.

Goal: compute $\underset{q(z) \in Q}{\arg\min} \; KL\left[q(z) \,\|\, p(z|x) \right]$

Problem: this is not even computable:

$$KL\left(q(z) \,\|\, p(z|x) \right) = E_q \left(\log \frac{q(z)}{p(x|z)} \right)$$

$$* \qquad = E_q \left[\log q(z) \right] - E_q \left[\log p(x|z) \right]$$

$$= E_q \left[\log q(z) \right] - E_q \left[\log p(x,z) \right] + \log p(x)$$

This depends on $\log p(x)$, which we can't compute.

So instead consider the functional

$$ELBO(q) = E_q \left[\log p(x,z) \right] - E_q \left[\log q(z) \right]$$

Clearly $\underbrace{ELBO(q)}_{\leq \log p(x)} = -KL\left[q(z) \,\|\, p(z|x) \right] + \log p(x)$

~~So maximizing~~ (Hence ELBO)

So maximizing ELBO is equivalent to minimizing the KL divergence in (*)

Let's rewrite the ELBO:

$$\mathrm{ELBO}(q) = E[\log p(z)] + E[\log p(x|z)] - E[\log q(z)]$$

$$= \underbrace{E[\log p(x|z)]}_{\text{Expected log-likelihood of data under } q} - \underbrace{KL[q(z)\,\|\,p(z)]}_{\text{KL b/t } q(z) \text{ and prior } p(z)},$$

~~"complete data"~~

"Explain the data"

"Don't stray from the prior."

Mean-field family

Key question: what is Q?

The mean-field family are all $q(z)$ of the form

$$q(z) = \prod_{j=1}^{m} q_j(z_j)$$

Note: <u>not</u> a model of the observed data (no x!)
The ELBO connects x to the $q(z_j)$'s.

- In principle, can take any parametric form to match desired form of each parameter
- Sometimes, we can find an "optimal" form for q_j based on the model

Back to mixture of normals example

The mean-field family consists of models of this form:

$$q(\mu, \sigma^2, c) = q(\sigma^2) \cdot \prod_{k=1}^{K} \underbrace{q(\mu_k \mid m_k, s_k^2)}_{\text{Gaussian \cancel{prior} distn for } \mu_k} \cdot \prod_{i=1}^{N} q(c_i \mid \phi_i)$$

ϕ_i: K vector of probabilities

- We have simply asserted that these are Gaussians / categorical
- But in fact these choices are provably optimal (later)

$$\boxed{\log q_j^*(z_j) = \mathbb{E}_{-j}\left[\log p(z_j \mid z_{-j}, x)\right]}$$

Algorithms = coordinate ascent variational inference.

Consider the j^{th} latent variable z_j.

It's complete conditional is $p(z_j \mid z_{-j}, x)$,

Fact: with z_{-j} fixed, the optimal $q_j(z_j)$ is of the form

(prove later)

$$q_j^*(z_j) \propto \exp\left\{ E_{-j}\left[\log p(z_j \mid z_{-j}, x)\right]\right\}$$

or Equivalently, working with the joint,

$$q_j^*(z_j) \propto \exp\left\{ E_{-j}\left[\log p(z_j, z_{-j}, x)\right]\right\}$$

This expectation is w.r.t. the variational density over z_{-j}, ~~with~~ ie. $\prod_{l \neq j} q_l(z_l)$

Remember: all latent variables are independent. So these expectations don't involve the j^{th} ~~variable~~. variational factor.

- So coordinate ascent cycles through these variational factors, updating them one at a time. (Remember, we're updating the variational parameters)

Derivation of the key fact:

$$ELBO(q) = E_q[\log p(x,z)] - E_q[\log q(z)]$$

And $q(z) = \prod_{j=1}^{M} q_j(z_j)$

So write this isolating $q_j(z_j)$:

$$ELBO(q_j) = E_{q_j}\left[E_{q_{-j}}[\log p(x, z_j, z_{-j})]\right] - E_{q_j}[\log q_j(z_j)] + \text{constant}$$

(Iterated expectations)

all $E_{q_{-j}}$'s in here

Focus on this. This is, up to constant,

$$ELBO(q_j) = E_{q_j}[\log q_j^*(z_j)] - E_{q_j}[\log q_j(z_j)]$$

where $q_j^*(z_j) = \exp\{E_{-j}[\log p(x, z_j, z_{-j})]\}$

$$= -KL(q_j \,\|\, q_j^*)$$

So clearly we maximize $ELBO(q_j)$ by minimizing $KL(q_j \| q_j^*)$, which happens when $q_j(z_j) = q_j^*(z_j)$

QED

Return to the Gaussian mixture-model example

$x_i \in \mathbb{R}$, data point

μ_k : k^{th} mean

c_i : indicator vector w/ $c_{ik} = 1$ if x_i came from component k

τ^2: assume fixed $\mu_k \sim N(0, \tau^2)$

Mean-field family: $q(z) = \prod_{j=1}^{M} q_j(z_j)$

$$= \prod_{k=1}^{K} N(\mu_k \mid m_k, s_k^2) \cdot \prod_{i=1}^{N} p(c_i \mid \phi_i)$$

$$\mathrm{ELBO}(q) = E_q\left[\log\left(\prod_{i=1}^{N}\left[N(x_i \mid \mu, c)\right] p(\mu, c)\right)\right] - E_q\left[\log\left[\prod_{k=1}^{K} q(\mu_k \mid m_k, s_k^2)\right] + \log \prod_{i=1}^{N} q(c_i \mid \phi_i)\right]$$

(careful) This is a prior density... a number, whose expectation is taken

$$= \sum_{k=1}^{K} E_q\left[\log p(\mu_k) ; m_k, s_k^2\right]$$

$$+ \sum_{i=1}^{N}\left\{E\left(\log p(c_i) ; \phi_i\right) + E\left[\log p(x_i \mid c_i, \mu) ; \phi_i, m, s^2\right]\right\}$$

$$- \sum_{i=1}^{N} E\left\{\log q(c_i ; \phi_i)\right\} - \sum_{k=1}^{K} E\left\{\log q(\mu_k ; m_k, s_k^2)\right\}$$

① Update for c_i (cluster assignment)

From our earlier key fact, we have

$$q^*(c_i \mid \phi_i) \propto \exp\left\{\log p(c_i) + E\left[\log p(x_i \mid c_i, \mu) ; m, s^2\right]\right\}$$

all other terms not involving c_i are constants.

Take each term:

$$\log p(c_i) = \log \text{prior}$$
$$= \log\left(\frac{1}{K}\right) = \cancel{0} - \log K \qquad \text{regardless of } c_i$$

2nd term

$$E\left[\log p(x_i \mid c_i, \mu) \,;\, m, s^2\right]$$

Well, we can write $p(x_i \mid c_i, \mu)$ as follows:

$$p(x_i \mid c_i, \mu) = \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}}$$

So
$$\log p(x_i \mid c_i, \mu) = \sum_{k=1}^{K} c_{ik} \underbrace{\log p(x_i \mid \mu_k)}_{}$$
$$= -\frac{(x_i - \mu_k)^2}{2} + \text{constant}$$

So our second term is

$$E\left\{\log p(x_i \mid c_i, \mu) \,;\, \mu, s^2\right\} = E\left[\sum_{k=1}^{K} -c_{ik} \frac{(x_i - \mu_k)^2}{2}\right]$$

$$= \sum_K c_{ik}\, E\left[-\frac{(x_i - \mu_k)^2}{2} \,;\, m_k, s_k^2\right] + \text{constant}$$

$$= \cancel{\sum_K c_{ik} E} = \sum_K c_{ik}\left[E\left(\mu_k \,;\, m_k, s_k^2\right) x_i - \frac{E\left(\mu_k^2 \,;\, m_k, s_k^2\right)}{2}\right] + \text{con}$$

Note: $\sum_{k=1}^{K} c_{ik} = 1$ independently of c_i

Thus we need $E(\mu_k)$ and $E(\mu_k^2)$ under the variational Gaussian for each component:

$$E(\mu_k \,;\, m_k, s_k^2) = m_k$$
$$E(\mu_k^2 \,;\, m_k, s_k^2) = s_k^2 + m_k^2$$

(~~$E(x^2) = \text{var}(x)$~~) $\text{var}(x) = E(x^2) - E(x)^2$

So our optimal q looks like a multinomial:

$$q^*(c_i \mid \phi_i) \propto \prod_{k=1}^{K} \left[e^{\psi_{ik}}\right]^{c_{ik}}$$

where $\psi_{ik} = x_i E[\mu_k ; m_k, s_k^2] - \frac{E[\mu_k^2 ; m_k, s_k^2]}{2}$

So $\phi_{ik} \propto e^{\psi_{ik}}$

② Update for the mixture-component means.

Recall $q(\mu_k ; m_k, s_k^2) = N(\mu_k ; m_k, s_k^2)$

Again, use our basic fact:

$$q^*(\mu_k) \propto \exp\left\{ \log p(\mu_k) + \sum_{i=1}^{N} E\left[\log p(x_i \mid c_i, \mu) \; ; \; m_{-k}, s_{-k}^2\right] \right\}$$

So

$$\log q^*(\mu_k) = \log p(\mu_k) + \sum_{i=1}^{N} E\left[\log p(x_i \mid c_i, \mu) \; ; \phi_i, m_{-k}, s_{-k}^2\right] + \text{const.}$$

$$= \frac{-\mu_k^2}{2\tau^2} + \sum_i E\left[\sum_{l=1}^{K} c_{il} \log p(x_i \mid \mu_l) \; ; \; \phi_i, m_{-k}, s_{-k}^2\right] + \text{constant}$$

$$= \frac{-\mu_k^2}{2\tau^2} + \sum_i E\left[c_{ik} \log p(x_i \mid \mu_k) \; ; \; \phi_i, m_{-k}, s_{-k}^2\right] + \text{constant}$$

(all other terms in sum constant in μ_k)

$$= \frac{-\mu_k^2}{2\tau^2} + \sum_i \log p(x_i \mid \mu_k) \cdot E(c_{ik} ; \phi_i) + \text{constant}$$

$$= \frac{-\mu_k^2}{2\tau^2} + \sum_{i=1}^{N} \phi_{ik} \left(\frac{-(x_i - \mu_k)^2}{2} \right) + \text{constant}$$

$$= \frac{-\mu_k^2}{2\tau^2} + \sum_i \left[\phi_{ik} x_i \mu_k - \frac{\phi_{ik}\mu_k^2}{2} \right]$$

$$= \mu_k \left(\sum_{i=1}^{N} \phi_{ik} x_i \right) - \mu_k^2 \left[\frac{1}{2\tau^2} + \sum_i \frac{\phi_{ik}}{2} \right]$$

This is ~~the~~ log density of an exponential family with "natural sufficient statistics" (μ_k, μ_k^2) and "natural parameters" $\left(\sum_i \phi_{ik} x_i , \frac{-1}{2}\left[\frac{1}{\tau^2} + \sum_i \phi_{ik} \right] \right)$

This is a Gaussian (m_k, s_k^2)

$$\frac{m_k}{s_k^2} = \sum_i \phi_{ik} x_i$$

$$-\frac{1}{2s_k^2} = -\frac{1}{2} \left[\frac{1}{\tau^2} + \sum_i \phi_{ik} \right]$$

So $$s_k^2 = \frac{1}{\frac{1}{\tau^2} + \sum_i \phi_{ik}}$$

$$m_k = \frac{\sum_i \phi_{ik} x_i}{\frac{1}{\tau^2} + \sum_i \phi_{ik}}$$

How conjugate looking!

weights/precisions on each data point = variational probability (ϕ_{ik}) of being assigned to each cluster

Calculating the ELBO:

$$\text{ELBO}(q) = \sum_{k=1}^{K} \overset{①}{E_q\left[\log p(\mu_k) ; m_k, s_k^2\right]}$$

$$+ \sum_{i=1}^{N}\left[\overset{②}{E\left(\log p(c_i) ; \phi_i\right)} + \overset{③}{E\left[\log p(x_i \mid c_i, \mu) ; \phi_i, m, s^2\right]}\right]$$

$$- \sum_{i=1}^{N} \overset{④}{E\left\{\log q(c_i ; \phi_i)\right\}} - \sum_{k=1}^{K} \overset{⑤}{E\left\{\log q(\mu_k ; m_k, s_k^2)\right\}}$$

① $E_q\left[\log p(\mu_k) ; m_k, s_k^2\right]$

$$= E_q\left[-\frac{\mu_k^2}{2\tau^2} ; m_k, s_k^2\right]$$

$$= -\frac{1}{2\tau^2} E_q(\mu_k^2)$$

Well, under q, $\mu_k \sim N(m_k, s_k^2)$

so $E_q(\mu_k^2) = \text{var}_q(\mu_k) + E_q(\mu_k)^2$

$$= s_k^2 + m_k^2$$

So the whole term is

$$\sum_{k=1}^{K}\left(-\frac{1}{2\tau^2}\right)(s_k^2 + m_k^2)$$

② $E\{\log p(c_i) ; \phi_i] = \log \frac{1}{K}$ ... constant

③ $E\left\{\log p(x_i \mid c_i, \mu) ; \phi_i, m, s^2\right\} =$

$$\sum_{k} E(c_{ik})\left(E(\mu_k ; m_k, s_k^2) \cdot x_i - \frac{E(\mu_k^2 ; m_k, s_k^2)}{2}\right)$$

④ $E\{\log q(c_i; \phi_i)\}$

$$= E\left[\sum_{k=1}^{K} c_{ik} \log(\phi_{ik})\right]$$

$$= \sum_{k=1}^{K} (\log \phi_{ik}) E(c_{ik})$$

$$= \sum_{k=1}^{K} \phi_{ik} \log \phi_{ik}$$

⑤ $E\{\log q(\mu_k; m_k, s_k^2)\}$

$$= E\left\{-\frac{1}{2s_k^2}(\mu_k - m_k)^2\right\}$$

$$= \frac{-1}{2s_k^2} E(\mu_k^2 - 2\mu_k m_k + m_k^2)$$

$$= -\frac{1}{2s_k^2}\left[E(\mu_k^2) - 2m_k^2 + m_k^2\right]$$

$$= -\frac{1}{2s_k^2}\left\{s_k^2 + m_k^2 - 2m_k^2 + m_k^2\right\} = 0$$

Five-minute math: exponential families

Suppose that $f(x|\theta)$ takes the form

$$f(x|\theta) = \cancel{h(x)\cdot g(\theta)} \; h(x)\cdot \exp\left\{\eta(\theta)^T T(x) - A(\theta)\right\}$$

where

$$\theta \in \mathbb{R}^D$$
$$\eta : \mathbb{R}^D \to \mathbb{R}^D$$
$$T(x) : \mathbb{R} \to \mathbb{R}^D$$

$\boxed{\text{Exponential family}}$

If $\eta(\theta)$ is the identity, we have

$$f(x|\eta) = h(x)\cdot \exp\left\{\eta^T T(x) - A(\eta)\right\}$$

η: "natural parameter"

$T(x)$: "natural sufficient statistics"

$A(\eta)$ = log of normalization factor

Example: $X \sim N(\mu, \sigma^2)$ $\qquad \theta = (\mu, \sigma^2)$

Then $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{\eta^T T(x) - A(\eta)\right\}$$

where $\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T$, $T(x) = (x, x^2)^T$

Variational inference in exponential families

Consider our generic model $p(z, x)$, where x = data and z = hidden parameters.

Suppose that each complete conditional is in ~~an~~ an exponential family, with sufficient statistic z_j:

$$p(z_j \mid z_{-j}, x) = h(z_j) \exp\left\{ \eta_j(z_{-j}, x)^T z_j - a(\eta_j(z_{-j}, x)) \right\}$$

"parameters" = function of conditioning variables

"statistic"

Now consider a mean-field variational approximation, where $q(z) = \prod_j q_j(z_j)$.

~~The exponential~~

And recall our key fact, that in coordinate ascent, the optimal q_j takes the form

$$\log q_j^*(z_j) = E_{-j}\left[\log p(z_j \mid z_{-j}, x) \right] \quad \text{under } q_{-j}(z) = \prod_{\ell \neq j} q_\ell(z_\ell)$$

The exponential-family assumption makes this especially simple:

$$\log q_j^*(z_j) = E\left(\log p(z_j \mid z_{-j}, x)\right)$$

$$= \log h(z_j) + E\left[\eta_j(z_{-j}, x)\right]^T z_j - \underbrace{E\left[a(\eta_j(z_{-j}, x))\right]}_{\text{constant in } z_j}$$

So $q_j^*(z_j) \propto h(z_j) \cdot \exp\left\{ E\left[\eta_j(z_{-j}, x)\right] \cdot z_j \right\}$

- In the same exponential family as its complete conditional $p(z_j \mid z_{-j}, x)$.

- Each update: set parameter

$$\nu_j = E\left[\eta_j(z_{-j}, x)\right]$$

Any "conditionally conjugate" model in Bayes looks like this.

Suppose that: - β is a vector of "global" parameters
- z is a set of local parameters for each "context"

Thus $p(\beta, z, x) = p(\beta) \cdot \prod_{i=1}^{N} p(z_i, x_i \mid \beta)$

Mixture of Gaussians is a clear example.