MSA101/MVE187 2021 Lecture 13
Variational Bayes

Petter Mostad

Chalmers University

October 11, 2021

Review / overview

- Last time: The EM algorithm: Using Kullback-Leibler divergence to find a maximal posterior estimate.
- This time: Variational Bayes: Using Kullback-Leibler divergence to find an optimally fitting density to the posterior.

An extension of the KL notation

- Let us define

$$\mathrm{KL}[q||p] = \mathrm{E}_q\left[-\log\frac{p(z)}{q(z)}\right] = -\int q(z)\log\frac{p(z)}{q(z)}\,dz$$

  for any density $q(z)$ and *any function* $p(z)$ so that the integral exists. (For standard KL p must be a density).
- Consequence: If $p_2(z) = Cp_1(z)$ then for any q

$$\mathrm{KL}[q||p_2] = \mathrm{E}_q\left[-\log\frac{Cp_1(z)}{q(z)}\right] = -\log C + \mathrm{KL}[q||p_1].$$

- For example, if $\int p_2(z)\,dz = C$ then for any q

$$\mathrm{KL}[q||p_2] \geq -\log C$$

  because $\mathrm{KL}[q||p_2/C] \geq 0$, with minimum occurring when $q \propto_z p_2$.
- We still have

$$\mathrm{KL}[q||p] = \mathrm{E}_q[-\log p(z)] - H_q[Z]$$

  where $H_q[Z]$ is the entropy of a random variable Z with density q .

Example 1: Approximating the posterior

- In the identity

$$\pi(\mathrm{data}, \theta) = \pi(\theta \mid \mathrm{data})\pi(\mathrm{data})$$

  $\pi(\mathrm{data})$ is a constant as a function of θ.

- Thus for a density q for θ,

$$\mathrm{KL}[q||\pi(\mathrm{data}, \cdot)] = -\log \pi(\mathrm{data}) + \mathrm{KL}[q||\pi(\cdot \mid \mathrm{data})].$$

- We may try to find a q minimizing $\mathrm{KL}[q||\pi(\cdot \mid \mathrm{data})]$ by finding a q minimizing $\mathrm{KL}[q||\pi(\mathrm{data}, \cdot)]$: This is part of the Variational Bayes idea.

Example 2: The EM algorithm

- Consider the identity

$$\pi(x, z \mid \theta) = \pi(x \mid \theta)\pi(z \mid x, \theta).$$

  Considering this as a function of z, $\pi(x \mid \theta)$ is a constant.

- For a density q for z we get

$$\mathrm{KL}[q||\pi(x, \cdot \mid \theta)] = -\log \pi(x \mid \theta) + \mathrm{KL}[q||\pi(\cdot \mid x, \theta)]$$

- The above equation is in the core of the proof of the EM algorithm:
  - Set $q(z) = \pi(z \mid x, \theta^{OLD})$ for some θ^{OLD}. (E step)
  - Find a θ^{NEW} that minimizes the left-hand side. (M step)
  - Then, moving from θ^{OLD} to θ^{NEW}, the left-hand side will decrease, and $\mathrm{KL}(q||\pi(\cdot \mid x, \theta)]$ will increase. Thus $-\log \pi(x \mid \theta)$ will decrease.

Approximations using Variational Bayes

- Idea: Finding an approximation to the posterior $\pi(\theta \mid \text{data})$ in some family of densities $\mathcal{Q}$ that does not necessarily contain the posterior.
- More specifically find the $q \in \mathcal{Q}$ minimizing the Kullback Leibler divergence from q to the posterior.
- Writing as above

  $$\mathrm{KL}[q||\pi(\text{data}, \cdot)] = -\log \pi(\text{data}) + \mathrm{KL}[q||\pi(\cdot \mid \text{data})].$$

  we instead find the $\hat{q}$ minimizing $\mathrm{KL}[q||\pi(\text{data}, \cdot)]$.
- As $\log \pi(\text{data}) \geq -\mathrm{KL}[q||\pi(\text{data}, \cdot)]$ the value $-\mathrm{KL}[\hat{q}||\pi(\text{data}, \cdot)]$ is called the *evidence lower bound*, or ELBO.
- Thus we want to *maximize*

  $$\begin{aligned} \mathcal{L}(q) &= -\mathrm{KL}[q||\pi(\text{data}, \cdot)] = \int q(\theta) \log \frac{\pi(\text{data}, \theta)}{q(\theta)} \, d\theta \\ &= \mathrm{E}_q[\log \pi(\text{data}, \theta)] + H_q[\theta] \end{aligned}$$

  where $H_q[\theta]$ is the entropy of a variable θ with density q.

Splitting θ into components (or subvectors)

- Let us look for densities q that can be written as products

$$q(\theta) = \prod_{i=1}^{n} q_i(\theta_i)$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is split into (groups of) dimensions.

- For the entropy term we get that

$$H_q[\theta] = \sum_{i=1}^{n} H_{q_i}[\theta_i]$$

where θ_i are variables with densities q_i.

- For any $i \in 1, \ldots, n$ the first term of $\mathcal{L}(q)$ may be rewritten

$$\mathrm{E}_q[\log \pi(\text{data}, \theta)] = \mathrm{E}_{q_i}\left[\mathrm{E}_{q_j, j\neq i}[\log \pi(\text{data}, \theta)]\right]$$

- So if we fix all q_j with $j \neq i$, the optimal q_i maximizing $\mathcal{L}(q)$ is the q_i maximizing

$$\begin{aligned} & \mathrm{E}_{q_i}\left[\mathrm{E}_{q_j, j\neq i}[\log \pi(\text{data}, \theta)]\right] + H_{q_i}[\theta_i] \\ = \quad & -\mathrm{KL}\left[q_i \,||\, \exp\left(\mathrm{E}_{q_j, j\neq i}[\log \pi(\text{data}, \cdot)]\right)\right] \end{aligned}$$

First option: Solving simultaneous equations

- We have seen that $\mathrm{KL}\left[q_i || \exp\left(\mathrm{E}_{q_j, j\neq i}\left[\log \pi(\mathrm{data}, \cdot)\right]\right)\right]$ is minimized when
$$q_i(\theta_i) \propto_{\theta_i} \exp\left(\mathrm{E}_{q_j, j\neq i}\left[\log \pi(\mathrm{data}, \cdot)\right]\right)$$
- If we write out these n equations for $i = 1, \ldots, n$, they become n equations in the n unknowns $q_1, q_2, \ldots, q_n$.
- Sometimes it is possible to simultaneously solve these equations.
- NOTE: The solution we get is the optimal using the assumption that the posterior splits as independent distributions over $\theta_1, \theta_2, \ldots, \theta_n$, but *making no other assumptions*, e.g., about parametric classes.

Variational Bayes: Toy example

- Consider the following example:

$$\begin{aligned} y_1, \ldots, y_n &\sim \text{Normal}(\mu, \tau^{-1}) \\ \pi(\mu) &\propto 1 \\ \pi(\tau) &\propto 1/\tau \end{aligned}$$

- Using conjugacy, we get that the exact posterior is given by

$$\begin{aligned} \tau \mid y_1, \ldots, y_n &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2} s^2\right) \\ \mu \mid \tau, y_1, \ldots, y_n &\sim \text{Normal}\left(\overline{y}, (n\tau)^{-1}\right) \end{aligned}$$

  where s^2 is the sample variance.
- As an illustration, we find the Variational Bayes approximate posterior. Note:

$$\begin{aligned} \pi(y_1, \ldots, y_n, \mu, \tau) &\propto \frac{1}{\tau} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi/\tau}} \exp\left(-\frac{\tau}{2}(y_i - \mu)^2\right) \\ \log \pi(y_1, \ldots, y_n, \mu, \tau) &= C + \left(\frac{n}{2} - 1\right) \log \tau - \frac{\tau}{2}(n-1)s^2 - \frac{n\tau}{2}(\overline{y} - \mu)^2 \end{aligned}$$

Variational Bayes: Toy example continued

- We use as approximation for the posterior the family of densities $q(\mu, \tau) = q_1(\mu) q_2(\tau)$, so that we assume μ and τ are independent, but we do not make additional restrictions on q_1 and q_2.
- We get

$$\begin{aligned} & \exp\left(\mathrm{E}_\mu\left[\log \pi(\text{data}, \mu, \tau)\right]\right) \\ \propto_\tau \quad & \exp\left(\left(\frac{n}{2} - 1\right)\log \tau - \frac{\tau}{2}(n-1)s^2 - \frac{n\tau}{2}\,\mathrm{E}_\mu\left[(\overline{y} - \mu)^2\right]\right) \end{aligned}$$

- From this we see that

$$q_2(\tau) = \text{Gamma}\left(\tau; \frac{n}{2}, \frac{1}{2}(n-1)s^2 + \frac{n}{2}\,\mathrm{E}_\mu\left[(\overline{y} - \mu)^2\right]\right)$$

- We get

$$\exp\left(\mathrm{E}_\tau\left[\log \pi(\text{data}, \mu, \tau)\right]\right) \propto_\mu \exp\left(-\frac{n}{2}\,\mathrm{E}_\tau[\tau](\overline{y} - \mu)^2\right)$$

- From this we see that

$$q_1(\mu) = \text{Normal}\left(\mu; \overline{y}, \left(n\,\mathrm{E}_\tau[\tau]\right)^{-1}\right).$$

Variational Bayes: Toy example continued

- Taking expectations using these two densities leads to

$$\begin{aligned} \mathsf{E}_\tau[\tau] &= \frac{n/2}{(n-1)s^2/2 + n/2 \cdot \mathsf{E}_\mu\left[(\bar{y}-\mu)^2\right]} \\ \mathsf{E}_\mu\left[(\bar{y}-\mu)^2\right] &= \left(n\,\mathsf{E}_\tau[\tau]\right)^{-1} \end{aligned}$$

- This is two equations with two unknowns; solving gives

$$\begin{aligned} \mathsf{E}_\tau[\tau] &= \frac{1}{s^2} \\ \mathsf{E}_\mu\left[(\bar{y}-\mu)^2\right] &= \frac{s^2}{n} \end{aligned}$$

- The final solution is

$$\begin{aligned} q_2(\tau) &= \text{Gamma}\left(\tau; \frac{n}{2}, \frac{n}{2}s^2\right) \\ q_1(\mu) &= \text{Normal}\left(\mu; \bar{y}, \frac{s^2}{n}\right) \end{aligned}$$

- **Presentation break for R illustration**

Second option: Iterative solution

- We would like to minimize
$$\mathrm{KL}\left[q_i || \exp\left(\mathrm{E}_{q_j, j\neq i}\left[\log \pi(\text{data}, \cdot)\right]\right)\right]$$
for $i = 1, \dots, n$.
- If a simultaneous solution cannot be found, we can start with a reasonable solutions $q_1, q_2, \dots, q_n$ and then repeatedly cycle through $i = 1, \dots, n$ minimizing the KL divergence above for q_i (keeping q_j, $j \neq i$ fixed).
- Generally this is done by assuming that q_i is in some parametric family for each i, so that one can optimize over the values of the parameters.
- In this case, we assume that the posterior is approximated as splitting in independent factors over the θ_i, we assume that the q_i are in particular parametric families, and we may get approximation errors.
- However, the method may scale well in very high dimensions.
- The *mean field* variational Bayes approximation of the posterior.

What if we minimize $\mathrm{KL}[\pi(\text{data} \mid \cdot)||q]$ instead of $\mathrm{KL}[q||\pi(\text{data} \mid \cdot)]$?

- We have

$$\mathrm{KL}[\pi(\cdot \mid \text{data})||q] = -\int \pi(\theta \mid \text{data}) \log \frac{q(\theta)}{\pi(\theta \mid \text{data})}\, d\theta$$

$$= \int \pi(\theta \mid \text{data}) \log \pi(\theta \mid \text{data})\, d\theta - \int \pi(\theta \mid \text{data}) \log q(\theta)\, d\theta$$

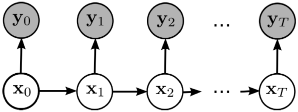
  so we only need to find the q maximizing the last term.
- If we assume that $q(\theta) = q(\theta \mid \eta) = \prod_{i=1}^{n} q_i(\theta_i \mid \eta_i)$ we get that

$$\begin{aligned} \int \pi(\theta \mid \text{data}) \log q(\theta \mid \eta)\, d\theta &= \sum_{i=1}^{n} \int \pi(\theta \mid \text{data}) \log q_i(\theta_i \mid \eta_i)\, d\theta \\ &= \sum_{i=1}^{n} \int \pi(\theta_i \mid \text{data}) \log q_i(\theta_i \mid \eta_i)\, d\theta_i. \end{aligned}$$

  So we optimize by setting $q_i(\theta_i \mid \eta_i)$ equal to the marginal posterior $\pi(\theta_i \mid \text{data})$ for each i (or choose η_i to minimize the KL divergence).
- Less useful approximations in practice.

From last time: The Baum-Welch algorithm (as EM example)

We consider an HMM where all the x_i have a finite state spaces



but where some of the parameters of the distributions $\pi(X_0)$, $\pi(X_i \mid X_{i-1})$, and $\pi(Y_i \mid X_i)$ are unknown. Objective: Given fixed values for the y_i, find maximum likelihood estimates for the parameters in the model.

- Note: If assuming flat priors the problem becomes that of computing the parameters maximizing the posterior, i.e., finding the MAP.
- Idea: Use the EM algorithm, with the values of the x_i as the augmented data.
- The E step of the EM algorithm is computed using (a small generalization of) the Forward-Backward algorithm.

The Baum-Welch algorithm: Simplified example

- For simplicity we assume each X_i can have values $1, \dots, M$, and we assume $X_0 = 1$. We assume there is one unknown parameter θ (with flat prior) with

$$\Pr(X_i = k \mid X_{i-1} = j) = \begin{cases} \theta/2 & |j-k| = 1 \text{ and } 1 < j < M \\ \theta & |j-k| = 1 \text{ and } j = 1 \text{ or } j = M \\ 1-\theta & j = k \end{cases}$$

- Assuming observed data is compatible with the model, the full loglikelihood given θ becomes

$$\begin{aligned} & \log\left(\pi(x_0, \dots, x_T, y_0, \dots, y_T \mid \theta)\right) \\ = \quad & \log \pi(x_0) + \sum_{i=1}^{T} \log \pi(x_i \mid x_{i-1}, \theta) + \sum_{i=0}^{T} \log \pi(y_i \mid x_i) \\ = \quad & C + c_1 \log \theta + c_2 \log(1-\theta) \end{aligned}$$

where c_1, c_2 are counts of one-step transitions, and stays in the same value, respectively, while C is a constant not involving θ.

Example continued

- In the E step, we would like to compute the expectation of the full loglikelihood under the distribution $\pi(x_0, \ldots, x_T \mid y_0, \ldots, y_T, \theta^{old})$ for some parameter θ^{old}.
- Thus we need to compute the expectations of the counts c_1 and c_2 under this distribution.
- Fixing θ^{old}, we can use the Forward-Backward algorithm to compute the densities $\pi(x_i \mid y_0, \ldots, y_i)$ and $\pi(y_{i+1}, \ldots, y_T \mid x_i)$. Further we have that

$$\begin{aligned}
& \pi(x_i, x_{i+1} \mid y_0, \ldots, y_T) \\
\propto \quad & \pi(y_{i+1}, \ldots, y_T \mid x_i, x_{i+1})\pi(x_i, x_{i+1} \mid y_0, \ldots, y_i) \\
\propto \quad & \pi(y_{i+2}, \ldots, y_T \mid x_{i+1})\pi(y_{i+1} \mid x_{i+1})\pi(x_{i+1} \mid x_i)\pi(x_i \mid y_0, \ldots, y_i)
\end{aligned}$$

  making it possible to compute the joint posterior for x_i and x_{i+1} from these densities.

Example continued

The algorithm can now be summed up as

- Choose starting parameter θ^{old}.
- Run the Forward-Backward algorithm on the Markov model with parameter θ^{old} to compute the numbers $\mathrm{E}[c_1]$ and $\mathrm{E}[c_2]$.
- Find the θ maximizing the expected loglikelihood

  $$\mathrm{E}[c_1] \log \theta + \mathrm{E}[c_2] \log(1 - \theta).$$

  In fact, we get

  $$\theta^{new} = \frac{1}{T} \mathrm{E}[c_1]$$

- Iterate until convergence.
- **See implementation in R**