MSA101/MVE187 2020 Lecture 12.1 Variational Bayes: Theory

Petter Mostad

Chalmers University

October 4, 2020

An extension of the KL notation

- The Kullback Leibler divergence from a density $q(x)$ to a density $p(x)$ is defined as

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

- By abuse of notation we extend the definition to cases where p is only proportional to a density. Then, if $p_1(x) = p_2(x)/C$ we get

$$\mathrm{KL}[q||p_1] = \log C + \mathrm{KL}[q||p_2].$$

- If $p(x)$ is a density and $p(x) = u(x)/C$, we get $\mathrm{KL}[q||u] \geq -\log C$, with the minimal value occuring when $u(x)$ is proportional to $q(x)$.
- Example: For a posterior $\pi(\theta \mid \text{data})$ we have

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

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

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

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

  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.
- Usually one uses the notation

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

$q \in \mathcal{Q}$ factorizing over subspaces for θ

- Assume $\mathcal{Q}$ consists of densities on the form $q(\theta \mid \eta)$ where $\eta \in \Omega$ for some set Ω.
- Assume there is a split $\theta = (\theta_1, \dots, \theta_n)$ of θ into (groups of) parameters and a corresponding split $\eta = (\eta_1, \dots, \eta_n)$ so that we can write for any $\eta \in \Omega$

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

- Note for example for the entropy of a variable with density $q(\theta \mid \eta)$: $-\int q(\theta \mid \eta) \log q(\theta \mid \eta)\, d\theta = -\sum_{i=1}^{n} \int q_i(\theta_i \mid \eta_i) \log q_i(\theta_i \mid \eta_i)\, d\theta_i$.
- We get

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

Optimizing one q_i at a time

- Assume we fix all q_j with $j \neq i$ and want to find the q_i maximizing $\mathcal{L}(q)$ under this restriction.
- Using the expression for $\mathcal{L}(q)$ above we find we must maximize

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

  where we take the expectation over all $q_j(\theta_j \mid \eta_j)$ for $j \neq i$.
- If it exists, use the η_i so that

$$q_i(\theta_i \mid \eta_i) \propto_{\theta_i} \exp\left(\mathrm{E}_{j\neq i}\left[\log \pi(\mathrm{data}, \theta)\right]\right)$$

  otherwise use an η_i minimizing

$$\mathrm{KL}\left[q_i || \exp\left(\mathrm{E}_{j\neq i}[\log \pi(\mathrm{data}, \cdot)]\right)\right].$$

Mean field variational Bayes approximation

- Sometimes, the set of equations

$$q_i(\theta_i \mid \eta_i) \propto_{\theta_i} \exp\left(\mathrm{E}_{j\neq i}\left[\log \pi(\text{data}, \theta)\right]\right)$$

  can be solved simultaneously for $i = 1, \ldots, n$.
- More commonly we set up an iterative algorithm where, for all $i = 1, \ldots, n$, we optimize each q_i given fixed values for q_j with $j \neq i$, and then make repeated cycles of these updates. This creates an algorithm that converges to an $\eta \in \Omega$ giving rise to a local maximum for $\mathcal{L}(q)$.
- This is 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

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

  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.

MSA101/MVE187 2020 Lecture 12.2
Variational Bayes: An example

Petter Mostad

Chalmers University

October 4, 2020

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

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

- This is two equations with two unknowns; solving gives

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

- The final solution is

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