MAST90125: Bayesian Statistical learning

Lecture 21: Expectation propagation

Feng Liu and Guoqi Qian

What have we covered so far

- In the last lecture, we discussed one method for approximate Bayesian inference, Variational Bayes. This method was based on partitioning the parameter vector $\boldsymbol{\theta}$ into sub-vectors $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_K$ and finding approximate posteriors $Q(\boldsymbol{\theta}_i); i = 1, \ldots, K$ for each sub-vector.

- In this lecture, we will discuss another method for approximate Bayesian inference, Expectation propagation. This method is based on partitioning the data, rather than the parameters.

Partitioning the posterior

- In expectation propagation, it is assumed that $p(\boldsymbol{\theta}|y)$ can be factorised as,

$$p(\boldsymbol{\theta}|y_1, \ldots y_n) = \prod_{i=0}^{n} f_i(\boldsymbol{\theta}),$$

  where i represents a datapoint.
- Question: How would you choose $f_i(\theta)$?
  - In this course, we have typically assumed that conditional on θ, observations are $i.i.d$. This means that,

$$p(\boldsymbol{\theta}|y_1, \ldots y_n) = \frac{p(y_1, \ldots y_n|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y_1, \ldots y_n)} = \frac{p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})}{p(y_1, \ldots y_n)}.$$

  - More importantly, for Bayesian computing it is usually sufficient to work with the known un-normalised density. Therefore a natural choice for $f_i(\boldsymbol{\theta})$ is $p(\boldsymbol{\theta})$ if $i = 0$ and $p(y_i|\boldsymbol{\theta})$ if $i = 1, \ldots, n$.

Approximating the posterior

- As we have been discovering since lecture 8, it is often not practical to determine the exact posterior analytically. So, does working with $f_i(\boldsymbol{\theta})$ as defined on the previous slide, and nothing else, allow us to determine the exact posteriors?
  - In general, no.
- To get around this, in expectation propagation we define $g(\boldsymbol{\theta})$ such that

$$g(\boldsymbol{\theta}) = \prod_{i=0}^{n} g_i(\boldsymbol{\theta}).$$

- We then estimate $g_i(\boldsymbol{\theta})$ such that $g_i(\boldsymbol{\theta}) \approx f_i(\boldsymbol{\theta}) \Rightarrow g(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta}|y_1, \ldots, n)$.

How do we learn $g_i(\boldsymbol{\theta})$?

- To determine $g_i(\boldsymbol{\theta})$, we need to define a *cavity distribution*

$$g_{-i}(\boldsymbol{\theta}) \propto \frac{g(\boldsymbol{\theta})}{g_i(\boldsymbol{\theta})}$$

  and *tilted distribution* assumed proportional to

$$g_{-i}(\boldsymbol{\theta}) f_i(\boldsymbol{\theta}).$$

- Having defined these distributions, the approximation of $g_{-i}(\boldsymbol{\theta}) f_i(\boldsymbol{\theta})$ is the new estimate of $g(\boldsymbol{\theta})$. Then the new estimate of $g_i(\boldsymbol{\theta})$ is defined as $g(\boldsymbol{\theta})/g_{-i}(\boldsymbol{\theta})$.
- This leaves us with unanswered questions:
  - How do we measure that the distribution $g(\boldsymbol{\theta})$ approximates $g_{-i}(\boldsymbol{\theta}) f_i(\boldsymbol{\theta})$?
  - How many choices do we have over the distributional form of $g(\boldsymbol{\theta})$?

How do we define distribution similarity?

- We have already encountered the Kullback-Leibler divergence many times.
- In expectation propagation, the divergence we are considering is

$$D_{KL}\{cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})||g(\boldsymbol{\theta})\} = \int cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})\{\log(cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})) - \log(g(\boldsymbol{\theta}))\}d\boldsymbol{\theta}.$$

- As D_{KL} is a strictly non-negative measure such that $D_{KL}(g||g) = 0$, we want to minimise this function.
  - Note the order of approximate and exact distributions has been reversed to that of D_{KL} in variational Bayes.
  - D_{KL} as used in expectation propagation is data-point specific. This means the final result of expectation propagation is not necessarily a global (i.e. over all data-points/prior) approximation to the posterior.

Choosing the distributional form of $g(\boldsymbol{\theta})$

- In expectation propagation, we need to choose the distributional form of the approximating distribution, $g(\boldsymbol{\theta})$.
  - Note: This is unlike variational Bayes, where the distributional form of $Q(\boldsymbol{\theta}_i)$ was dictated by the distributional form of the joint distribution.
- We will assume that $g(\boldsymbol{\theta})$ is a member of the exponential family of distributions. Hence we can write,

$$g(\boldsymbol{\theta}|\boldsymbol{\eta}) = h(\boldsymbol{\theta})a(\boldsymbol{\eta})e^{\boldsymbol{\eta}' u(\boldsymbol{\theta})} = h(\boldsymbol{\theta})e^{\boldsymbol{\eta}' u(\boldsymbol{\theta}) - A(\boldsymbol{\eta})},$$

where $A(\boldsymbol{\eta}) = -\log(a(\boldsymbol{\eta}))$, $\boldsymbol{\eta}$ is a vector of natural parameters, $u(\boldsymbol{\theta})$ is a vector of sufficient statistics with a first derivative satisfying

$$\frac{dg(\boldsymbol{\theta}|\boldsymbol{\eta})}{d\boldsymbol{\eta}} = (u(\boldsymbol{\theta}) - A'(\boldsymbol{\eta}))h(\boldsymbol{\theta})e^{\boldsymbol{\eta}' u(\boldsymbol{\theta}) - A(\boldsymbol{\eta})} = (u(\boldsymbol{\theta}) - A'(\boldsymbol{\eta}))g(\boldsymbol{\theta}|\boldsymbol{\eta}).$$

Minimising $D_{KL}\{g_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})||g(\boldsymbol{\theta})\}$

- To minimise the KL divergence, we need to solve $dD_{KL}/d\boldsymbol{\eta}$ =0.

$$
\begin{aligned}
0 = \frac{dD_{KL}\{g_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})||g(\boldsymbol{\theta}|\boldsymbol{\eta})\}}{d\boldsymbol{\eta}} &= \frac{d}{d\boldsymbol{\eta}}\int cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})\{\log(cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})) - \log(g(\boldsymbol{\theta}|\boldsymbol{\eta}))\}d\boldsymbol{\theta} \\
&= c\int g_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})\frac{d\{\log(cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})) - \log(g(\boldsymbol{\theta}|\boldsymbol{\eta}))\}}{d\boldsymbol{\eta}}d\boldsymbol{\theta} \\
&= -c\int g_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})(u(\boldsymbol{\theta}) - A'(\boldsymbol{\eta}))d\boldsymbol{\theta} \\
A'(\boldsymbol{\eta}) &= E_{cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})}(u(\boldsymbol{\theta}))
\end{aligned}
$$

- In addition, we know that

$$
\int \frac{dg(\boldsymbol{\theta}|\boldsymbol{\eta})}{d\boldsymbol{\eta}}d\boldsymbol{\theta} = \frac{d}{d\boldsymbol{\eta}}\int g(\boldsymbol{\theta}|\eta)d\boldsymbol{\theta} = \frac{d}{d\eta}1 = 0.
$$

Minimising $D_{KL}\{g_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})||g(\boldsymbol{\theta})\}$

- At the same time, we know that

$$\int \frac{dg(\boldsymbol{\theta}|\boldsymbol{\eta})}{d\boldsymbol{\eta}} d\boldsymbol{\theta} = \int (u(\boldsymbol{\theta}) - A'(\boldsymbol{\eta}))g(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta} = E_{\boldsymbol{\theta}|\boldsymbol{\eta}}(u(\boldsymbol{\theta})) - A'(\boldsymbol{\eta}).$$

- Combining the two results for $\int \frac{dg(\boldsymbol{\theta}|\boldsymbol{\eta})}{d\boldsymbol{\eta}} d\boldsymbol{\theta}$, and the result for D_{KL} we find that

$$\begin{aligned} E_{\boldsymbol{\theta}|\boldsymbol{\eta}}(u(\boldsymbol{\theta})) &= A'(\boldsymbol{\eta}) \qquad (1) \\ E_{cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})}(u(\boldsymbol{\theta})) &= A'(\boldsymbol{\eta}) \qquad (2) \end{aligned}$$

Aims of the algorithm

- Combining (1) and (2) we get the equality

$$E_{cg_{-i}(\boldsymbol{\theta})f_i(\boldsymbol{\theta})}(u(\boldsymbol{\theta})) = E_{\boldsymbol{\theta}|\boldsymbol{\eta}}(u(\boldsymbol{\theta})).$$

  Thus determining $g(\boldsymbol{\theta})$ requires iterative moment matching until convergence.

- Before writing up the algorithm, let's choose a specific distribution for $g(\boldsymbol{\theta})$ with desirable properties. Usually, $g(\boldsymbol{\theta})$ is assumed multivariate normal, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with natural parameters $\boldsymbol{\Sigma}^{-1}, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. This implies that $g_i(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. If we match the kernels of $g(\boldsymbol{\theta})$ and $\prod_{i=0}^{n} g_i(\boldsymbol{\theta})$,

$$e^{-\frac{\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}}{2}} e^{\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = \prod_{i=0}^{n} e^{-\frac{\boldsymbol{\theta}'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\theta}}{2}} e^{\boldsymbol{\theta}'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i} = e^{-\frac{\boldsymbol{\theta}'\sum_{i=0}^{n}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\theta}}{2}} e^{\boldsymbol{\theta}'\sum_{i=0}^{n}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i},$$

  we find $\boldsymbol{\Sigma}^{-1} = \sum_{i=0}^{n}\boldsymbol{\Sigma}_i^{-1}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \sum_{i=0}^{n}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$.

Expectation propagation for a regression with known error variance.

- This algorithm assumes
  - $f_i(\boldsymbol{\beta}) = p(y_i|\theta_i) = p(y_i|\mathbf{x}_i'\boldsymbol{\beta}) = d(y_i, \eta^{-1}(\mathbf{x}_i'\boldsymbol{\beta}))$, where d is some arbitrary distribution and η is a link function.
  - $g(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $g_i(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.
  - $g_0(\boldsymbol{\beta})$ is fixed to the prior $p(\boldsymbol{\beta})$, so does not need updating.

- For $t = 1, \ldots,$
  - For $i = 1, \ldots, n$
    1. Compute the (natural) parameters of the cavity distribution $g_{-i}(\boldsymbol{\beta})$,
       $\boldsymbol{\Sigma}_{-i}^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_i^{-1}$, $\boldsymbol{\Sigma}_{-i}^{-1}\boldsymbol{\mu}_{-i} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \Rightarrow \boldsymbol{\mu}_{-i} = \boldsymbol{\Sigma}_{-i}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i)$.
    2. Compute the parameters of $g_{-i}(\theta_i)$
       $M_{-i} = \mathbf{x}_i'\boldsymbol{\mu}_{-i}$, $V_{-i} = \mathbf{x}_i'\boldsymbol{\Sigma}_{-i}\mathbf{x}_i$

Expectation propagation for a regression with known error variance.

3 Construct the un-normalised tilted distribution and calculate,

$$E_k = \int_{-\infty}^{\infty} \theta_i^k g_{-i}(\theta_i) f_i(\theta_i) d\theta_i \quad \text{for } k = 0, 1, 2,$$

where E_0 is the normalising constant for tilted distribution. Then moment match and set $M = \frac{E_1}{E_0}$ and $V = \frac{E_2}{E_0} - M^2$ (M and V are two moments that $g(\boldsymbol{\theta})$ *should have*, i.e., the moment matching). Note in practice, we will need to specify finite bounds on the integral. The simplest choice would be $M_{-i} \pm \delta\sqrt{V_{-i}}$ for suitably large δ.

Expectation propagation for a regression with known error variance.

3 Construct the un-normalised tilted distribution and calculate,

$$E_k = \int_{-\infty}^{\infty} \theta_i^k g_{-i}(\theta_i) f_i(\theta_i) d\theta_i \quad \text{for } k = 0, 1, 2,$$

where E_0 is the normalising constant for tilted distribution. Then moment match and set $M = \frac{E_1}{E_0}$ and $V = \frac{E_2}{E_0} - M^2$ (M and V are two moments that $g(\boldsymbol{\theta})$ *should have*, i.e., the moment matching). Note in practice, we will need to specify finite bounds on the integral. The simplest choice would be $M_{-i} \pm \delta\sqrt{V_{-i}}$ for suitably large δ.

4 Determine the natural parameters of $g_i(\theta_i)$,
$M_i/V_i = M/V - M_{-i}/V_{-i}, \quad 1/V_i = 1/V - 1/V_{-i}$.

5 Transform the natural parameters found in step 4 to those of $g_i(\boldsymbol{\beta})$,
$\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i = \mathbf{x}_i M_i/V_i, \boldsymbol{\Sigma}_i^{-1} = \mathbf{x}_i(1/V_i)\mathbf{x}_i'$

6 Update the natural parameters of $g(\boldsymbol{\beta})$,
$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_{-i}^{-1}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{-i}^{-1}\boldsymbol{\mu}_{-i}$.

Stop once estimates have converged.

Comments

- The differencing method used in the expectation propagation algorithm for finding the natural parameters of $g_{-i}()$, would hold for any g in the exponential family. This is because the definition $g(\boldsymbol{\theta}) = \prod_{i=0}^{n} g_i(\boldsymbol{\theta})$ implies

$$e^{\boldsymbol{\eta}' u(\boldsymbol{\theta})} = e^{\sum_{i=0}^{n} \boldsymbol{\eta}_i' u(\boldsymbol{\theta})}.$$

- In previous lectures, we considered at hierarchical models and normal based regression. To fit such models in a fully Bayesian way, we needed to estimate parameters additional to $\boldsymbol{\beta}$, such as the error variance, σ_e^2.
  - How do you think we could do this using expectation propagation?
    - In the case of a normal regression, we could do this by assuming $g(\boldsymbol{\beta}, \sigma_e^2) = \prod_{i=0}^{n} g_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}) g_{i,\sigma_e^2}(\sigma_e^2)$. This would double the number of factors g_i we need to estimate in the algorithm.

Extending expectation propagation to more complicated models

- In the case of a normal regression, we could do this by assuming $g(\beta, \sigma_e^2) = \prod_{i=0}^{n} g_{i,\beta}(\beta) g_{i,\sigma_e^2}(\sigma_e^2)$. This would double the number of factors g_i we need to estimate in the algorithm.
- In the write-up of the expectation propagation algorithm, we assumed g_i was normal. Would you necessarily want $g_{i,\sigma_e^2}(\sigma_e^2)$ to be normally distributed?

Extending expectation propagation to more complicated models

- In the case of a normal regression, we could do this by assuming $g(\beta, \sigma_e^2) = \prod_{i=0}^{n} g_{i,\beta}(\beta) g_{i,\sigma_e^2}(\sigma_e^2)$. This would double the number of factors g_i we need to estimate in the algorithm.

- In the write-up of the expectation propagation algorithm, we assumed g_i was normal. Would you necessarily want $g_{i,\sigma_e^2}(\sigma_e^2)$ to be normally distributed?
  - Given that $\sigma^2 \in (0, \infty)$ and a normal random variable can take any value from $(-\infty, \infty)$, a normal distribution would be inappropriate. A log-normal might be a good choice though.

- While we will not look into such examples, there is nothing to stop you in practice from considering more complicated models. In so doing, you may want to
  - Allow factor blocks to differ by distribution $\Rightarrow$ do not feel restricted to normals.

Example of expectation propagation

- To demonstrate expectation propagation, we will return to the logistic regression example.
- As a reminder, the basic logistic regression from a Bayesian perspective corresponds to assuming a flat prior $p(\boldsymbol{\beta}) \propto 1$.
- For the purposes of comparison, we will compare the approximate inference obtained using expectation propagation to
  - fitting a standard glm $\Leftrightarrow$ Normal approximation at posterior mode.
  - Bayesian estimation using Metropolis-Hasting algorithm.
  - Where possible, exact posterior inference.
- The code required for this example is contained in a separate R markdown document.