Computing the Posterior in Probabilistic Linear Regression

Piyush Rai

CS771 Supplementary Notes/Slides

August 16, 2018

Inferring the Posterior Distribution (fully Bayesian Inference)

- Inferring the full posterior is straightforward if the hyperparams β and λ to be known/fixed
  - Basically, the conjugacy helps here (Gaussian prior is conjugate to Gaussian likelihood)
- The posterior over the weight vector $\boldsymbol{w}$ (with β and λ known)

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) = \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)}{p(\boldsymbol{y}|\mathbf{X}, \beta, \lambda)}$$

- Computing $P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ (like Bernoulli-Beta case, doing it only upto proportionality constant)

$$P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto P(\boldsymbol{w}|\lambda)P(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)$$

- After some algebra, this gets simplified into the following (proof on the next two slides)

$$\begin{aligned} P(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{(The posterior must be Gaussian due to conjugacy)} \\ \text{where} \quad \boldsymbol{\Sigma} &= (\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)^{-1} = (\beta \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma}(\beta \sum_{n=1}^{N} y_n \boldsymbol{x}_n) = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta} \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y} \end{aligned}$$

The "Completing The Square" Trick for Gaussian Posterior

- Plugging in the respective distributions for $p(\boldsymbol{w}|\lambda)$ and $p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta)$, we will get

$$
\begin{aligned}
p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda) \propto p(\boldsymbol{w}|\lambda)p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0},\lambda^{-1}\mathbf{I}_D)\mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w},\beta^{-1}\mathbf{I}_N) \\
&\propto \exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)\exp\left(-\frac{\beta}{2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right) \\
&= \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}-\frac{\beta}{2}(\boldsymbol{y}^\top\boldsymbol{y}+\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&\propto \exp\left[-\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}-\frac{\beta}{2}(\boldsymbol{w}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{w}-2\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y})\right] \\
&= \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda\mathbf{I}_D+\beta\mathbf{X}^\top\mathbf{X})\boldsymbol{w}-2\beta\boldsymbol{w}^\top\mathbf{X}^\top\boldsymbol{y}\right)\right]
\end{aligned}
$$

- We will now try to bring the exponent into a quadratic form to see if it corresponds to some Gaussian. So basically, we will use the "complete the square" trick

The "Completing The Square" Trick for Gaussian Posterior

- So we had.. $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$
- Let's see if we can bring the above posterior into the form of the following Gaussian

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right] = \exp\left[-\frac{1}{2}(\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]$$

- Let's multiply and divide $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y}\right)\right]$ by $\exp\left[-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]$
- This gives the following up to a prop. constant (remember $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is constant w.r.t. $\boldsymbol{w}$):

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{w}^\top(\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})\boldsymbol{w} - 2\beta \boldsymbol{w}^\top \mathbf{X}^\top \boldsymbol{y} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right]$$

- Finally comparing with the expression of $\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can see that

$$\begin{aligned} \boldsymbol{\Sigma} &= (\lambda \mathbf{I}_D + \beta \mathbf{X}^\top \mathbf{X})^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \beta \mathbf{X}^\top \boldsymbol{y} \quad \Rightarrow \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}(\beta \mathbf{X}^\top \boldsymbol{y}) = (\mathbf{X}^\top \mathbf{X} + \frac{\lambda}{\beta}\mathbf{I}_D)^{-1}\mathbf{X}^\top \boldsymbol{y} \end{aligned}$$

- Note: The above expression for the posterior can also be directly obtained using properties of Gaussian distributions (Refer to the maths refresher slides on "reverse conditionals", or MLAPP 4.3-4.4)