Bayesian Inference for the Gaussian Distribution

Shantanu Kodgirwar

November 11, 2020

1 Maximum likelihood estimation

Input vectors are given as $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ and the output/target variables as $\boldsymbol{t} = (t_1, \ldots, t_N)^T$ and the polynomial coefficients as $\boldsymbol{w} = (w_1, \ldots, w_M)^T$.

$$\begin{aligned} t_i &= \sum_{i=1}^{N} y(x_i, \boldsymbol{w}) && (1) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{M} y_k(x_i) w_k && (2) \\ &= (\boldsymbol{X}\boldsymbol{w})_i + n_i && (3) \end{aligned}$$

Assumed distribution of n_i:

$$n_i \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

The values of t, given the values of x follows a Gaussian distribution.

$$t_i \sim \mathcal{N}(y(x_i, \boldsymbol{w}), \sigma^2) \tag{5}$$

The likelihood function is defined as follows.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \Big] \tag{6}$$

A precision parameter β is defined which is given as $\beta^{-1} = \sigma^2$. Thus, we have the following modified likelihood function as follows.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\Big[-\frac{\beta}{2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \Big] \tag{7}$$

Assuming the data is drawn independently, the likelihood is the joint probability given as the product of individual marginal probabilities. It is also assumed the value of β is known or assumed.

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \tag{8}$$

The log likelihood of equation 8 is given as

$$\begin{aligned}
\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) && (9)\\
&= \sum_{i=1}^{N} \ln \left\{ \sqrt{\frac{\beta}{2\pi}} \exp \left[\frac{-\beta}{2} \left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)^2 \right] \right\} && (10)\\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \sum_{i=1}^{N} \left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)^2 && (11)\\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \left\| \boldsymbol{t} - \boldsymbol{X}\boldsymbol{w} \right\|^2 && (12)
\end{aligned}$$

The posterior probability to determine the parameters is given as the product of likelihood function and prior.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w}) \tag{13}$$

The value of the prior $p(\boldsymbol{w}) = 1$.

By maximizing the negative likelihood (or posterior distribution with prior as one) with respect to $\boldsymbol{w}$,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}} \left\{ - \ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \right\} &\stackrel{!}{=} 0 && (14)\\
\frac{\partial}{\partial \boldsymbol{w}} \left\{ - \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \right\} &\stackrel{!}{=} 0 && (15)\\
\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \left\| \boldsymbol{t} - \boldsymbol{X}\boldsymbol{w} \right\|^2 \right\} &\stackrel{!}{=} 0 && (16)\\
\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \left(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right)^T \left(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right) \right\} &\stackrel{!}{=} 0 && (17)\\
\beta \left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t} \right) &\stackrel{!}{=} 0 && (18)
\end{aligned}$$

Therefore, $\boldsymbol{w}_{\mathrm{ML}}$ is evaluated.

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{19}$$

It can be seen that the maximum likelihood results into least square estimator. We can similarly estimate β_{ML} by maximizing the posterior with respect to β. The known value of $\boldsymbol{w}_{\text{ML}}$ can now be utilized here.

Taking the log likelihood in equation 9, the following could be shown.

$$\frac{\partial}{\partial \beta} \{ - \ln p(\boldsymbol{w}_{\text{ML}} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \} \stackrel{!}{=} 0 \tag{20}$$

$$\frac{\partial}{\partial \beta} \{ - \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}_{\text{ML}}, \beta) \} \stackrel{!}{=} 0 \tag{21}$$

$$\frac{\partial}{\partial \beta} \left\{ -\frac{N}{2} \ln \beta + \frac{\beta}{2} \| \boldsymbol{t} - \boldsymbol{X} \boldsymbol{w}_{\text{ML}} \|^2 \right\} \stackrel{!}{=} 0 \tag{22}$$

$$\| \boldsymbol{t} - \boldsymbol{X} \boldsymbol{w}_{\text{ML}} \|^2 - \frac{N}{\beta} \stackrel{!}{=} 0 \tag{23}$$

Therefore, the value β_{ML} is determined to be

$$\beta_{\text{ML}} = \frac{N}{\| \boldsymbol{t} - \boldsymbol{X} \boldsymbol{w}_{\text{ML}} \|^2} \tag{24}$$

2 Maximum a posteriori estimation

In the case of maximum a posteriori (MAP) estimation, the distribution of prior over parameters is known.

2.1 Gaussian distribution of prior

The prior distribution is given as follows

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/\alpha) = \left(\frac{\alpha}{2\pi} \right)^{M/2} \exp \left\{ -\frac{\alpha}{2} \| \boldsymbol{w} \|^2 \right\} \tag{25}$$

The posterior distribution is shown as follows.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \alpha) \tag{26}$$

where β as defined earlier is the precision parameter of the likelihood, α is the hyperparameter (also a precision parameter of the prior distribution) which controls the distribution of model parameters. It is assumed that the value of α and β is known.

The log of the posterior is given as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\boldsymbol{w} \mid \alpha) \tag{27}$$

The log likelihood is known from equation 9. Therefore,

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{M}{2} \ln \left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2} \|\boldsymbol{w}\|^2 \tag{28}$$

Maximizing the negative log of posterior with respect to $\boldsymbol{w}$.

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)\} \stackrel{!}{=} 0 \tag{29}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) \right] + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} \right\} \stackrel{!}{=} 0 \tag{30}$$

$$\beta \left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t} \right) + \alpha \boldsymbol{w} \stackrel{!}{=} 0 \tag{31}$$

Let us assign a regularization parameter $\lambda = \alpha/\beta$. Therefore, the value of parameter using maximum a posteriori estimation $\boldsymbol{w}_{\text{MAP}}$ is given as

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{32}$$

2.2 Jeffreys prior

2.2.1 Evaluation for precision parameter β

Jeffreys prior is given as

$$p(\sigma) = \frac{1}{\sigma} \tag{33}$$

Parameter transformation: $\sigma \rightarrow \beta = 1/\sigma^2$

$$\begin{aligned} p_\beta(\beta) &= p_\sigma(\sigma) \left|_{\sigma=1/\sqrt{\beta}} \left| \frac{d\sigma}{d\beta} \right| \right. \\ &\propto \sqrt{\beta} \frac{1}{1/\sigma^3} \left|_{\sigma=1/\sqrt{\beta}} \right. \\ &= \sqrt{\beta} \sigma^3 \left|_{\sigma=1/\sqrt{\beta}} \right. \\ &= \sqrt{\beta} \beta^{-3/2} = 1/\beta \end{aligned}$$

Therefore,

$$p(\beta) = \frac{1}{\beta} \tag{34}$$

The posterior distribution is given as

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\beta) \tag{35}$$

The log likelihood of the posterior distribution is given as

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\beta) \tag{36}$$

Using equation 9 to give the log likelihood, the above equation would be as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \ln \beta \tag{37}$$

Maximizing the log posterior with respect to $\boldsymbol{w}$,

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{38}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 \right\} \stackrel{!}{=} 0 \tag{39}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) \right\} \stackrel{!}{=} 0 \tag{40}$$

$$\beta \left(\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t} \right) \stackrel{!}{=} 0 \tag{41}$$

Therefore, we find that, $\boldsymbol{w}_{\text{MAP}}$ is

$$\boldsymbol{w}_{\text{MAP}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{42}$$

We can see that $\boldsymbol{w}_{\text{MAP}} = \boldsymbol{w}_{\text{ML}}$ (refer equation 19).
Now, maximizing the log posterior with respect to β, equation 37 is used.

$$\frac{\partial}{\partial \beta} \{-\ln p(\boldsymbol{w}_{\text{MAP}} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{43}$$

$$\frac{\partial}{\partial \beta} \left\{ \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 \right\} - \frac{\partial}{\partial \beta} \left\{ \frac{N}{2} \ln \beta \right\} + \frac{\partial}{\partial \beta} \{\ln \beta\} \stackrel{!}{=} 0 \tag{44}$$

$$\frac{1}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{N}{2\beta} + \frac{1}{\beta} \stackrel{!}{=} 0 \tag{45}$$

Therefore,

$$\beta_{\text{MAP}} = \frac{N - 2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2} \tag{46}$$

2.2.2 Evaluation for hyperparameter α

Jeffreys prior for the hyperparameter α is given as follows,

$$p(\alpha) = \frac{1}{\alpha} \tag{47}$$

The joint distribution $p(w, \alpha)$ is given as follows

$$p(\boldsymbol{w}, \alpha) = p(\boldsymbol{w} \mid \alpha) p(\alpha) \tag{48}$$

The distribution $p(\boldsymbol{w} \mid \alpha)$ is known from equation 25. Therefore, the joint probability $p(w, \alpha)$ is evaluated as follows.

$$p(\boldsymbol{w}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} \times \frac{1}{\alpha} \tag{49}$$

$$p(\boldsymbol{w}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{(M-2)/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} \tag{50}$$

The posterior distribution is given as follows.

$$p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta) p(\boldsymbol{w}, \alpha) p(\beta) \tag{51}$$

The distribution of $p(\beta)$ is known from equation 34 to be $p(\beta) = 1/\beta$. Therefore, the log of the posterior distribution is

$$\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta) + \ln p(\boldsymbol{w}, \alpha) + \ln p(\beta) \tag{52}$$

Evaluating the above equation results to the following.

$$\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) = \frac{N}{2} \ln \beta - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{(M-2)}{2} \ln \alpha - \frac{\alpha}{2} \|\boldsymbol{w}\|^2 - \ln \beta \tag{53}$$

Initially, maximizing the posterior (refer equation 53) with respect to $\boldsymbol{w}$, we get

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X})\} \stackrel{!}{=} 0 \tag{54}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{\frac{\beta}{2} \left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})\right] + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w}\right\} \stackrel{!}{=} 0 \tag{55}$$

$$\beta \left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t}\right) + \alpha \boldsymbol{w} \stackrel{!}{=} 0 \tag{56}$$

Given the regularization parameter $\lambda = \alpha/\beta$, $\boldsymbol{w}_{\text{MAP}}$ is the same as per the equation 32.

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{57}$$

Finally, maximizing posterior(refer equation 53) with respect to β, we get the following

$$\frac{\partial}{\partial \beta} \{-\ln p(\boldsymbol{w}_{\text{MAP}}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X})\} \stackrel{!}{=} 0 \tag{58}$$

$$\frac{\partial}{\partial \beta} \left\{\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2\right\} - \frac{\partial}{\partial \beta} \left\{\frac{N}{2} \ln \beta\right\} + \frac{\partial}{\partial \beta} \{\ln \beta\} \stackrel{!}{=} 0 \tag{59}$$

$$\frac{1}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{N}{2\beta} + \frac{1}{\beta} \stackrel{!}{=} 0 \tag{60}$$

Therefore, β_{MAP} is as follows.

$$\beta_{\text{MAP}} = \frac{N-2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2} \tag{61}$$

It can also be seen that $\beta_{\text{MAP}} = \beta_{\text{ML}}$ (refer equation 24).

Maximizing the posterior (refer equation 53) with respect to α, we get the following.

$$\frac{\partial}{\partial\alpha}\{-\ln p(\boldsymbol{w}_{\text{MAP}}, \alpha, \beta_{\text{MAP}} \mid \boldsymbol{t}, \boldsymbol{X})\} \stackrel{!}{=} 0 \tag{62}$$

$$\|\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{M-2}{\alpha} \stackrel{!}{=} 0 \tag{63}$$

Therefore, α_{MAP} is given as follows.

$$\alpha_{\text{MAP}} = \frac{M-2}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{64}$$

3 Gamma Distribution

The likelihood function for the target values given β (precision parameter) can be written as

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta^{-1}) \tag{65}$$

$$\propto \beta^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2\right\} \tag{66}$$

The corresponding conjugate prior[1] should be proportional to the product of the power of β and the exponential of the linear function of β.

This corresponds to the gamma distribution which is given as follows.

$$\text{Gam}(\beta \mid a, b) = \frac{b^a}{\Gamma(a)}\beta^{a-1}\exp(-b\beta) \tag{67}$$

where, $\Gamma(a)$ is a gamma function that ensures the gamma distribution is properly normalized, a is called as the shape parameter and b is called as the rate parameter.

The expectation $\mathbb{E}[\beta]$ is given as follows.

$$\mathbb{E}[\beta] = \frac{a}{b} \tag{68}$$

[1] posterior distribution is in the same probability distribution family as the prior distribution

The mode mode[β] which is equivalent to maximizing the posterior with respect to β is given as follows.

$$\text{mode}[\beta] = \frac{a-1}{b} \tag{69}$$

Let us consider the posterior distribution considering the jeffreys prior for $p(\beta)$ as given in equation 34.

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) &\propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\beta) && (70) \\ &\propto \beta^{\frac{N}{2}} \exp\left\{ -\frac{\beta}{2} \sum_{i=1}^{N} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \right\} \times \frac{1}{\beta} && (71) \\ &\propto \beta^{\frac{N}{2}-1} \exp\left\{ -\frac{\beta}{2} \sum_{i=1}^{N} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \right\} && (72) \end{aligned}$$

By comparing this to the gamma distribution in equation 67, mode[$\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}$] is as follows

$$\text{mode}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}] = \frac{N-2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2} \tag{73}$$

This is equivalent to β_{MAP} from the result in equation 46. Similarly, $\mathbb{E}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}]$ is given as follows.

$$\mathbb{E}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}] = \frac{N}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2} \tag{74}$$

This can also be shown to hold true for the joint distribution $p(\boldsymbol{w}, \alpha)$ in equation 48.

$$p(\boldsymbol{w}, \alpha) \propto \alpha^{\frac{M}{2}-1} \exp\left\{ -\frac{\alpha}{2} \|\boldsymbol{w}\|^2 \right\} \tag{75}$$

$$\text{mode}[\alpha \mid \boldsymbol{w}] = \frac{M-2}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{76}$$

This is same as α_{MAP} in equation 64. Similarly $\mathbb{E}[\alpha \mid \boldsymbol{w}]$ is

$$\mathbb{E}[\alpha \mid \boldsymbol{w}] = \frac{M}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{77}$$