Bayesian Curve Fitting with Gaussian Distribution

1 Maximum likelihood estimation

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

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

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.

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

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

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

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.

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \tag{7}$$

The log likelihood of equation 7 is given as

$$\begin{aligned}
\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) && (8)\\
&= \sum_{i=1}^{N} \ln \left\{ \sqrt{\frac{\beta}{2\pi}} \exp\left[\frac{-\beta}{2}\left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)^2\right]\right\} && (9)\\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\sum_{i=1}^{N}\left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)^2 && (10)\\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\left\|\left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)\right\|^2 && (11)
\end{aligned}$$

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

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w}) \tag{12}$$

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

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

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}}\left\{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\right\} &\stackrel{!}{=} 0 && (13)\\
\frac{\partial}{\partial \boldsymbol{w}}\left\{-\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta)\right\} &\stackrel{!}{=} 0 && (14)\\
\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\left\|\left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)\right\|^2\right\} &\stackrel{!}{=} 0 && (15)\\
\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\left(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right)^T\left(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right)\right\} &\stackrel{!}{=} 0 && (16)\\
\beta\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{t}\right) &\stackrel{!}{=} 0 && (17)
\end{aligned}$$

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

$$\boldsymbol{w}_{ML} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{18}$$

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 W_{ML} can now be utilized here.

Taking the log likelihood in equation 8, the following could be shown.

$$\frac{\partial}{\partial\beta}\left\{-\ln p(\boldsymbol{w}_{ML} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\right\} \stackrel{!}{=} 0 \tag{19}$$

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

$$\frac{\partial}{\partial\beta}\left\{\frac{N}{2}\ln\beta - \frac{\beta}{2}\left\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{ML})_i)\right\|^2\right\} \stackrel{!}{=} 0 \tag{21}$$

$$\frac{N}{\beta} - \left\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{ML})_i)\right\|^2 \stackrel{!}{=} 0 \tag{22}$$

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

$$\beta_{ML}^{-1} = \frac{1}{N}\left\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{ML})_i)\right\|^2 \tag{23}$$

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

$$p(\boldsymbol{w} \mid \alpha) \propto p(\alpha \mid \boldsymbol{w})p(\boldsymbol{w}) \tag{24}$$

However, $p(\boldsymbol{w} \mid \alpha)$ is known as follows.

$$p(\boldsymbol{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\left\|\boldsymbol{w}\right\|^2\right\} \tag{25}$$

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 8. Therefore,

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\left\|(t_i - (\boldsymbol{X}\boldsymbol{w})_i)\right\|^2 + \frac{M+1}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\left\|\boldsymbol{w}\right\|^2 \tag{28}$$

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 $\boldsymbol{w}_{MAP}$ is given as

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

2.2 Jeffreys prior

2.2.1 Evaluation for precision parameter β

Jeffreys prior is given as

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

Under the assumption of $\beta^{-1} = \sigma^2$,

$$p(\beta) = \beta^{-1/2} \tag{34}$$

Therefore, the prior distribution $p(\sigma \mid \boldsymbol{w})$ is proportional to Jeffreys prior.

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

where, $p(\boldsymbol{w} \mid \sigma)$ will be some initial distribution (example: a vector of ones of size M+1).

Therefore,

$$p(\beta \mid \boldsymbol{w}) = \beta^{-1/2} \tag{36}$$

The posterior distribution is given as

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \beta) \tag{37}$$

The log likelihood of the posterior distribution is given as

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\boldsymbol{w} \mid \beta) \tag{38}$$

Using equation 8 to give the log likelihood, the above equation 39 would be as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\|(t_i - (\boldsymbol{X}\boldsymbol{w})_i)\|^2 - \frac{1}{2}\ln\beta \tag{39}$$

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

$$\frac{\partial}{\partial \boldsymbol{w}}\{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{40}$$

$$\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\|(t_i - (\boldsymbol{X}\boldsymbol{w})_i)\|^2\right\} \stackrel{!}{=} 0 \tag{41}$$

$$\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})\right\} \stackrel{!}{=} 0 \tag{42}$$

$$\beta\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{t}\right) \stackrel{!}{=} 0 \tag{43}$$

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

$$\boldsymbol{w}_{MAP} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{44}$$

We can see that $w_{MAP} = w_{ML}$ (refer equation 18).

Now, maximizing the log posterior with respect to β, equation 39 is used.

$$\frac{\partial}{\partial\beta}\{-\ln p(\boldsymbol{w}_{MAP} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{45}$$

$$\frac{\partial}{\partial\beta}\left\{\frac{\beta}{2}\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{MAP})_i)\|^2\right\} - \frac{\partial}{\partial\beta}\left\{\frac{N}{2}\ln\beta\right\} + \frac{\partial}{\partial\beta}\left\{\frac{1}{2}\ln\beta\right\} \stackrel{!}{=} 0 \tag{46}$$

$$\frac{N}{\beta} - \frac{1}{\beta} - \frac{1}{2}\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{MAP})_i)\|^2 \stackrel{!}{=} 0 \tag{47}$$

Therefore,

$$\beta_{MAP}^{-1} = \frac{1}{(N-1)}\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{MAP})_i)\|^2 \tag{48}$$

2.2.2 Evaluation for hyperparameter α

Jeffreys prior for the hyperparameter α is given as follows,

$$p(\alpha) = \frac{1}{\alpha} \tag{49}$$

The prior distribution $p(\alpha \mid \boldsymbol{w})$ is given as follows,

$$p(\alpha \mid \boldsymbol{w}) \propto p(\boldsymbol{w} \mid \alpha)p(\alpha) \tag{50}$$

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

$$\begin{aligned} p(\alpha \mid \boldsymbol{w}) &= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\|\boldsymbol{w}\|^2\right\} \times \frac{1}{\alpha} && (51) \\ p(\alpha \mid \boldsymbol{w}) &= \left(\frac{\alpha}{2\pi}\right)^{(M)/2} \exp\left\{-\frac{\alpha}{2}\|\boldsymbol{w}\|^2\right\} && (52) \end{aligned}$$

The posterior distribution is given as

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta)p(\alpha \mid \boldsymbol{w}) \tag{53}$$

The log likelihood of the posterior distribution is

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta) + \ln p(\alpha \mid \boldsymbol{w}) \tag{54}$$

Using equation 8 to give the log likelihood and the log of the prior $p(\alpha \mid \boldsymbol{w})$, the equation becomes

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\|(t_i - (\boldsymbol{X}\boldsymbol{w})_i)\|^2 + \frac{M}{2}\ln\alpha - \frac{M}{2}\ln 2\pi - \frac{\alpha}{2}\|\boldsymbol{w}\|^2 \tag{55}$$

Initially, maximizing the posterior (refer equation 55) with respect to $\boldsymbol{w}$, we get

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}}\{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)\} &\overset{!}{=} 0 && (56) \\ \frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})\right] + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\} &\overset{!}{=} 0 && (57) \\ \beta\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{t}\right) + \alpha\boldsymbol{w} &\overset{!}{=} 0 && (58) \end{aligned}$$

Therefore, w_{MAP} is the same as per the equation 32.

$$w_{MAP} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{59}$$

where, $\lambda = \alpha/\beta$.

Finally, maximizing posterior(refer equation 55) with respect to β, we get the following

$$\frac{\partial}{\partial \beta}\{-\ln p(\boldsymbol{w}_{MAP} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha_{MAP}, \beta)\} \stackrel{!}{=} 0 \tag{60}$$

$$\frac{\partial}{\partial \beta}\left\{\frac{N}{2}\ln \beta - \frac{\beta}{2}\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{MAP})_i)\|^2\right\} \stackrel{!}{=} 0 \tag{61}$$

$$\frac{N}{\beta} - \|(t_i - (\boldsymbol{X}\boldsymbol{w}_{MAP})_i)\|^2 \stackrel{!}{=} 0 \tag{62}$$

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

$$\beta_{MAP}^{-1} = \frac{1}{N}\|(t_i - (\boldsymbol{X}\boldsymbol{w}_{MAP})_i)\|^2 \tag{63}$$

It can also be seen that $\beta_{MAP} = \beta_{ML}$ (refer equation 23).
Maximizing the posterior (refer equation 55) with respect to α, we get

$$\frac{\partial}{\partial \alpha}\{-\ln p(\boldsymbol{w}_{MAP} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)\} \stackrel{!}{=} 0 \tag{64}$$

$$\|\boldsymbol{w}_{MAP}\|^2 - \frac{M}{\alpha} \stackrel{!}{=} 0 \tag{65}$$

Therefore, α_{MAP} is given as follows.

$$\alpha_{MAP} = \frac{M}{\|\boldsymbol{w}_{MAP}\|^2} \tag{66}$$