

Bayesian Inference for the Gaussian Distribution

Shantanu Kodgirwar

November 11, 2020

1 Maximum likelihood estimation

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

$$t_i = \sum_{i=1}^N y(x_i, \mathbf{w}) \quad (1)$$

$$= \sum_{i=1}^N \sum_{k=1}^M y_k(x_i) w_k \quad (2)$$

$$= (\mathbf{X}\mathbf{w})_i + n_i \quad (3)$$

Assumed distribution of n_i :

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

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

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

The likelihood function is defined as follows.

$$p(t_i | \mathbf{X}, \mathbf{w}, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (t_i - (\mathbf{X}\mathbf{w})_i)^2 \right] \quad (6)$$

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

$$p(t_i | \mathbf{X}, \mathbf{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp \left[-\frac{\beta}{2} (t_i - (\mathbf{X}\mathbf{w})_i)^2 \right] \quad (7)$$

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.

$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^N p(t_i | \mathbf{X}, \mathbf{w}, \beta) \quad (8)$$

The log likelihood of equation 8 is given as

$$\ln p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^N \ln p(t_i | \mathbf{X}, \mathbf{w}, \beta) \quad (9)$$

$$= \sum_{i=1}^N \ln \left\{ \sqrt{\frac{\beta}{2\pi}} \exp \left[-\frac{\beta}{2} (t_i - (\mathbf{X}\mathbf{w})_i)^2 \right] \right\} \quad (10)$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \sum_{i=1}^N (t_i - (\mathbf{X}\mathbf{w})_i)^2 \quad (11)$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 \quad (12)$$

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

$$p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \beta) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}) \quad (13)$$

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

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

$$\frac{\partial}{\partial \mathbf{w}} \{-\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \beta)\} \stackrel{!}{=} 0 \quad (14)$$

$$\frac{\partial}{\partial \mathbf{w}} \{-\ln p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta)\} \stackrel{!}{=} 0 \quad (15)$$

$$\frac{\partial}{\partial \mathbf{w}} \left\{ \frac{\beta}{2} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 \right\} \stackrel{!}{=} 0 \quad (16)$$

$$\frac{\partial}{\partial \mathbf{w}} \left\{ \frac{\beta}{2} (\mathbf{t} - \mathbf{X}\mathbf{w})^T (\mathbf{t} - \mathbf{X}\mathbf{w}) \right\} \stackrel{!}{=} 0 \quad (17)$$

$$\beta (\mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{t}) \stackrel{!}{=} 0 \quad (18)$$

Therefore, \mathbf{w}_{ML} is evaluated.

$$\mathbf{w}_{\text{ML}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{t} \quad (19)$$

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 \mathbf{w}_{ML} can now be utilized here.

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

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

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

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

$$\|\mathbf{t} - \mathbf{X} \mathbf{w}_{\text{ML}}\|^2 - \frac{N}{\beta} \stackrel{!}{=} 0 \quad (23)$$

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

$$\beta_{\text{ML}} = \frac{N}{\|\mathbf{t} - \mathbf{X} \mathbf{w}_{\text{ML}}\|^2} \quad (24)$$

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

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

The posterior distribution is shown as follows.

$$p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w} | \alpha) \quad (26)$$

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

$$\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta) = \ln p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) + \ln p(\mathbf{w} | \alpha) \quad (27)$$

The log likelihood is known from equation 9. Therefore,

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

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

$$\frac{\partial}{\partial \mathbf{w}} \{-\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta)\} \stackrel{!}{=} 0 \quad (29)$$

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

$$\beta (\mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{t}) + \alpha \mathbf{w} \stackrel{!}{=} 0 \quad (31)$$

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

$$\mathbf{w}_{\text{MAP}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{t} \quad (32)$$

2.2 Jeffreys prior

2.2.1 Evaluation for precision parameter β

Jeffreys prior is given as

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

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

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

Therefore,

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

The posterior distribution is given as

$$p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \beta) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) p(\beta) \quad (35)$$

The log likelihood of the posterior distribution is given as

$$\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \beta) = \ln p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) + \ln p(\beta) \quad (36)$$

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

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

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

$$\frac{\partial}{\partial \mathbf{w}} \{-\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \beta)\} \stackrel{!}{=} 0 \quad (38)$$

$$\frac{\partial}{\partial \mathbf{w}} \left\{ \frac{\beta}{2} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 \right\} \stackrel{!}{=} 0 \quad (39)$$

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

$$\beta (\mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{t}) \stackrel{!}{=} 0 \quad (41)$$

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

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

We can see that $\mathbf{w}_{\text{MAP}} = \mathbf{w}_{\text{ML}}$ (refer equation 19).

Now, maximizing the log posterior with respect to β , equation 37 is used.

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

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

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

Therefore,

$$\beta_{\text{MAP}} = \frac{N - 2}{\|\mathbf{t} - \mathbf{X}\mathbf{w}_{\text{MAP}}\|^2} \quad (46)$$

2.2.2 Evaluation for hyperparameter α

Jeffreys prior for the hyperparameter α is given as follows,

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

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

$$p(\mathbf{w}, \alpha) = p(\mathbf{w} | \alpha)p(\alpha) \quad (48)$$

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

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

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

The posterior distribution is given as follows.

$$p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \alpha, \beta)p(\mathbf{w}, \alpha)p(\beta) \quad (51)$$

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

$$\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta) = \ln p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \alpha, \beta) + \ln p(\mathbf{w}, \alpha) + \ln p(\beta) \quad (52)$$

Evaluating the above equation results to the following.

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

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

$$\frac{\partial}{\partial \mathbf{w}} \{-\ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \alpha, \beta)\} \stackrel{!}{=} 0 \quad (54)$$

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

$$\beta (\mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{t}) + \alpha \mathbf{w} \stackrel{!}{=} 0 \quad (56)$$

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

$$\mathbf{w}_{\text{MAP}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{t} \quad (57)$$

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

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

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

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

Therefore, β_{MAP} is as follows.

$$\beta_{\text{MAP}} = \frac{N - 2}{\|\mathbf{t} - \mathbf{X}\mathbf{w}_{\text{MAP}}\|^2} \quad (61)$$

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.

$$\frac{\partial}{\partial \alpha} \{-\ln p(\mathbf{w}_{\text{MAP}} | \mathbf{t}, \mathbf{X}, \alpha, \beta)\} \stackrel{!}{=} 0 \quad (62)$$

$$\|\mathbf{w}_{\text{MAP}}\|^2 - \frac{M - 2}{\alpha} \stackrel{!}{=} 0 \quad (63)$$

Therefore, α_{MAP} is given as follows.

$$\alpha_{\text{MAP}} = \frac{M - 2}{\|\mathbf{w}_{\text{MAP}}\|^2} \quad (64)$$

3 Gamma Distribution

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

$$p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^N \mathcal{N}(t_i | \mathbf{X}, \mathbf{w}, \beta^{-1}) \quad (65)$$

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

The corresponding conjugate prior¹ 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.

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

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.

$$\mathbb{E}[\beta] = \frac{a}{b} \quad (68)$$

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

The mode $\text{mode}[\beta]$ which is equivalent to maximizing the posterior with respect to β is given as follows.

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

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

$$p(\mathbf{w} | \mathbf{t}, \mathbf{X}, \beta) \propto p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) p(\beta) \quad (70)$$

$$\propto \beta^{\frac{N}{2}} \exp \left\{ -\frac{\beta}{2} \sum_{i=1}^N (t_i - (\mathbf{X}\mathbf{w})_i)^2 \right\} \times \frac{1}{\beta} \quad (71)$$

$$\propto \beta^{\frac{N}{2}-1} \exp \left\{ -\frac{\beta}{2} \sum_{i=1}^N (t_i - (\mathbf{X}\mathbf{w})_i)^2 \right\} \quad (72)$$

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

$$\text{mode}[\beta | \mathbf{t}, \mathbf{X}, \mathbf{w}] = \frac{N-2}{\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2} \quad (73)$$

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

$$\mathbb{E}[\beta | \mathbf{t}, \mathbf{X}, \mathbf{w}] = \frac{N}{\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2} \quad (74)$$

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

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

$$\text{mode}[\alpha | \mathbf{w}] = \frac{M-2}{\|\mathbf{w}_{\text{MAP}}\|^2} \quad (76)$$

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

$$\mathbb{E}[\alpha | \mathbf{w}] = \frac{M}{\|\mathbf{w}_{\text{MAP}}\|^2} \quad (77)$$