[EECM30060] Machine Learning

Homework 1

Po-Chuan, Chen
Student ID: 311511052
`present90308.ee11@nycu.edu.tw`

National Yang Ming Chiao Tung University

October 14, 2022

Contents

1 Bayesian Linear Regression

1.1 Question

A linear regression function can be expressed as below where the $\boldsymbol{\phi}$ is a basis function:

$$y(x, \mathbf{w}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathrm{x})$$

In order to make prediction of t for new test data x from the learned $\mathbf{w}$, we may

- multiply the likelihood function of new data $p(t|x, \mathbf{w})$and the posterior distribution of training set with label set.
- take the integral over $\mathbf{w}$ to find the predictive distribution

$$\begin{aligned} p(t|x, \mathbf{x}, \mathbf{t}) &= \int_{-\infty}^{\infty} p(t, \mathbf{w}|\mathrm{x}, \mathbf{x},\ \mathbf{t})\mathrm{d}\mathbf{w} \\ &= \int_{-\infty}^{\infty} p(t|\mathbf{w}, \mathrm{x}, \mathbf{x},\ \mathbf{t})\mathrm{p}(\mathbf{w}|\mathrm{x}, \mathbf{x},\ \mathbf{t})\mathrm{d}\mathbf{w} \\ &= \int_{-\infty}^{\infty} p(t|x, \mathbf{w})\mathrm{p}(\mathbf{w}|\mathbf{x},\ \mathbf{t})\mathrm{d}\mathbf{w}. \end{aligned}$$

Prove that the predictive distribution just mentioned is the same with the form

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(\mathrm{t}|\mathrm{m}(\mathrm{x}), \mathrm{s}^2(\mathrm{x}))$$

where

$$\begin{aligned} m(x) &= \beta \boldsymbol{\phi}(x)^\top \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \\ s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^\top \mathbf{S} \boldsymbol{\phi}(x). \end{aligned}$$

1.2 Proof

Based on the marginal and conditional Gaussian, a marginal Gaussian distribution for **x** and conditional
Gaussian distribution for **y** given **x** in the form

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \mu, \Lambda^{-1}) \tag{1}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \tag{2}$$

and the marginal distribution of **y** and the conditional distribution of **x** given **y** are given by

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^{T}) \tag{3}$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x} \mid \sum\{\boldsymbol{A}^{T}\boldsymbol{L}(\boldsymbol{y}\text{-}\boldsymbol{b}) + \Lambda\boldsymbol{\mu}\}, \sum) \tag{4}$$

where

$$\sum = (\Lambda + \boldsymbol{A}^{T}\boldsymbol{L}\boldsymbol{A})^{-1}$$

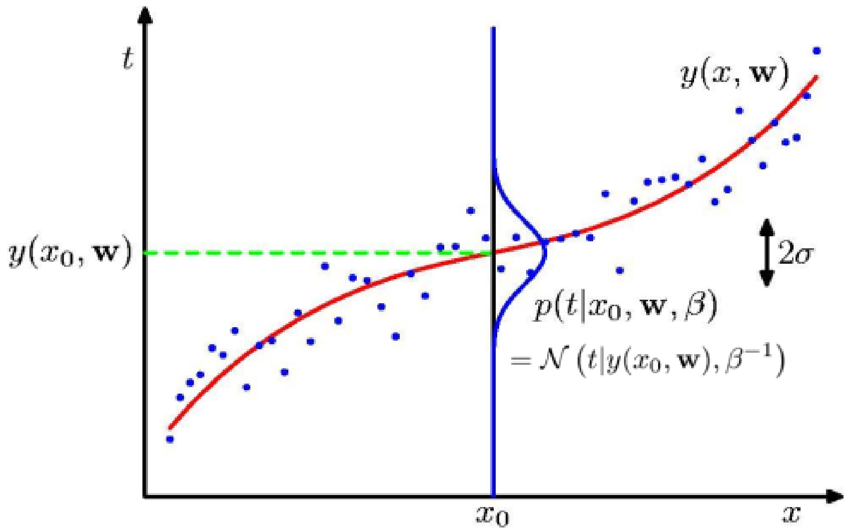

Figure 1: Likelihood model

We set the likelihood as

$$p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1}) = \mathcal{N}(t \mid \boldsymbol{\phi}(x)^{T}\boldsymbol{w}, \beta^{-1})$$

and the posterior as

$$p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = \mathcal{N}(\boldsymbol{w} \mid \beta\boldsymbol{S}_N \sum_{n=1}^{N} t_n\boldsymbol{\phi}\ (x_n), \boldsymbol{S}_N).$$

Based on the problem, we can take the predictive distribution to

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\Lambda^{-1}\boldsymbol{A}^T) \Longrightarrow$$

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t \mid \boldsymbol{\phi}(x)^T \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x))$$

which

$$\boldsymbol{x} \leftarrow \boldsymbol{w}, \mu = \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \Lambda^{-1} = \boldsymbol{S}_N$$

$$\boldsymbol{y} \leftarrow t, \boldsymbol{A} = \boldsymbol{\phi}(x)^T, \boldsymbol{b} = 0, \boldsymbol{L}^{-1} = \beta^{-1}$$

and S_N is covariance matrix with N dimension, α which is precision of prior, and β is precision in the data.

If we want to calculate the predictive distribution in full Bayesian treatment with training data $\mathbf{x}$ and $\mathbf{t}$ and a new test point x, the distribution will turn into

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t \mid x, \boldsymbol{w}) p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{w}, \; p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1}) \Longrightarrow$$

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t \mid m(x), s^2(x))$$

where the mean $m(x)$ and the variance $s^2(x)$ is like

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^T$$

$$\boldsymbol{\phi}(x_n) = \left\{ \begin{array}{l} 1 \\ x_n \\ x_n^2 \\ \vdots \\ x_n^{M-1} \end{array} \right\} \text{ with polynomial regression, } \boldsymbol{I} = \textit{unit matrix } M \times M$$

and then we can use predictive distribution to predict t based on the new coming $\boldsymbol{x}$.

And the distribution can be transfer into

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t \mid m(x), \sigma_N^2(x))$$

where the mean and variance is

$$m(x) = \boldsymbol{\phi}(x)^T m_N,\ m_N = \beta \boldsymbol{S}_N \Phi^T t$$

$$\sigma_N^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x) \xrightarrow{N\to\infty} \beta^{-1} \textit{ and } \sigma_{N+1}^2(x) \leq \sigma_N^2(x) \tag{5}$$

$$\boldsymbol{S}_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

As we can see in the equation (5), we can know that

$$\boldsymbol{S}_{N+1}^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N+1} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T = \boldsymbol{S}_N^{-1} + \beta \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T$$

, note that we can see $\alpha \to 0$ implies $m_N \to w = (\boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \boldsymbol{t}$, and

$$(\boldsymbol{M} + \boldsymbol{v}\ \boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T \boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T \boldsymbol{M}^{-1} \boldsymbol{v}}$$

So, based on the results we get, we can know that

$$\begin{aligned}
\sigma_{N+1}^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_{N+1} \boldsymbol{\phi}(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T (\boldsymbol{S}_N^{-1} + \beta \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T) \\
&= \beta^{-1} + \boldsymbol{\phi}(x)^T [\boldsymbol{S}_N - \frac{\beta(\boldsymbol{S}_N \boldsymbol{\phi}(x_n))(\boldsymbol{\phi}(x_n)^T \boldsymbol{S}_N)}{1 + \beta \boldsymbol{\phi}(x_n)^T \boldsymbol{S}_N} \boldsymbol{\phi}(x_n)] \boldsymbol{\phi}(x) \\
&= \sigma_N^2(x) - \frac{\beta(\boldsymbol{\phi}(x_n)^T \boldsymbol{S}_N \boldsymbol{\phi}(x))^2}{1 + \beta \boldsymbol{\phi}(x_n)^T \boldsymbol{S}_N \boldsymbol{\phi}(x_n)} \leq \sigma_N^2(x)
\end{aligned}$$

So the predictive distribution used as follows

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t \mid x, \boldsymbol{w}) p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{w},\ p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1}) \Longrightarrow$$

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t \mid m(x), \sigma^2(x))$$

where the mean $m(x)$ and the variance $\sigma^2(x)$ is like

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$\sigma^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^T$$

$$\boldsymbol{\phi}(x_n) = \left\{ \begin{array}{l} \phi_0(x_n) \\ \phi_1(x_n) \\ \phi_2(x_n) \\ \vdots \\ \phi_{M-1}(x_n) \end{array} \right\}, \boldsymbol{I} = unit\ matrix\ M \times M,$$

$$\boldsymbol{\phi}(x_n) = \left\{ \begin{array}{l} 1 \\ x_n \\ x_n^2 \\ \vdots \\ x_n^{M-1} \end{array} \right\} \text{with polynomial regression}$$

which β^{-1} is noise in data, $\boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)$ is an uncertainty in $\boldsymbol{w}$.