Lecture 03. Bayesian Linear Regression

Lecturer: Jie Wang Date: Sep 20, 2022

The major reference of this lecture is [1].

1 Introduction

We have studied the linear regression from the perspectives of least squares and maximum likelihood. In this lecture, we shall study linear regression from a quite different approach, that is, Bayesian linear regression.

2 The Problem Settings

Regression aims at predicting the value of one or more *continuous* target variables Y given a set of observed (input/control) variables $X \in \mathbb{R}^D$.

Linear regression is to model the relation between the input features $X \in \mathbb{R}^D$ and its corresponding response $Y \in \mathbb{R}$ by a linear model:

$$Y = f(X; \mathbf{w}) = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_D X_D + \epsilon, \tag{1}$$

where $X = (X_1, X_2, \ldots, X_D)^\top$, $\mathbf{w} = (w_0, w_1, \ldots, w_D)^\top \in \mathbb{R}^{D+1}$, and

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \tag{2}$$

To simplify the model in (1), we can introduce a dummy variable $X_0 = 1$ and define $\bar{X} = (X_0, X_1, X_2, \ldots, X_D)^\top = (1, X_1, X_2, \ldots, X_D)^\top$. Then, the model in (1) becomes

$$Y = f(X; \mathbf{w}) = \mathbf{w}^\top \bar{X}. \tag{3}$$

We would use the training data $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ to find appropriate values for $\mathbf{w}$.

3 Posterior Distribution

Suppose that, the model parameter $\mathbf{w}$ has a Gaussian prior of the form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_0, \mathbf{\Sigma}_0) = \frac{1}{(2\pi)^{(D+1)/2}} \frac{1}{|\mathbf{\Sigma}_0|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \mu_0)^\top \mathbf{\Sigma}_0^{-1}(\mathbf{w} - \mu_0)\right\}. \tag{4}$$

Then, given a set of input data instances $\{\mathbf{x}_i\}_{i=1}^n$, the joint distribution of the corresponding target variables is a Gaussian

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\sigma^2\mathbf{I}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\mathbf{w})\right\}. \tag{5}$$

We would like to find **the posterior distribution** $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$, and then we can estimate the model parameter $\mathbf{w}$ by maximizing the posterior distribution. To do so, we first find the joint distribution of $\mathbf{w}$ and $\mathbf{y}$, and then the conditional distribution of $\mathbf{w}$ given $\mathbf{X}$ and $\mathbf{y}$.

Remark 1. As the input data instances are given (observed), we shall treat them as constants. Thus, to simplify notations, we will denote the conditional distribution $p(\mathbf{y}|\mathbf{w}, \mathbf{X})$ by $p(\mathbf{y}|\mathbf{w})$.

Remark 2. Considering the model parameters as random variables and estimating them by maximizing the corresponding posterior distribution is roughly the idea of Bayesian approach to parameter estimation.

3.1 Joint distribution

We first find the joint distribution over $\mathbf{w}$ and $\mathbf{y}$. Let

$$\mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix}. \tag{6}$$

The log of the joint distribution is

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{w}) + \ln p(\mathbf{y}|\mathbf{w}) \\ &= -\frac{1}{2}(\mathbf{w}-\mu_0)^\top \mathbf{\Sigma}_0^{-1}(\mathbf{w}-\mu_0) \\ &\quad -\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\mathbf{w}) + \text{const}, \end{aligned} \tag{7}$$

where "const" denotes terms that are independent of $\mathbf{w}$ and $\mathbf{y}$. Eq. (7) shows that, the log of the joint distribution over $\mathbf{z}$ is a quadratic function of $\mathbf{z}$. Thus, the joint random variable $\mathbf{z}$ has a Gaussian distribution.

To find $\mathbb{E}[\mathbf{z}]$ and $\text{Cov}[\mathbf{z}]$, we write the first two terms on the RHS of Eq. (7) in the form of

$$-\frac{1}{2}(\mathbf{z}-\mu_\mathbf{z})^\top(\Sigma_\mathbf{z})^{-1}(\mathbf{z}-\mu_\mathbf{z}) = -\frac{1}{2}\mathbf{z}^\top(\Sigma_\mathbf{z})^{-1}\mathbf{z} + \mathbf{z}^\top(\Sigma_\mathbf{z})^{-1}\mu_\mathbf{z} - \frac{1}{2}(\mu_\mathbf{z})^\top(\Sigma_\mathbf{z})^{-1}\mu_\mathbf{z}. \tag{8}$$

The second and first order terms in Eq. (7) are

$$\begin{aligned} &-\frac{1}{2}\mathbf{w}^\top\mathbf{\Sigma}_0^{-1}\mathbf{w} - \frac{1}{2}\mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\mathbf{w}^\top\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^\top\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} \\ =& -\frac{1}{2}\mathbf{w}^\top\left(\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\right)\mathbf{w} - \frac{1}{2}\mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^\top\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{y} \\ =& -\frac{1}{2}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}^\top\begin{pmatrix}\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix} \\ =& -\frac{1}{2}\mathbf{z}^\top\begin{pmatrix}\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1}\end{pmatrix}\mathbf{z}, \end{aligned} \tag{9}$$

and

$$\mathbf{w}^\top\mathbf{\Sigma}_0^{-1}\mu_0 = \begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}^\top\begin{pmatrix}\mathbf{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0}\end{pmatrix} = \mathbf{z}^\top\begin{pmatrix}\mathbf{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0}\end{pmatrix}, \tag{10}$$

respectively. The only remaining term is

$$-\frac{1}{2}(\mu_0)^\top\mathbf{\Sigma}_0^{-1}\mu_0. \tag{11}$$

Thus, combining Eqs. (9), (10), and (11), we can rewrite Eq. (7) by

$$\begin{aligned} \ln p(\mathbf{z}) = &-\frac{1}{2}\mathbf{z}^\top\begin{pmatrix}\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1}\end{pmatrix}\mathbf{z} + \mathbf{z}^\top\begin{pmatrix}\mathbf{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0}\end{pmatrix} - \frac{1}{2}(\mu_0)^\top\mathbf{\Sigma}_0^{-1}\mu_0 \\ &+\text{const}. \end{aligned} \tag{12}$$

Comparing the first two terms in Eq. (12) with those in Eq. (8) leads to

$$(\Sigma_\mathbf{z})^{-1} = \begin{pmatrix}\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1}\end{pmatrix}, \tag{13}$$

and

$$(\Sigma_{\mathbf{z}})^{-1}\mu_{\mathbf{z}} = \begin{pmatrix} \boldsymbol{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0} \end{pmatrix}. \tag{14}$$

To solve for $\Sigma_{\mathbf{z}}$, we introduce the result as follows.

Lemma 1. *Suppose that the involved matrices are invertible. Then,*

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{MBD^{-1}} \\ -\mathbf{D^{-1}CM} & \mathbf{D}^{-1} + \mathbf{D^{-1}CMBD^{-1}} \end{pmatrix},$$

where

$$\mathbf{M} = (\mathbf{A} - \mathbf{BD^{-1}C})^{-1}.$$

Applying Lemma 1 to Eq. (13) yields

$$\Sigma_{\mathbf{z}} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0\mathbf{X}^\top \\ \mathbf{X}\boldsymbol{\Sigma}_0 & \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top \end{pmatrix}. \tag{15}$$

Moreover, in view of Eq. (14), it is easy to see that

$$\mu_{\mathbf{z}} = \Sigma_{\mathbf{z}} \begin{pmatrix} \boldsymbol{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mathbf{X}\mu_0 \end{pmatrix}. \tag{16}$$

All together concludes that

$$\mathrm{Cov}[\mathbf{z}] = \Sigma_{\mathbf{z}}, \tag{17}$$
$$\mathbb{E}[\mathbf{z}] = \mu_{\mathbf{z}}. \tag{18}$$

Thus, the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ is a Gaussian, which takes the form of

$$p(\mathbf{w}, \mathbf{y}) = p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mu_{\mathbf{z}}, \Sigma_{\mathbf{z}}). \tag{19}$$

Remark 3. We can easily check that, the remaining term in Eq. (8) is also consistent with the third term in Eq. (12),

$$(\mu_{\mathbf{z}})^\top(\Sigma_{\mathbf{z}})^{-1}\mu_{\mathbf{z}} = (\mu_0)^\top\boldsymbol{\Sigma}_0^{-1}\mu_0. \tag{20}$$

3.2 Marginal distribution

As we have derived the joint distribution of $\mathbf{w}$ and $\mathbf{y}$ in Eq. (19), we can derive many related distributions, e.g., the marginal distributions.

Recall that, we assume that the model parameter $\mathbf{w}$ has a Gaussian distribution given by Eq. (4). Clearly, we have

$$\mathbb{E}[\mathbf{w}] = \mu_0, \tag{21}$$
$$\mathrm{Cov}[\mathbf{w}] = \boldsymbol{\Sigma}_0. \tag{22}$$

In view of Eq. (17) and Eq. (18), we can see that

$$\mathbb{E}[\mathbf{y}] = \mathbf{X}\mu_0, \tag{23}$$
$$\mathrm{Cov}[\mathbf{y}] = \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top. \tag{24}$$

This implies that $\mathbf{y}$ is also Gaussian, and

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mu_0, \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top). \tag{25}$$

To simplify notations, let $\mu_{\mathbf{y}} = \mathbb{E}[\mathbf{y}]$ and $\Sigma_{\mathbf{y}} = \mathrm{Cov}[\mathbf{y}]$.

3.3 Conditional distribution

Our goal is to find the posterior distribution of the model parameter $\mathbf{w}$, which is the conditional distribution of $\mathbf{w}$ given $\mathbf{y}$. Indeed, the conditional distribution of $\mathbf{w}$ given $\mathbf{y}$ is also a Gaussian (why?). Let

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}|\mathbf{y}}, \Sigma_{\mathbf{w}|\mathbf{y}}).$$

Our goal is to find $\mu_{\mathbf{w}|\mathbf{y}}$ and $\Sigma_{\mathbf{w}|\mathbf{y}}$. Notice that

$$\begin{aligned}\ln p(\mathbf{z}) &= \ln p(\mathbf{w}|\mathbf{y}) + \ln p(\mathbf{y}) \\ &= -\frac{1}{2}(\mathbf{w}-\mu_{\mathbf{w}|\mathbf{y}})^\top \Sigma_{\mathbf{w}|\mathbf{y}}^{-1}(\mathbf{w}-\mu_{\mathbf{w}|\mathbf{y}}) \\ &\quad -\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{y}})^\top \Sigma_{\mathbf{y}}^{-1}(\mathbf{y}-\mu_{\mathbf{y}}) + \text{const.}\end{aligned} \tag{26}$$

To simplify notations, we denote the covariance matrix $\Sigma_{\mathbf{z}}$ and the **precision matrix** $\Lambda_{\mathbf{z}} = \Sigma_{\mathbf{z}}^{-1}$ by

$$\boldsymbol{\Sigma}_{\mathbf{z}} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{ww}} & \boldsymbol{\Sigma}_{\mathbf{wy}} \\ \boldsymbol{\Sigma}_{\mathbf{yw}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0\mathbf{X}^\top \\ \mathbf{X}\boldsymbol{\Sigma}_0 & \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\top\end{pmatrix} \tag{27}$$

and

$$\boldsymbol{\Lambda}_{\mathbf{z}} = \begin{pmatrix}\boldsymbol{\Lambda}_{\mathbf{ww}} & \boldsymbol{\Lambda}_{\mathbf{wy}} \\ \boldsymbol{\Lambda}_{\mathbf{yw}} & \boldsymbol{\Lambda}_{\mathbf{yy}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1}\end{pmatrix}. \tag{28}$$

In view of Eq. (19), we have

$$\begin{aligned}\ln p(\mathbf{z}) &= -\frac{1}{2}(\mathbf{z}-\mu_{\mathbf{z}})^\top \Sigma_{\mathbf{z}}^{-1}(\mathbf{z}-\mu_{\mathbf{z}}) + \text{const} \\ &= -\frac{1}{2}\left(\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix} - \begin{pmatrix}\mu_0 \\ \mu_{\mathbf{y}}\end{pmatrix}\right)^\top \begin{pmatrix}\boldsymbol{\Lambda}_{\mathbf{ww}} & \boldsymbol{\Lambda}_{\mathbf{wy}} \\ \boldsymbol{\Lambda}_{\mathbf{yw}} & \boldsymbol{\Lambda}_{\mathbf{yy}}\end{pmatrix}\left(\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix} - \begin{pmatrix}\mu_0 \\ \mu_{\mathbf{y}}\end{pmatrix}\right) + \text{const} \qquad (29) \\ &= -\frac{1}{2}(\mathbf{w}-\mu_0)^\top \boldsymbol{\Lambda}_{\mathbf{ww}}(\mathbf{w}-\mu_0) \\ &\quad -\frac{1}{2}(\mathbf{w}-\mu_0)^\top \boldsymbol{\Lambda}_{\mathbf{wy}}(\mathbf{y}-\mu_{\mathbf{y}}) \\ &\quad -\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{y}})^\top \boldsymbol{\Lambda}_{\mathbf{yw}}(\mathbf{w}-\mu_0) \\ &\quad -\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{y}})^\top \boldsymbol{\Lambda}_{\mathbf{yy}}(\mathbf{y}-\mu_{\mathbf{y}}) + \text{const.} \qquad (30)\end{aligned}$$

The only quadratic term of $\mathbf{w}$ in Eq. (29) is

$$-\frac{1}{2}\mathbf{w}^\top \boldsymbol{\Lambda}_{\mathbf{ww}}\mathbf{w}.$$

In view of Eq. (26), we have

$$\Sigma_{\mathbf{w}|\mathbf{y}}^{-1} = \boldsymbol{\Lambda}_{\mathbf{ww}} = \boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X},$$

and thus

$$\Sigma_{\mathbf{w}|\mathbf{y}} = (\boldsymbol{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1}$$

where $\beta = 1/\sigma^2$. Similarly, as the linear terms of $\mathbf{w}$ in Eq. (29) is

$$\mathbf{w}^\top \{\mathbf{\Lambda_{ww}}\mu_0 - \mathbf{\Lambda_{wy}}(\mathbf{y} - \mu_\mathbf{y})\},$$

and the linear terms of $\mathbf{w}$ in Eq. (26) is

$$\mathbf{w}^\top \Sigma_{\mathbf{w}|\mathbf{y}}^{-1}\mu_{\mathbf{w}|\mathbf{y}},$$

we have

$$\mu_{\mathbf{w}|\mathbf{y}} = \mathbf{\Sigma_{w|y}}\{\mathbf{\Lambda_{ww}}\mu_0 - \mathbf{\Lambda_{wy}}(\mathbf{y} - \mu_\mathbf{y})\} = \mathbf{\Sigma_{w|y}}\{\mathbf{\Lambda_{ww}}\mu_0 - \mathbf{\Lambda_{wy}}(\mathbf{y} - \mathbf{X}\mu_0)\}.$$

Notice that, the covariance matrix $\mathbf{\Sigma_{w|y}}$ and the expectation $\mu_{\mathbf{w}|\mathbf{y}}$ depend on the data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with N data instances. Thus, we denote $\mathbf{\Sigma_{w|y}}$ and $\mu_{\mathbf{w}|\mathbf{y}}$ by

$$\mathbf{\Sigma}_N = (\mathbf{\Sigma}_0^{-1} + \beta\mathbf{X}^\top\mathbf{X})^{-1}, \tag{31}$$

and

$$\mu_N = \mathbf{\Sigma}_N\{\mathbf{\Lambda_{ww}}\mu_0 - \mathbf{\Lambda_{wy}}(\mathbf{y} - \mathbf{X}\mu_0)\} = \mathbf{\Sigma}_N(\mathbf{\Sigma}_0^{-1}\mu_0 + \beta\mathbf{X}^\top\mathbf{y}), \tag{32}$$

respectively.

All together, the posterior distribution of $\mathbf{w}$ given $\mathbf{y}$ (after observing $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$) is

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mu_N, \mathbf{\Sigma}_N) \tag{33}$$

with μ_N and $\mathbf{\Sigma}_N$ given by Eq. (32) and Eq. (31), respectively.

4 Maximum a Posterior

Suppose that $\mu_0 = 0$ and $\mathbf{\Sigma}_0 = \mathbf{I}/\alpha$. Then

$$\mathbf{\Sigma}_N = (\alpha\mathbf{I} + \beta\mathbf{X}^\top\mathbf{X})^{-1} \tag{34}$$

$$\mu_N = \beta\mathbf{\Sigma}_N\mathbf{X}^\top\mathbf{y}. \tag{35}$$

The log of the posterior distribution is

$$\ln p(\mathbf{w}|\mathbf{y}) = -\frac{\beta}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 - \frac{\alpha}{2}\|\mathbf{w}\|^2 + \text{const} \tag{36}$$

Maximization of this posterior distribution leads to the same solution by quadratic regularized least squares.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.