Lecture 03. Bayesian Linear Regression

Lecturer: Jie Wang Date: Sep 22, 2021

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 $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ to find appropriate values for $\mathbf{w}$.

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

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

where

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

Lemma 2. *When all inverses exist,*

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}.$$

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 $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$. 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 $p(\mathbf{y}|\mathbf{w})$.

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 $\mathrm{Cov}[\mathbf{z}]$, 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

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

and

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

respectively.

Comparing Eq. (9) and Eq. (10) with the second and first order terms in Eq. (8), we have

$$(\Sigma_{\mathbf{z}})^{-1} = \begin{pmatrix} \mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1} \end{pmatrix} \Rightarrow \Sigma_{\mathbf{z}} = \begin{pmatrix} \mathbf{\Sigma}_0 & \mathbf{\Sigma}_0\mathbf{X}^\top \\ \mathbf{X}\mathbf{\Sigma}_0 & \sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top \end{pmatrix} \tag{11}$$

and

$$(\Sigma_{\mathbf{z}})^{-1}\mu_{\mathbf{z}} = \begin{pmatrix} \mathbf{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0} \end{pmatrix} \Rightarrow \mu_{\mathbf{z}} = \Sigma_{\mathbf{z}} \begin{pmatrix} \mathbf{\Sigma}_0^{-1}\mu_0 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mathbf{X}\mu_0 \end{pmatrix}. \tag{12}$$

Moreover, it is easy to check that

$$(\mu_{\mathbf{z}})^\top(\Sigma_{\mathbf{z}})^{-1}\mu_{\mathbf{z}} = (\mu_0)^\top\mathbf{\Sigma}_0^{-1}\mu_0. \tag{13}$$

Therefore, we can rewrite the first two terms on the RHS of Eq. (7) in the form of Eq. (8) with $\mathbf{z}$, $\Sigma_{\mathbf{z}}$, and $\mu_{\mathbf{z}}$ defined by Eq. (6), Eq. (11), and Eq. (12), respectively. This shows that

$$\begin{aligned} \mathrm{Cov}[\mathbf{z}] &= \Sigma_{\mathbf{z}}, \qquad (14) \\ \mathbb{E}[\mathbf{z}] &= \mu_{\mathbf{z}}. \qquad (15) \end{aligned}$$

3.2 Inverse of partitioned matrix

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

$$\mathbf{\Sigma}_{\mathbf{z}} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{ww}} & \mathbf{\Sigma}_{\mathbf{wy}} \\ \mathbf{\Sigma}_{\mathbf{yw}} & \mathbf{\Sigma}_{\mathbf{yy}} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_0 & \mathbf{\Sigma}_0\mathbf{X}^\top \\ \mathbf{X}\mathbf{\Sigma}_0 & \sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top \end{pmatrix} \tag{16}$$

and

$$\mathbf{\Lambda}_{\mathbf{z}} = \begin{pmatrix} \mathbf{\Lambda}_{\mathbf{ww}} & \mathbf{\Lambda}_{\mathbf{wy}} \\ \mathbf{\Lambda}_{\mathbf{yw}} & \mathbf{\Lambda}_{\mathbf{yy}} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} & -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ -(\sigma^2\mathbf{I})^{-1}\mathbf{X} & (\sigma^2\mathbf{I})^{-1} \end{pmatrix}. \tag{17}$$

Lemma 1 leads to some interesting properties. For example, let $\mathbf{\Sigma}_{\mathbf{ww}}$, $\mathbf{\Sigma}_{\mathbf{wy}}$, $\mathbf{\Sigma}_{\mathbf{yw}}$, and $\mathbf{\Sigma}_{\mathbf{yy}}$ be the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$, respectively. Then

$$\begin{aligned} \mathbf{\Lambda}_{\mathbf{ww}} &= \mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X} \\ &= (\mathbf{\Sigma}_{\mathbf{ww}} - \mathbf{\Sigma}_{\mathbf{wy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yw}})^{-1} \\ &= (\mathbf{\Sigma}_0 - (\mathbf{\Sigma}_0\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{\Sigma}_0)^{-1}. \end{aligned} \tag{18}$$

$$\begin{aligned} \mathbf{\Lambda}_{\mathbf{wy}} &= -\mathbf{X}^\top(\sigma^2\mathbf{I})^{-1} \\ &= -(\mathbf{\Sigma}_{\mathbf{ww}} - \mathbf{\Sigma}_{\mathbf{wy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{\Sigma}_{\mathbf{yw}})^{-1}\mathbf{\Sigma}_{\mathbf{wy}}\mathbf{\Sigma}_{\mathbf{yy}}{}^{-1}. \end{aligned} \tag{19}$$

3.3 Marginal distribution

In view of Eq. (14) and Eq. (15), we can see that

$$\begin{aligned} \mathbb{E}[\mathbf{y}] &= \mathbf{X}\mu_0, \qquad (20) \\ \mathrm{Cov}[\mathbf{y}] &= \sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top. \qquad (21) \end{aligned}$$

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

3.4 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 $\mathbf{y}$. Notice that

$$\begin{aligned}\ln p(\mathbf{z}) &= \ln p(\mathbf{w}|\mathbf{y}) + \ln p(\mathbf{y}) \\ &= -\frac{1}{2}(\mathbf{w} - \mu_{\mathbf{w}|\mathbf{y}})^\top \Sigma_{\mathbf{w}|\mathbf{y}}^{-1}(\mathbf{w} - \mu_{\mathbf{w}|\mathbf{y}}) \\ &\quad - \frac{1}{2}(\mathbf{y} - \mu_{\mathbf{y}})^\top \Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - \mu_{\mathbf{y}}) + \text{const}.\end{aligned} \tag{22}$$

In view of Eq. (8), Eq. (14), Eq. (15), and Eq. (6), we have

$$\begin{aligned}\ln p(\mathbf{z}) &= -\frac{1}{2}\left(\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix} - \begin{pmatrix}\mu_0\\ \mu_{\mathbf{y}}\end{pmatrix}\right)^\top \begin{pmatrix}\mathbf{\Lambda}_{\mathbf{ww}} & \mathbf{\Lambda}_{\mathbf{wy}}\\ \mathbf{\Lambda}_{\mathbf{yw}} & \mathbf{\Lambda}_{\mathbf{yy}}\end{pmatrix}\left(\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix} - \begin{pmatrix}\mu_0\\ \mu_{\mathbf{y}}\end{pmatrix}\right) + \text{const} \\ &= -\frac{1}{2}(\mathbf{w} - \mu_0)^\top \mathbf{\Lambda}_{\mathbf{ww}}(\mathbf{w} - \mu_0) \\ &\quad - \frac{1}{2}(\mathbf{w} - \mu_0)^\top \mathbf{\Lambda}_{\mathbf{wy}}(\mathbf{y} - \mu_{\mathbf{y}}) \\ &\quad - \frac{1}{2}(\mathbf{y} - \mu_{\mathbf{y}})^\top \mathbf{\Lambda}_{\mathbf{yw}}(\mathbf{w} - \mu_0) \\ &\quad - \frac{1}{2}(\mathbf{y} - \mu_{\mathbf{y}})^\top \mathbf{\Lambda}_{\mathbf{yy}}(\mathbf{y} - \mu_{\mathbf{y}}) + \text{const}.\end{aligned} \tag{23}$$

The only quadratic term of $\mathbf{w}$ in Eq. (23) is

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

In view of Eq. (22), we have

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

which leads to

$$\Sigma_{\mathbf{w}|\mathbf{y}} = (\mathbf{\Sigma}_0^{-1} + \mathbf{X}^\top(\sigma^2\mathbf{I})^{-1}\mathbf{X})^{-1} = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{\Sigma}_0. \tag{24}$$

Similarly, the linear terms of $\mathbf{w}$ in Eq. (23) is

$$\mathbf{w}^\top\left\{\mathbf{\Lambda}_{\mathbf{ww}}\mu_0 - \mathbf{\Lambda}_{\mathbf{wy}}(\mathbf{y} - \mu_{\mathbf{y}})\right\}.$$

Moreover, as the linear terms of $\mathbf{w}$ in Eq. (22) is

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

we have

$$\begin{aligned}\mu_{\mathbf{w}|\mathbf{y}} &= \mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}}\left\{\mathbf{\Lambda}_{\mathbf{ww}}\mu_0 - \mathbf{\Lambda}_{\mathbf{wy}}(\mathbf{y} - \mu_{\mathbf{y}})\right\} \\ &= \mathbf{\Lambda}_{\mathbf{ww}}^{-1}\left\{\mathbf{\Lambda}_{\mathbf{ww}}\mu_0 - \mathbf{\Lambda}_{\mathbf{wy}}(\mathbf{y} - \mu_{\mathbf{y}})\right\} \\ &= \mu_0 - \mathbf{\Lambda}_{\mathbf{ww}}^{-1}\mathbf{\Lambda}_{\mathbf{wy}}(\mathbf{y} - \mu_{\mathbf{y}}) \\ &= \mu_0 + \mathbf{\Sigma}_{\mathbf{wy}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y} - \mu_{\mathbf{y}}) \\ &= \mu_0 + \mathbf{\Sigma}_0\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top)^{-1}(\mathbf{y} - \mu_{\mathbf{y}}) \\ &= \mu_0 + \mathbf{\Sigma}_0\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{\Sigma}_0\mathbf{X}^\top)^{-1}(\mathbf{y} - \mathbf{X}\mu_0).\end{aligned} \tag{25}$$

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

$$\mathbf{\Sigma}_N = (\mathbf{\Sigma}_0^{-1} + \beta \mathbf{X}^\top \mathbf{X})^{-1}, \tag{26}$$

and

$$\mu_N = \mathbf{\Sigma}_N \left\{ \mathbf{\Lambda}_{\mathbf{ww}} \mu_0 - \mathbf{\Lambda}_{\mathbf{wy}} (\mathbf{y} - \mathbf{X}\mu_0) \right\} = \mathbf{\Sigma}_N (\mathbf{\Sigma}_0^{-1} \mu_0 + \beta \mathbf{X}^\top \mathbf{y}), \tag{27}$$

with $\beta = 1/\sigma^2$, respectively.

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

$$p(\mathbf{w}|\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mu_N, \mathbf{\Sigma}_N) \tag{28}$$

with μ_N and $\mathbf{\Sigma}_N$ given by Eq. (27) and Eq. (26), respectively.

4 Maximum a Posterior

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

$$\mathbf{\Sigma}_N = (\alpha \mathbf{I} + \beta \mathbf{X}^\top \mathbf{X})^{-1} \tag{29}$$

$$\mu_N = \beta \mathbf{\Sigma}_N \mathbf{X}^\top \mathbf{y}. \tag{30}$$

The log of the posterior distribution is

$$\ln p(\mathbf{w}|\mathbf{y}) = -\frac{\beta}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 - \frac{\alpha}{2} \|\mathbf{w}\|^2 + \text{const} \tag{31}$$

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.