CS189/289A SPRING 2023: LINEAR REGRESSION

LECTURE NOTE

Chawin Sitawarin *

University of California, Berkeley

Demo: https://colab.research.google.com/drive/1Pv1BUScUd6fHm2s2pQ9oLPHwg60QSYi8

1 Optimization View of Linear Regression

One very nice thing about linear regression is that it can be tackled or studied from various perspectives and domains. One can simply look at linear regression as a standalone optimization problem. Specifically, given some matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and vector $\boldsymbol{y} \in \mathbb{R}^n$, we wish to find the parameters $\boldsymbol{w} \in \mathbb{R}^d$ that minimize the loss function $loss(w; \boldsymbol{X}, \boldsymbol{y})$ which is chosen to be the mean (or sum) of squared errors (MSE):

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{R}^d}{\arg\min}\ loss(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y}) \tag{1}$$

$$\text{where} \quad L(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}\right)^2 = \frac{1}{n} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right\|_2^2 . \tag{2}$$

A simple interpretation is that we want to best predict or "reconstruct" any variable y_i given only its feature vector $\boldsymbol{x}_i$ by making an assumption that their relationship is linear.

2 Probabilistic View of Linear Regression

A different way to motivate linear regression is as a probabilistic model for the data. Here, we assume that the data is generated by a linear model with additive Gaussian noise. Or more specifically, we assume that the data is generated by the following model:

$$p(y \mid \boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(y \mid \boldsymbol{x}^\top \boldsymbol{w}, \sigma^2), \tag{3}$$

or in other words,

$$y = \boldsymbol{x}^\top \boldsymbol{w} + \epsilon \qquad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

- This view shows another example of MLE.
- This view will be useful when we study extensions of linear regression such as ridge regression and LASSO, particularly as a solution of the Maximum A Posteriori (MAP) problem.

In this setup, we can view linear regression as a way to estimate the parameters $\boldsymbol{w}$ and σ^2 from the data via *Maximum Likelihood Estimation* (MLE).

*Corresponding email: chawins@berkeley.edu

We write the Likelihood function as

$$L(\boldsymbol{w}, \sigma \mid \boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{w}, \sigma^2) \tag{5}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}\right)^2\right) \tag{6}$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right\|_2^2\right) \tag{7}$$

We can miximize the likelihood directly, but more often than not, it is more convenient to minimize the *Negative Log-Likelihood* (NLL) instead. This term is frequently used in machine learning as well as most of the deep learning models.

$$\arg\max_{\boldsymbol{w},\sigma} \; L(\boldsymbol{w}, \sigma \mid \boldsymbol{X}, \boldsymbol{y}) = \arg\max_{\boldsymbol{w},\sigma} \; \log L(\boldsymbol{w}, \sigma \mid \boldsymbol{X}, \boldsymbol{y}) \qquad \text{(Log-Likelihood)} \tag{8}$$

$$= -\arg\min_{\boldsymbol{w},\sigma} \; \log L(\boldsymbol{w}, \sigma \mid \boldsymbol{X}, \boldsymbol{y}) \qquad \text{(Negative Log-Likelihood)} \tag{9}$$

$$= \arg\min_{\boldsymbol{w},\sigma} \; \frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right\|_2^2 + \frac{n}{2}\log(2\pi\sigma^2) \tag{10}$$

Most of the times, we focus more on finding $\boldsymbol{w}^*$ and ignore σ^*. To find $\boldsymbol{w}^*$ we set the gradient to zero, i.e.,

$$\nabla_w L(\boldsymbol{w}^*, \sigma \mid \boldsymbol{X}, \boldsymbol{y}) = 0 \tag{11}$$

$$\frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^*\right\|_2^2 + \frac{n}{2}\log(2\pi\sigma^2)\right) = 0 \tag{12}$$

$$\frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{y}) = 0 \tag{13}$$

$$\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}^* = 0 \tag{14}$$

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{X}^\top\boldsymbol{y} \tag{15}$$

$$\boldsymbol{w}^* = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \tag{16}$$

We call the final MSE produced by $\boldsymbol{w}$ *Residual Sum of Squares* (RSS).

$$RSS(\boldsymbol{w}) = \left\|\boldsymbol{y} - \hat{\boldsymbol{y}}\right\|_2^2 = \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right\|_2^2 \tag{17}$$

Exercise 1

Show that the MLE solution for σ (i.e., σ^*) is given by $\sigma^* = \frac{1}{n}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}^*\right\|_2^2$.

2.1 Generative Model View of Linear Regression

Instead of viewing linear regression as a way to estimate $p(y \mid \boldsymbol{x})$ using the parameters $\boldsymbol{w}$ and σ^2, we can also view it as a "generative model" for the joint distribution $p(y, \boldsymbol{x})$ instead. In this case, we make an assumption that $p(y, \boldsymbol{x})$ is *jointly Guassian*, or precisely, we assume that they are generated from the following distribution:

$$p(y, \boldsymbol{x}) = \mathcal{N}\left(\begin{bmatrix} y \\ \boldsymbol{x} \end{bmatrix} \mid \begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}\right) \tag{18}$$

Then we will estimate all the parameters μ_y, $\boldsymbol{\mu}_x$, Σ_{yy}, $\boldsymbol{\Sigma}_{xy}$, $\boldsymbol{\Sigma}_{xx}$ from the data with MLE. Given that the joint distribution is Gaussian, we know that the marginal distribution for each variable is also Gaussian. As a result, we know

that the following empirical means and covariances are the MLE solutions:

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{19}$$

$$\hat{\boldsymbol{\mu}}_x = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \tag{20}$$

$$\hat{\Sigma}_{yy} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2 \tag{21}$$

$$\hat{\boldsymbol{\Sigma}}_{xy} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_x)(y_i - \hat{\mu}_y)^\top \tag{22}$$

$$\hat{\boldsymbol{\Sigma}}_{xx} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_x)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_x)^\top \tag{23}$$

Now we will show that $\mathbb{E}[y \mid x]$ takes the same form as the predicted value $\hat{y}$ in the previous section.

$$\mathbb{E}[y \mid x] = \mu_{y|x} \tag{24}$$

$$= \mu_y + \boldsymbol{\Sigma}_{xy}^\top \boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x) \tag{25}$$

Then plugging in the MLE solutions, we get:

$$\hat{y} \coloneqq f(\boldsymbol{x}) = \hat{\mu}_y + \hat{\boldsymbol{\Sigma}}_{xy}^\top \hat{\boldsymbol{\Sigma}}_{xx}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_x) \tag{26}$$

$$= \hat{\boldsymbol{\Sigma}}_{xy}^\top \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\boldsymbol{x} + \left(\hat{\mu}_y - \hat{\boldsymbol{\Sigma}}_{xy}^\top \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\mu}}_x\right) \tag{27}$$

$$= \boldsymbol{w}_1^\top \boldsymbol{x} + w_0 \tag{28}$$

$$= \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{xy}^\top \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \\ \hat{\mu}_y - \hat{\boldsymbol{\Sigma}}_{xy}^\top \hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\mu}}_x \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} \tag{29}$$

It takes some algebra to convince that $[\boldsymbol{w}_1; w_0] = \boldsymbol{w} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ in the previous section if the rows of $\boldsymbol{X}$ are $[\boldsymbol{x}_i \; 1]$'s (appended by 1).

Acknowledgement

This note is, for the most parts, based on Murphy [2022].

References

K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL `probml.ai`. 3