Lecture 1. Linear Regression

Lecturer: Jie Wang Date: March 19, 2021

The major reference of this lecture is [1].

1 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:

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

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

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

$$Y = f(X; \mathbf{w}) = \mathbf{w}^\top \bar{X}. \tag{2}$$

We would use the training data we have collected to find appropriate values for $\mathbf{w}$.

2 Linear Regression by Least Squares

Suppose that we are given a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$.

For the ideal case, we would like

$$y_i = f(\mathbf{x}_i; \mathbf{w}) = \mathbf{w}^\top \bar{\mathbf{x}}_i \tag{3}$$

holds for all $i = 1, \ldots, n$, where $\bar{\mathbf{x}}_i = (1, x_{i,1}, \ldots, x_{i,D})^\top$, that is, the model can perfectly explains the data (the input variables are also known as *explanatory variables*). In matrix form, we can write the system of the equations (3) as

$$\mathbf{y} = \mathbf{X}\mathbf{w}, \tag{4}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$ with its i^{th} row being $\mathbf{x}_i^\top$ with an additional 1 in the first position, i.e.,

$$\mathbf{X} = \begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \\ \vdots \\ \bar{\mathbf{x}}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,D} \end{pmatrix}.$$

For simplicity, we assume that $\mathbf{X}$ has full (column) rank in this lecture.

Notice that, in real applications, Eq. (4) often has no solutions.

Question 1.

1. When the problem in (4) has no solution, it is called an inconsistent system. How to tell if (4) is consistent or not?

2. If the problem in (4) admits a solution, when it is (not) unique?

In many real problems, we have more equations in (4) than the number of input variables, i.e., $n > D$, which implies that the system of equations in (4) is often inconsistent. We thus would like to find a model that can minimize the fitting errors.

We first define the fitting error for each data instance by

$$\ell(y_i, f(\mathbf{x}_i; \mathbf{w})) = \frac{1}{2}(y_i - f(\mathbf{x}_i; \mathbf{w}))^2.$$

Then, the average fitting error of the linear model over the whole data set is

$$L(\mathbf{w}) = \frac{1}{n}\sum_i^n \ell(y_i, f(\mathbf{x}_i; \mathbf{w})) = \frac{1}{2n}\sum_i^n (y_i - f(\mathbf{x}_i; \mathbf{w}))^2 = \frac{1}{2n}\sum_i^n (y_i - \mathbf{w}^\top \bar{\mathbf{x}}_i)^2 = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2.$$

We can thus estimate $\mathbf{w}$ by minimizing the total loss L, i.e.,

$$\hat{\mathbf{w}}_{LS} \in \underset{\mathbf{w}}{\mathbf{argmin}}\, L(\mathbf{w}),$$

which can be done by the first order optimization conditions. Thus, by setting the gradient of L to 0, we have

$$0 = \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}\bigg|_{\mathbf{w}=\hat{\mathbf{w}}_{LS}} = -\frac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_{LS}). \tag{5}$$

Then, solving (5) leads to

$$\hat{\mathbf{w}}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{6}$$

The solution $\hat{\mathbf{w}}_{LS}$ in (6) is the so-called *least-squares solution*.

Thus, for a new data instance $\mathbf{x}$, the estimated response by the model is

$$\hat{y} = f(\mathbf{x}; \hat{\mathbf{w}}_{LS}) = (\bar{\mathbf{x}})^\top \hat{\mathbf{w}}_{LS}.$$

2.1 The geometric interpretation of the least squares solution

In view of the model in (3) and (4,) and the estimated parameters in (6,) the estimated responses on the given data instances can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}_{LS} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{7}$$

There is a nice geometric meaning of the above equation, that is, $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$. Specifically,

$$\hat{\mathbf{y}} = \underset{\mathbf{z}\in\mathcal{C}(\mathbf{X})}{\mathbf{argmin}}\, \|\mathbf{z} - \mathbf{y}\|_2,$$

where $\mathcal{C}(\mathbf{X})$ is the column space of $\mathbf{X}$. Thus, we call the matrix

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

the projection matrix that projects an arbitrary vector to the column space of $\mathbf{X}$.

Question 2.

1. How to tell if $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ on $\mathcal{C}(\mathbf{X})$?
2. How to find $\hat{\mathbf{y}}$ when $\mathbf{X}$ is not invertible? Is $\hat{\mathbf{y}}$ unique? What about $\hat{\mathbf{w}}$ in Eq. (7)?

2.2 Regularized least squares

One of the most commonly used approach to alleviate *over-fitting* is by adding a regularization term to the loss function

$$L(\mathbf{w}) + \lambda\Omega(\mathbf{w}),$$

where λ is the positive regularization parameter. One of the simplest regularizers is given by the ℓ_2-norm of the model parameters:

$$\Omega(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{1}{2}\mathbf{w}^\top\mathbf{w}.$$

In this case, we estimate the weight parameters $\mathbf{w}$ by minimizing the total loss

$$\hat{\mathbf{w}}_{RLS} \in \operatorname*{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}. \tag{8}$$

By the first-order optimization condition, we have

$$0 = \frac{\partial(L(\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w})}{\partial\mathbf{w}}\bigg|_{\mathbf{w}=\hat{\mathbf{w}}_{RLS}} = -\frac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_{RLS}) + \lambda\hat{\mathbf{w}}_{RLS},$$

which leads to

$$\hat{\mathbf{w}}_{RLS} = (\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Remark 1. Notice that, the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is always invertable (why?), which implies that the problem in (8) always admits a unique solution. This is an extra benefit by the regularization.

The above regularized least squares regression is also known as the *ridge regression*. In general, the regularization term would encourage the weight values to shirink towards zero.

2.3 Basis function

We can also model nonlinear relationships between the input features and the corresponding response by introducing the so-called basis functions. Specifically, let

$$\phi_i(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R},\ i = 1, \ldots, m$$

be a set of transformations—that can be nonlinear—of the input data. Then,

$$f(\mathbf{x}; \mathbf{w}) = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_m\phi_m(\mathbf{x}).$$

Commonly used basis functions includes polynomials, Gaussians, sigmoid functions, etc. Notice that, the model is still linear with respect to the parameters.

3 Linear Regression by Maximum Likelihood

Consider real applications. We can never take into account all the influencing factors and our equipments are not perfect. This implies that even we sample exactly the same data instance, we may get as many different values of the response variables as the times we evalute the corresponding response. Thus, we can explicitly introduce this uncertainty into our model:

$$Y = f(X; \mathbf{w}) = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_D X_D + \epsilon, \tag{9}$$

where

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

Thus, for each input data instance, the output is

$$y_i = f(\mathbf{x}_i; \mathbf{w}) + \epsilon_i = \mathbf{w}^\top \bar{\mathbf{x}}_i + \epsilon_i.$$

Notice that, though $\{\mathbf{x}_i\}$ is fixed, the target variable $\{y_i\}$ is not (why?).

Thus, the probability density function of y_i conditioned on the model parameters and the input variables is

$$p(y_i | \mathbf{x}_i, \mathbf{w}, \sigma) = \mathcal{N}(\mathbf{w}^\top \bar{\mathbf{x}}_i, \sigma^2).$$

The key idea of the Maximum Likelihood Estimation (MLE) is that we try to find $\hat{\mathbf{w}}$ that can maximize the joint density (likelihood) of $\{y_i\}$ conditioned on *the model parameters* and *the input variables*, that is

$$\hat{\mathbf{w}}_{ML} = \underset{\mathbf{w}}{\mathbf{argmax}}\ p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{w}, \sigma) = \underset{\mathbf{w}}{\mathbf{argmax}}\ p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \sigma). \tag{10}$$

By assuming that $\{y_i\}$ are mutually independent given the model parameters and the input variables, the so-called likelihood function—that is, the objective function in (10)—takes the form of

$$L(\mathbf{w}) = p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \sigma) = \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \mathbf{w}, \sigma). \tag{11}$$

For computational convenience, we usually deal with the log-likelihood.

$$\begin{aligned}
\log L(\mathbf{w}) &= \sum_{n=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2 \right\} \right) \\
&= -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \bar{\mathbf{x}}_i)^2 \\
&= -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2.
\end{aligned}$$

The first-order optimization condition implies that

$$0 = \left. \frac{\partial \log L(\mathbf{w})}{\partial \mathbf{w}} \right|_{\mathbf{w} = \hat{\mathbf{w}}_{ML}} = \frac{1}{\sigma^2} \left(\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}}_{ML} \right).$$

Thus,

$$\hat{\mathbf{w}}_{ML} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

which is the same as the one derived from the least squares perspective.

References

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