CHAPTER 19

Linear Regression

The fundamental idea behind the linear regression algorithm is that it assumes a linear relationship between the features of the dataset. As a result of the pre-defined structure that is imposed on the parameters of the model, it is also called a parametric learning algorithm. Linear regression is used to predict targets that contain real values. As we will see later in Chapter 20 on logistic regression, the linear regression model is not adequate to deal with learning problems whose targets are categorical.

The Regression Model

In linear regression, the prevailing assumption is that the target variable (i.e., the unit that we want to predict) can be modeled as a linear combination of the features.

A linear combination is simply the addition of a certain number of vectors that are scaled (or adjusted) by some arbitrary constant. A vector is a mathematical construct for representing a set of numbers.

For example, let us assume a randomly generated dataset consisting of two features and a target variable. The dataset has 50 observations (see Figure 19-1).

© Ekaba Bisong 2019
E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_19

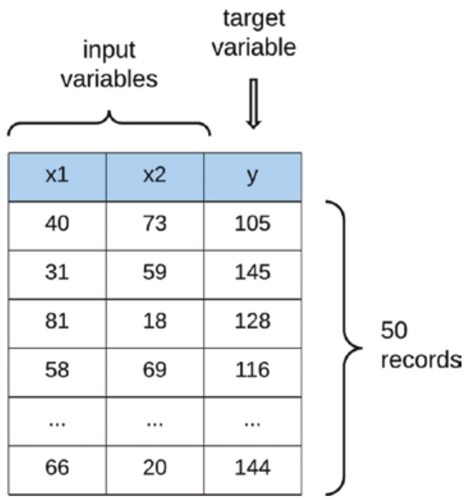

Figure 19-1. *Sample dataset*

The vectors of this dataset are

$$x1 = [40\ 31\ 81\ 57 \ldots 66\ ], \quad x2 = [73\ 59\ 18\ 69 \ldots 20\ ], \quad y = [105\ 145\ 128\ 116 \ldots 144\ ]$$

In a linear regression model, every feature has an assigned "weight." We can say that the weight parameterizes each feature in the dataset. The weights in the dataset are adjusted to take on values that capture the underlying relationship between the features that optimally approximate the target variable. The linear regression model is formally written as

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

where

- $\hat{y}$ (pronounced y-hat) is the approximate value of the output y that we want to predict.
- θ_i, where $i = \{1, 2, \ldots n\}$, is the weight assigned to each feature in the dataset. The notation n is the size of features of the dataset.
- θ_0 represents the "bias" term.

A Visual Representation of Linear Regression

To provide more intuition, let us draw a 2-D plot of the first feature x_1 and the target variable y of the dataset with all 50 records. We are using just one feature in this illustration because it is easier to visualize with a 2-D scatter plot (see Figure 19-2).

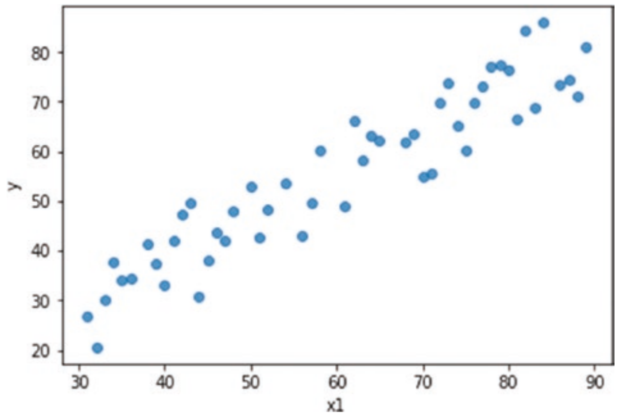

Figure 19-2. *Scatter plot of x_1 (on the x axis) and y (on the y axis)*

The goal of the linear model is to find a line that gives the best approximation or best fit to the data points. When found, this line will look like something in Figure 19-3. The line of best fit is known as the regression line.

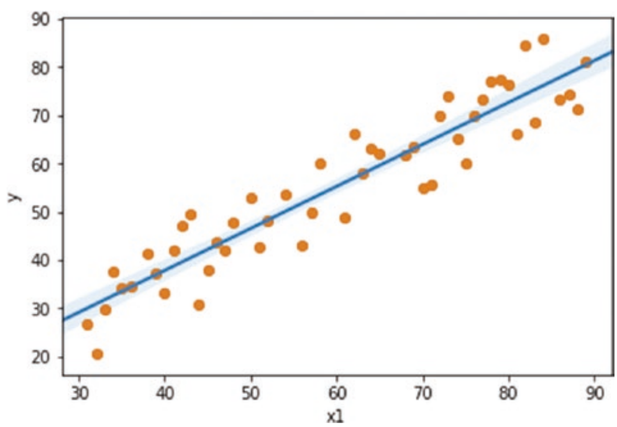

Figure 19-3. *Scatter plot of x_1 (on the x axis) and y (on the y axis) with regression line*