8.1 Finding the Parameters of the Model

After we choose the type of the model to be used we have to find its best parameters, given the data in the training set. Recall that, during the learning, we do not know the values of the target attributes of the instances in the test set, but only the values of their predictive attributes. In our example, the real heights of Omar and Patricia were "hidden from" the learning algorithm; we used them only to measure the predictive performance of the resulting model.

Each learning technique is, in essence, an optimization algorithm that finds the optimal parameters of the corresponding model given some objective function. It means that the type of the model determines the learning algorithm or, in other words, each learning algorithm is designed to optimize a specific type of model. Several regression algorithms have been proposed, most of them in the area of statistics. Next, we briefly describe the most popular ones.

8.1.1 Linear Regression

The linear regression (LR) algorithm is one of the oldest and simplest regression algorithms. Although simple, it is able to induce good regression models, which are easily interpretable.

Let's take a closer look at our model *height* = 128.017 + 0.611 × *weight*. This will allow us to understand the main idea behind LR. Given the notation above, each instance $\mathbf{x}$ is associated with only one attribute, the weight (that's why it is usually called univariate linear regression) and the target attribute y is associated with the height. As we have already shown, this model is the equation of a line in a two-dimensional space. We can see that there are two parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, such that $\hat{\beta}_1$ is associated with the importance of the attribute x_1, the weight. The other parameter, $\hat{\beta}_0$, is called the intercept and is the value of y when the linear model intercepts the y-axis, in other words when $x_1 = 0$.

The result of the optimization process is that this line goes "through the middle" of these instances, represented by points. The objective function, in this case, could be defined as follows. Find the parameters $\hat{\beta}_0, \hat{\beta}_1$ representing a line such that the mean of the squared distance of the points to this line is minimal. Or, in other words, find a model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times x_1$, called a univariate linear model, such that the MSE between y_i and $\hat{y}_i$ is minimal, considering all the instances $(\mathbf{x}_i, y_i)$ in the training set where $i = 1, \ldots, n$. Formally, this can be written as:

$$\underset{\hat{\beta}_0 \hat{\beta}_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y})^2 = \underset{\hat{\beta}_0 \hat{\beta}_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 \times x_i))^2_1 \tag{8.6}$$

Given the training data of our example, the optimal values for $\hat{\beta}_0$ and $\hat{\beta}_1$ are 128.017 and 0.611, respectively.

For multivariate linear regression – the linear model generalized for any number p of predictive attributes – the model is expressed as:

$$\hat{y} = \beta_0 + \sum_{j=1} \beta_j \times x_j \tag{8.7}$$

where p is the number of predictive attributes, $\hat{\beta}_0$ is the value of $\hat{y}$ when all $x_j = 0$ and $\hat{\beta}_i$ is the slope of the linear model according to the jth axis: the variation of $\hat{y}$ per unit of variation of x_j.

Take a quick look at the notation: x_j means the jth attribute of some object $\mathbf{x}$ represented as a tuple $\mathbf{x} = (x_1, \ldots, x_j, \ldots, x_p)$. Since our data con-sists of more instances $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, the ith instance will be denoted as $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_j}, \ldots, x_{i_p})$, x_{i_j} corresponding to its jth attribute.

The values of $\beta_0, \beta_1, \ldots, \beta_p$ are estimated using an appropriate optimization method to minimize the following objective function

$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \sum_{i=1} (y_i - \hat{y})^2 = \underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \sum_{i=1} \left[y_i - \left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta} \times x_{i_j} \right) \right]^2 \tag{8.8}$$

for n instances with p predictive attributes.

8.1.1.1 Empirical Error

Remember that we have only the training set available during the learning, thus, the error is also measured on this set while the model is learned. If we denote the (squared) deviation $(y_i - \hat{y})^2$ as $error(y_i, \hat{y})$, the objective functions, introduced in equations (8.6) and (8.8), considering the given instances $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, can be written as

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{\operatorname{argmin}} \sum_{i=1}^{n} error(y_i, \hat{y}) \tag{8.9}$$

and

$$\underset{\hat{\beta}_0,\ldots,\hat{\beta}_p}{\operatorname{argmin}} \sum_{i=1} error(y_i, \hat{y}) \tag{8.10}$$

respectively for univariate and multivariate linear regression.

The error measured on the training set is called the empirical error or empir- ical loss, and measures the deviation between the predicted ($\hat{y}$) and measured (y_i) values for training instances ($\mathbf{x}_i$).

The assumptions made about the errors – the unexplained variation of y – by the multivariate linear model are:

- they are independent and identically distributed, an assumption that sufferswhen there is collinearity between predictive attributes
- homoscedasticity: there is homogeneous variance, as depicted in Figure 8.7
- normal distribution: a condition that can be verified by comparing the theo-retical distribution and the data distribution.

The breach of these assumptions can result in an inaccurate definition of thecoefficients $\hat{\beta}_i$.

Assessing and evaluating results The main result of MLR is the estimates of the

$\hat{\beta}_i$ coefficients. The $\hat{\beta}_0$ coefficient is often named *coefficient* or $\hat{\alpha}$. Do not forget that knowing the values of all $\hat{\beta}_i$, for $i = 0, \ldots, p$, it is possible to estimate the target value of new unlabeled instances. Moreover, the estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ express the influence of the attribute values $x_1, x_2, \ldots, x_p$ of an instance $\mathbf{x}$ on itstarget value y. The sign of $\hat{\beta}_i$ corresponds to the importance of the attribute $x_i$on y. For positive $\hat{\beta}_i$, the value x_i positively influences the value of y, while for negative $\hat{\beta}_i$ the influence of x_i on y is negative.

Advantages and disadvantages of LR The interpretability of linear regres- sion models is one reason for their popularity. Another is that LR hasno hyper-parameters. Table 8.2 summarizes the main advantages and disadvantages of LR.

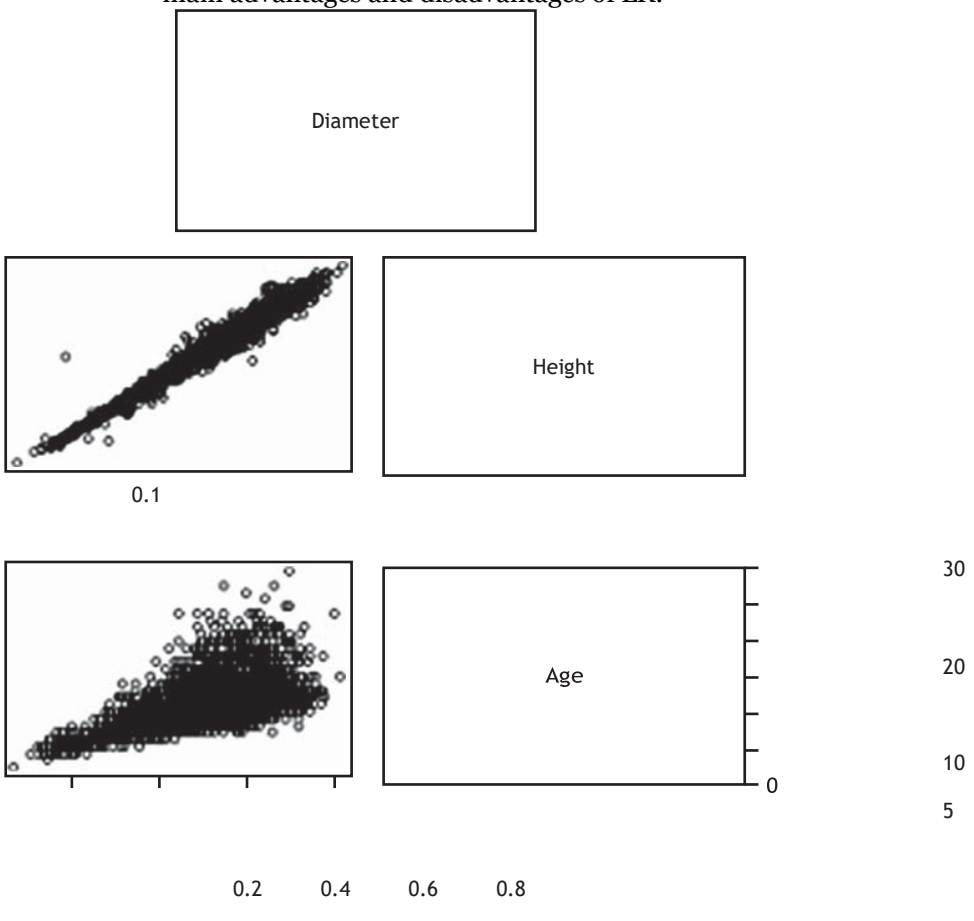

Figure 8.7 Approximately homogeneous variance between diameter and height and nonhomogeneous variance between diameter and age. This data is about the shellfish *Haliotis moluscus*, also known as abalone. It is a food source in several human cultures. The shells are used as decorative items.

Multivariate linear regression suffers when the data set has a large numberof predictive attributes. This can be overcome by using one of the following approaches:

- using an attribute selection method to reduce dimensionality (review theattribute selection methods described in Section 4.5.2)
- shrinking the weights
- using linear combinations of attributes instead of the original predictiveattributes.

Shrinkage methods and linear combination of attributes is discussed next. But before then, there is a discussion of the bias–variance trade-off, an impor- tant concept in understanding shrinkage methods.