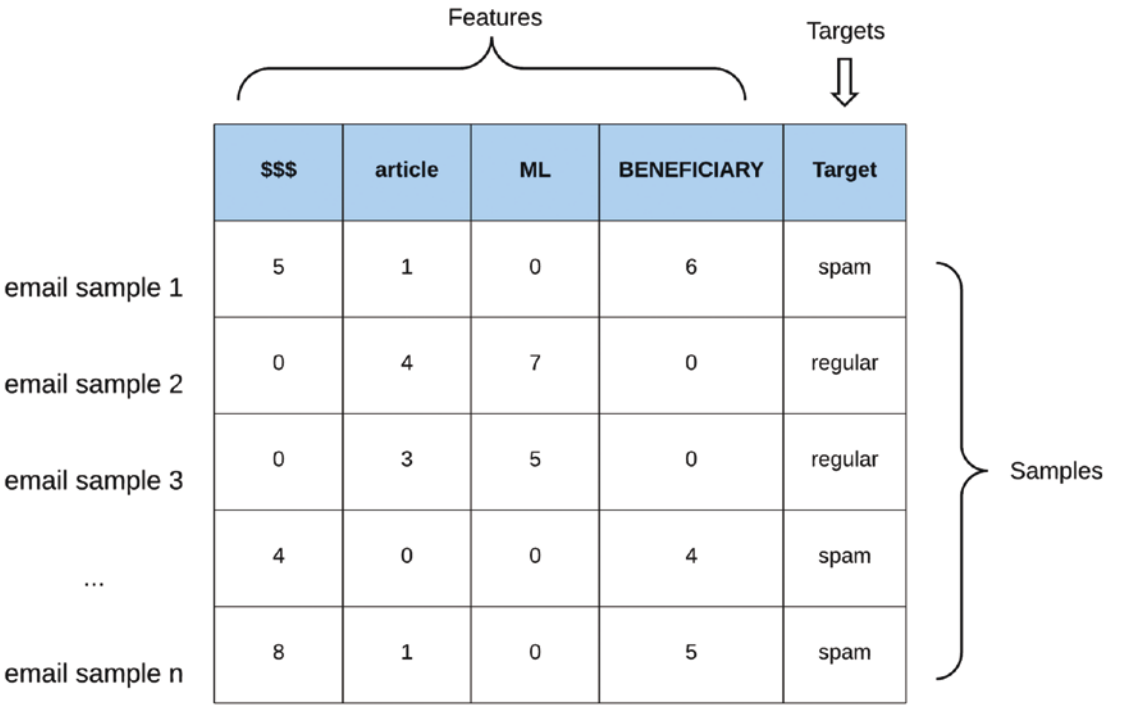

Figure 14-1. *Dataset representation*

The fundamental concept behind supervised machine learning is that each sample is associated with a target variable, and the goal is to teach the computer to learn the patterns from the dataset features that results in a target as a prediction outcome. The columns of a dataset in machine learning are referred to as features; other names you may find commonly used are **variables** or **attributes** of the dataset, but in this book, we will use the term features to describe the measurement units of a data sample. Moreover, the samples of a dataset are also referred to as rows, data points, or observations, but we will use the term samples throughout this book.

Hence, in supervised learning, a set of features are used to build a learning model that will predict the outcome of a target variable as shown in Figure 14-1.

Next, we will cover important modeling considerations for building supervised learning models.

Regression vs. Classification

In supervised learning, we typically have two types of modeling task, and they are **regression** and **classification**.

Regression

The supervised learning problem is a regression task when the values of the target variable are real-valued numbers.

Let's take, for example, that we are given a housing dataset and are asked to build a model that can predict the price of a house. The dataset, for example, has features such as the price of the house, the number of bedrooms, the number of bathrooms, and the total square feet. Let's illustrate how this dataset will look like with a contrived example in Figure 14-2.

Figure 14-2. *Regression problem: housing dataset*

From the learning problem, the features of the dataset are the number of bedrooms, the number of bathrooms, and the square foot of the floor area, while the target feature is the price of the house. The use case presented in Figure 14-3 is framed as a **regression task** because the target feature is a **real-valued number**.

Classification

In a classification task, the target feature is a label denoting some sort of class membership. These labels are also called categorical variables, because they represent labels that belong to two or more categories. Also, no natural ordering exists between the categories or labels.

As an example, suppose we are given a dataset containing the heart disease diagnosis of patients, and we are asked to build a model to predict if a patient has a heart disease or not. Like the previous example, let's assume the dataset has features blood pressure, cholesterol level, heart rate, and heart disease diagnosis. A contrived illustration of this example is shown in Figure 14-3.

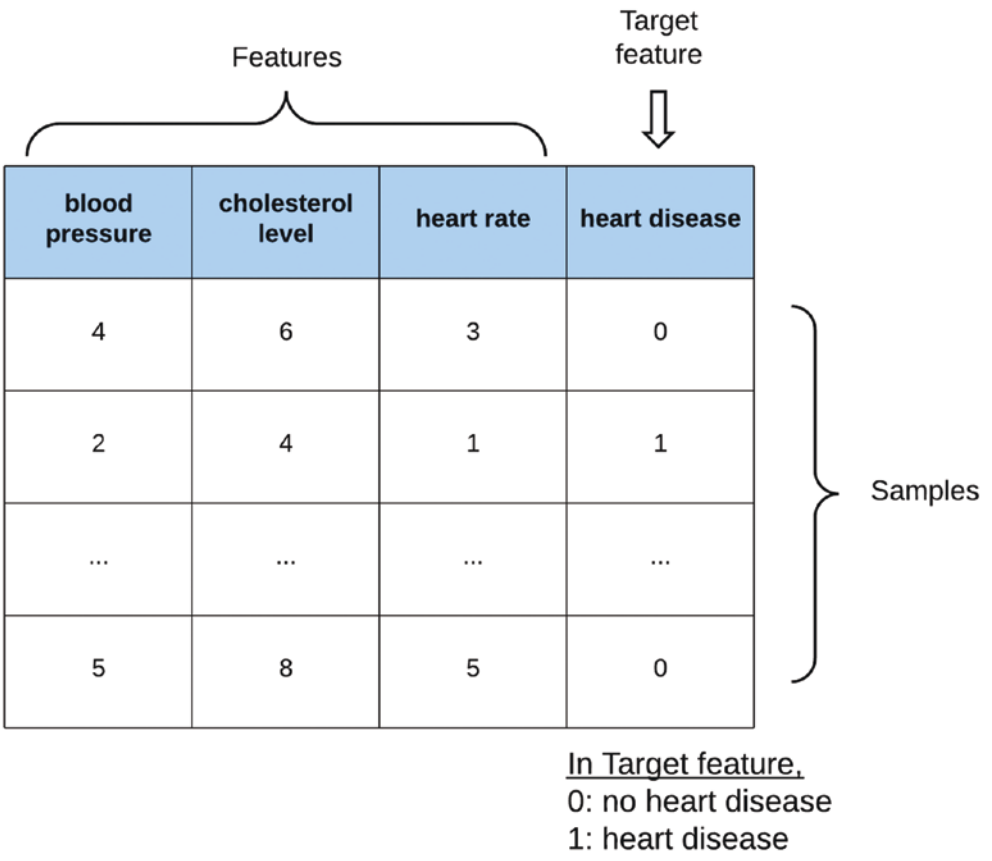

Figure 14-3. *Classification task: heart disease dataset*

From the table in Figure 14-3, the **target variable denotes a class membership of heart disease or no heart disease**; hence, the target is categorical and can be termed as a **classification problem**.

How Do We Know that Learning Has Occurred?

This question is vital to determine if the learning algorithm can learn a useful pattern between the input features and the targets. Let's create a scenario that will give us better insights into appraising the question of determining when learning has occurred.

Assume a teacher takes a physics class for 3 months, and at the end of each session, the teacher administers a test to ascertain if the student has learned anything.

Let's consider two different scenarios the teacher might use in evaluating the students:

1. The teacher evaluates the student with the exact word-for-word questions that were used as sample problems while teaching.
2. The teacher evaluates the student with an entirely different but similar set of sample problems that are based on the principles taught in class.

In which of these subplots can the teacher ascertain that the student has learned? To figure this out, we must consider the two norms of learning:

1. Memorization: In the first subplot, it will be incorrect for the teacher to form a basis for learning because the student has seen and most likely memorized the examples during the class sessions. Memorization is when the exact snapshot of a sample is stored for future recollection. Therefore, it is inaccurate to use samples used in training to carry out learning evaluation. In machine learning, this is known as *data snooping*.
2. Generalization: In the second subplot, the teacher can be confident that the assessment serves as an accurate test to evaluate if the student has learned from the session. The ability to use the principles learned to solve previously unseen samples is known as *generalization*.

Hence, we can conclude that learning is the ability to generalize to previously unseen samples.