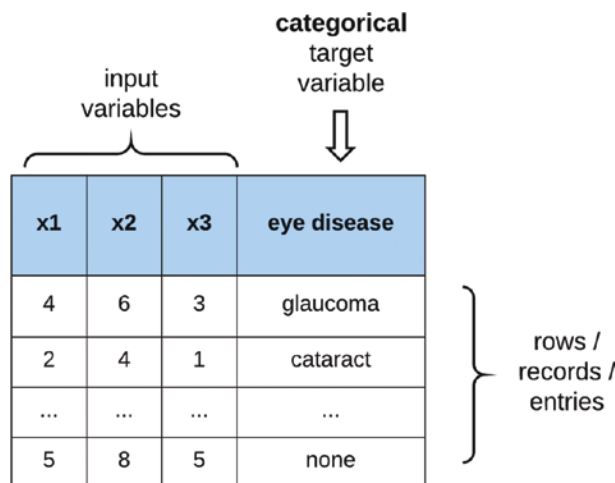


CHAPTER 20

Logistic Regression

Logistic regression is a supervised machine learning algorithm developed for learning classification problems. A classification learning problem is when the target variable is categorical. The goal of logistic regression is to map a function from the features of the dataset to the targets to predict the probability that a new example belongs to one of the target classes. Figure 20-1 is an example of a dataset with categorical targets.



The diagram shows a table with four columns. The first three columns are grouped under the label 'input variables' with a bracket above them. The fourth column is labeled 'categorical target variable' with a downward arrow pointing to it. The table has five rows. The first row is the header, and the following four rows are data entries. A bracket on the right side of the table groups the four data rows under the label 'rows / records / entries'.

input variables			categorical target variable
x1	x2	x3	eye disease
4	6	3	glaucoma
2	4	1	cataract
...
5	8	5	none

Figure 20-1. Dataset with qualitative variables as output

Why Logistic Regression?

To develop our understanding of classification with logistic regression and why linear regression is unsuitable for learning categorical outputs, let us consider a binary or two-class classification problem. The dataset illustrated in Figure 20-2 has the output y (i.e., eye disease) = {disease, no-disease} is an example of dataset with binary targets.

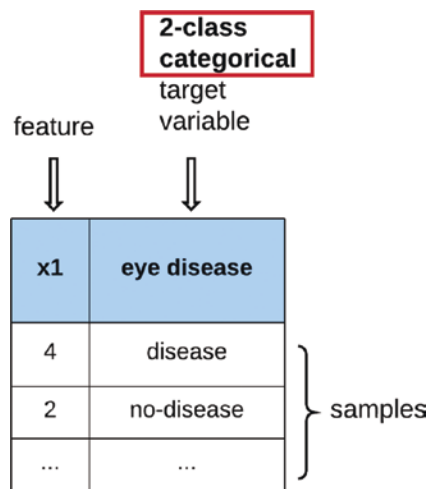


Figure 20-2. Two-class classification problem

From the illustration in Figure 20-3, the linear regression algorithm is susceptible to plot inaccurate decision boundaries especially in the presence of outliers (as seen toward the far right of the graph in Figure 20-3). Moreover, the linear regression model will be looking to learn a real-valued output, whereas a classification learning problem predicts the class membership of an observation using probability estimates.

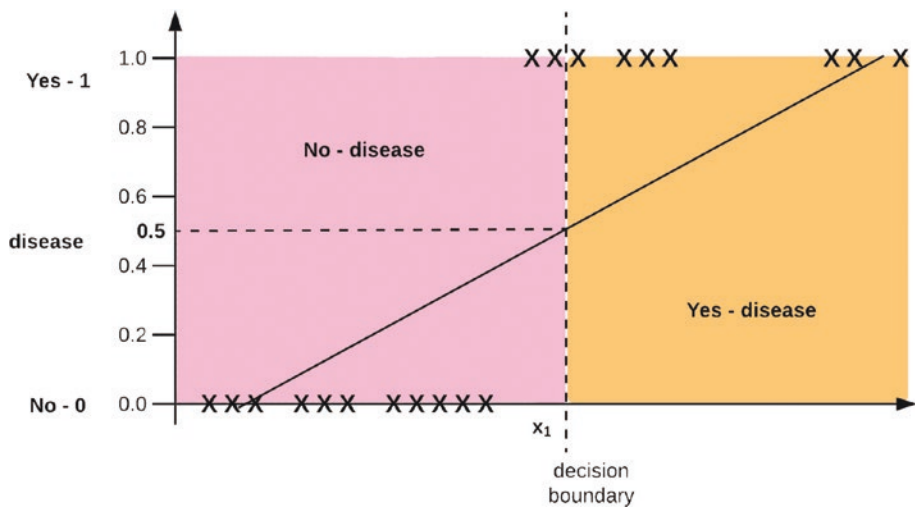


Figure 20-3. Linear regression on a classification dataset

Introducing the Logit or Sigmoid Model

The logistic function, also known as the logit or the sigmoid function, is responsible for constraining the output of the cost function so that it becomes a probability output between 0 and 1. The sigmoid function is formally written as

$$h(t) = \frac{1}{1 + e^{-t}}$$

The logistic regression model is formally similar to the linear regression model except that it is acted upon by the sigmoid model. The following is the formal representation:

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

$$h(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}}$$

where $0 \leq h(t) \leq 1$. The sigmoid function is graphically shown in Figure 20-4.

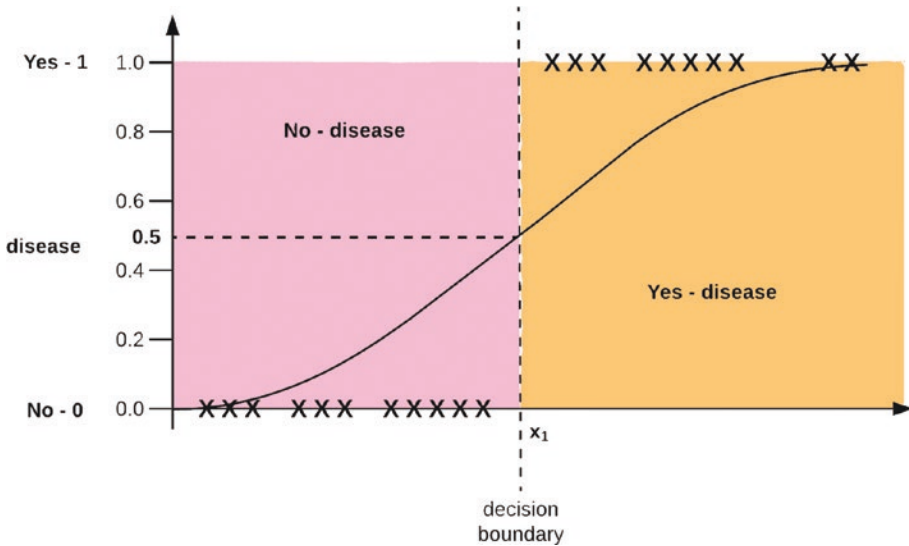


Figure 20-4. Logistic function