**Fully Connected Networks Applied to Multiclass Classification**

In the first three chapters, we used our neural network to solve simple problems that set a foundation for learning deep learning (DL). We reviewed the basic workings of a neuron, how multiple neurons can be connected, and how to devise a suitable learning algorithm. Combining this knowledge, we built a network that can act as an XOR gate – something that arguably can be done in a simpler way.

In this chapter, we finally get to the point where we build a network that does something nontrivial. We show how to build a network that can take an image of a handwritten digit as input, identify which one of the ten digits 0 through 9 the image represents, and present this information on its outputs.

Before showing how to build such a network, we introduce some concepts that are central to both traditional machine learning (ML) and deep learning (DL), namely, datasets and generalization.

91

The programming example also provides more details on how to modify both the networks and the learning algorithm to handle the case of multiclass classification. This modification is needed because recognizing handwritten digits implies that each input example needs to be classified as belonging to one of ten classes.

**Introduction to Datasets Used When Training Networks**

As we saw in the previous chapters, we train a neural network by presenting an input example to the network. We then compare the network output to the expected output and use gradient descent to adjust the weights to try to make the network provide the correct output for a given input. A reasonable question is from where to get these training examples that are needed to train the network. For our previous toy examples this was not an issue. A two-input XOR gate has only four input combinations, so we could easily create a list of all combinations. This assumes that we interpret the input and the output values as binary variables, which typically would not be the case but was true in our toy example.

In real applications of DL, obtaining these training examples can be a big challenge. One of the key reasons that DL has gained so much traction lately is that large online databases of images, videos, and natural language text have made it possible to obtain large sets of training data. If a supervised learning technique is used, it is not sufficient to obtain the input to the network. We also need to know the expected output, the ground truth, for each example. The process of associating each training input with an expected output is known as *labeling*, which is often a manual process. That is, a human must add a label to each example, detailing whether it is a dog, a cat, or a car. This process can be tedious because we often need many thousands of examples to achieve good results.

Starting to experiment with DL might be hard if the first step involved putting together a large collection of labeled training examples. Fortunately, other people have already done so and have made these examples publicly available. This is where the concept of datasets comes in. A (labeled) dataset consists of a collection of labeled training examples that can be used for training ML models. In this book, we will become familiar with a handful of different datasets within the fields of images, historical housing-price data, and natural

92

Languages. A section about datasets would not be complete without mentioning the classic Iris Dataset (Fisher, 1936), which is likely the first widely available dataset. It contains 150 instances of iris flowers, each instance belonging to one of three iris species. Each instance consists of four measurements (sepal length and width, petal length and width) of the particular plant. The Iris Dataset is extremely small and simple, so instead we start with a more complicated, although still simple, dataset: the Modified National Institute of Standards and Technology (MNIST) database of handwritten digits, also known simply as the MNIST dataset.

The MNIST dataset contains 60,000 training images and 10,000 test images. (We detail the differences between training and test images later in the chapter.) In addition to the images, the dataset consists of labels that describe which digit each image represents. The original images are 32×32 pixels, and the outermost two pixels around each image are blank, so the actual image content is found in the centered 28×28 pixels. In the version of the dataset that we use, the blank pixels have been stripped out, so each image is 28×28 pixels. Each pixel is represented by a grayscale value ranging from 0 to 255. The source of the handwritten digits is a mix of employees at the American Census Bureau and American high school students. The dataset was made available in 1998 (LeCun, Bottou, Bengio, et al., 1998). Some of the training examples are shown in Figure 4-1.

Figure 4-1 Images from the MNIST dataset. (Source: LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. “Gradient-Based Learning Applied to Document Recognition” in *Proceedings of the IEEE* vol. 86, no. 11 (Nov. 1998), pp. 2278-2324.)

93

**Exploring the Dataset**

We start with getting our hands dirty by exploring the dataset a little bit.