The Role of Data

Data is at the core of machine learning. It is central to the current evolution and further advancement of this field. Just as it is for humans, it is the same way for machines. Learning is not possible without data.

Humans learn how to perform tasks by collecting information from the Environment. This information is the data the brain uses to construct patterns and gain an understanding of the Environment. For a human being, data is captured through the sense organs. For example, the eyes capture visual data, the ears capture auditory data, the skin receives tactile data, while the nose and tongue detect olfactory and taste data, respectively.

As with humans, this same process of learning from data is replicated with machines. Let's take, for example, the task of identifying spam emails. In this example, the computer is provided email examples as data. It then uses an algorithm to learn to distinguish spam emails from regular emails.

The Cost of Data

Data is expensive to collect, and high-quality data is even more costly to capture due to the associated costs in storing and cleaning the data. Over the years, the paucity of data had limited the performance of machine learning methods. However, in the early 1990s, the Internet was born, and by the dawn of the century, it became a super highway for data distribution. As a result, large and diverse data became readily available for the research and development of machine learning products across various domains.

In this chapter, we covered the definition and history of machine learning and the importance of data. Next, we will take it further by discussing the principles of machine learning in Chapter 14.

CHAPTER 14

Principles of Learning

Machine learning is, for the most part, sub-divided into three components based on the approach to the learning problem. The three predominant categories of learning are the supervised, unsupervised, and reinforcement learning schemes. In this chapter, we will go over supervised learning schemes in detail and also touch upon unsupervised and reinforcement learning schemes to a lesser extent.

The focus on supervised learning is for a variety of reasons. Firstly, they are the predominant techniques used for building machine learning products in industry; secondly, as you will soon learn, they are easy to ground truth and assess their performances before being deployed as part of a large-scale production pipeline. Let's examine each of the three schemes.

Supervised Learning

To easily understand the concept of supervised learning, let's revisit the problem of identifying spam emails from a set of emails. We will use this example to understand key concepts that are central to the definition and the framing of a supervised learning problem, and they are

- Features
- Samples
- Targets

For this contrived example, let's assume that we have a dictionary of the top 4 words in the set of emails and we record the frequency of occurrence for each email sample. This information is represented in a tabular format, where each feature is a column and the rows are email samples. This tabular representation is called a dataset. Figure 14-1 illustrates this depiction.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_14