Loading Data

Loading data is an important process in the data analysis/machine learning pipeline. Data usually comes in **.csv** format. **csv** files can be loaded into Python by using the **loadtxt** method. The parameter **skiprows** skips the first row of the dataset – it is usually the header row of the data.

```
np.loadtxt(open("the_file_name.csv", "rb"), delimiter=",", skiprows=1)
```

Pandas is a preferred package for loading data in Python.
We will learn more about Pandas for data manipulation in the next chapter.

CHAPTER 11

Pandas

Pandas is a specialized Python library for data analysis, especially on humongous datasets. It boasts easy-to-use functionality for reading and writing data, dealing with missing data, reshaping the dataset, and massaging the data by slicing, indexing, inserting, and deleting data variables and records. Pandas also has an important **groupBy** functionality for aggregating data for defined conditions – useful for plotting and computing data summaries for exploration.

Another key strength of Pandas is in re-ordering and cleaning time series data for time series analysis. In short, Pandas is the go-to tool for data cleaning and data exploration.

To use Pandas, first import the Pandas module:

```
import pandas as pd
```

Pandas Data Structures

Just like NumPy, Pandas can store and manipulate a multi-dimensional array of data. To handle this, Pandas has the **Series** and **DataFrame** data structures.

Series

The **Series** data structure is for storing a 1-D array (or vector) of data elements. A series data structure also provides labels to the data items in the form of an **index**. The user can specify this label via the **index** parameter in the **Series** function, but if the **index** parameter is left unspecified, a default label of 0 to one minus the size of the data elements is assigned.

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_11