CHAPTER 1

Statistics and Probability

The purpose of this chapter is to instill in you the basic concepts of traditional statistics and probability. Certainly many of you might be wondering what it has to do with machine learning. Well, in order to apply a best fit model to your data, the most important prerequisite is for you to understand the data in the first place. This will enable you to find out distributions within data, measure the goodness of data, and run some basic tests to understand if some form of relationship exists between dependant and independent variables. Let's dive in.

■ **Note** This book incorporates **Python 2.7.11** as the de facto standard for coding examples. Moreover, you are required to have it installed it for the *Exercises* as well.

So why do I prefer Python 2.7.11 over Python 3x? Following are some of the reasons:

- Third-party library support for Python 2x is relatively better than support for Python 3x. This means that there are a considerable number of libraries in Python 2x that lack support in Python 3x.
- Some current Linux distributions and macOS provide Python 2x by default. The objective is to let readers, regardless of their OS version, apply the code examples on their systems, and thus this is the choice to go forward with.
- The above-mentioned facts are the reason why companies prefer to work with Python 2x or why they decide not to migrate their code base from Python 2x to Python 3x.

Case Study: Cycle Sharing Scheme—Determining Brand Persona

Nancy and Eric were assigned with the huge task of determining the brand persona for a new cycle share scheme. They had to present their results at this year's annual board meeting in order to lay out a strong marketing plan for reaching out to potential customers.

D. Haroon, *Python Machine Learning Case Studies*, DOI 10.1007/978-1-4842-2823-4_1

The cycle sharing scheme provides means for the people of the city to commute using a convenient, cheap, and green transportation alternative. The service has 500 bikes at 50 stations across Seattle. Each of the stations has a dock locking system (where all bikes are parked); kiosks (so customers can get a membership key or pay for a trip); and a helmet rental service. A person can choose between purchasing a membership key or short-term pass. A membership key entitles an annual membership, and the key can be obtained from a kiosk. Advantages for members include quick retrieval of bikes and unlimited 45-minute rentals. Short-term passes offer access to bikes for a 24-hour or 3-day time interval. Riders can avail and return the bikes at any of the 50 stations citywide.

Jason started this service in May 2014 and since then had been focusing on increasing the number of bikes as well as docking stations in order to increase convenience and accessibility for his customers. Despite this expansion, customer retention remained an issue. As Jason recalled, "We had planned to put in the investment for a year to lay out the infrastructure necessary for the customers to start using it. We had a strategy to make sure that the retention levels remain high to make this model self-sustainable. However, it worked otherwise (i.e., the customer base didn't catch up with the rate of the infrastructure expansion)."

A private service would have had three alternatives to curb this problem: get sponsors on board, increase service charges, or expand the pool of customers. Price hikes were not an option for Jason as this was a publicly sponsored initiative with the goal of providing affordable transportation to all. As for increasing the customer base, they had to decide upon a marketing channel that guarantees broad reach on low cost incurred.

Nancy, a marketer who had worked in the corporate sector for ten years, and Eric, a data analyst, were explicitly hired to find a way to make things work around this problem. The advantage on their side was that they were provided with the dataset of transaction history and thus they didn't had to go through the hassle of conducting marketing research to gather data.

Nancy realized that attracting recurring customers on a minimal budget required understanding the customers in the first place (i.e., persona). As she stated, "Understanding the persona of your brand is essential, as it helps you reach a targeted audience which is likely to convert at a higher probability. Moreover, this also helps in reaching out to sponsors who target a similar persona. This two-fold approach can make our bottom line positive."

As Nancy and Eric contemplated the problem at hand, they had questions like the following: Which attribute correlates the best with trip duration and number of trips? Which age generation adapts the most to our service?

Following is the data dictionary of the *Trips* dataset that was provided to Nancy and Eric:

Table 1-1. *Data Dictionary for the Trips Data from Cycles Share Dataset*

Feature name	Description
trip_id	Unique ID assigned to each trip
Starttime	Day and time when the trip started, in PST
Stoptime	Day and time when the trip ended, in PST
Bikeid	ID attached to each bike
Tripduration	Time of trip in seconds
from_station_name	Name of station where the trip originated
to_station_name	Name of station where the trip terminated
from_station_id	ID of station where trip originated
to_station_id	ID of station where trip terminated
Usertype	Value can include either of the following: short-term pass holder or member
Gender	Gender of the rider
Birthyear	Birth year of the rider

Exercises for this chapter required Eric to install the packages shown in Listing 1-1. He preferred to import all of them upfront to avoid bottlenecks while implementing the code snippets on your local machine.

However, for Eric to import these packages in his code, he needed to install them in the first place. He did so as follows:

1. Opened terminal/shell
2. Navigated to his code directory using terminal/shell
3. Installed `pip`:

   ```
   python get-pip.py
   ```

4. Installed each package separately, for example:

   ```
   pip install pandas
   ```

Listing 1-1. Importing Packages Required for This Chapter

```
%matplotlib inline

import random
import datetime
import pandas as pd
import matplotlib.pyplot as plt
import statistics
```

```
import numpy as np
import scipy
from scipy import stats
import seaborn
```

Performing Exploratory Data Analysis

Eric recalled to have explained Exploratory Data Analysis in the following words:

> *What do I mean by exploratory data analysis (EDA)? Well, by this I mean to see the data visually. Why do we need to see the data visually? Well, considering that you have 1 million observations in your dataset then it won't be easy for you to understand the data just by looking at it, so it would be better to plot it visually. But don't you think it's a waste of time? No not at all, because understanding the data lets us understand the importance of features and their limitations.*

Feature Exploration

Eric started off by loading the data into memory (see Listing 1-2).

Listing 1-2. Reading the Data into Memory

```
data = pd.read_csv('examples/trip.csv')
```

Nancy was curious to know how big the data was and what it looked like. Hence, Eric wrote the code in Listing 1-3 to print some initial observations of the dataset to get a feel of what it contains.

Listing 1-3. Printing Size of the Dataset and Printing First Few Rows

```
print len(data)
data.head()
```

Output

```
236065
```

Table 1-2. *Print of Observations in the First Seven Columns of Dataset*

trip_id	starttime	stoptime	bikeid	tripduration	from_station_name	to_station_name
431	10/13/2014 10:31	10/13/2014 10:48	SEA00298	985.935	2nd Ave & Spring St	Occidental Park/ Occidental Ave S & S Washing...
432	10/13/2014 10:32	10/13/2014 10:48	SEA00195	926.375	2nd Ave & Spring St	Occidental Park/ Occidental Ave S & S Washing...
433	10/13/2014 10:33	10/13/2014 10:48	SEA00486	883.831	2nd Ave & Spring St	Occidental Park/ Occidental Ave S & S Washing...
434	10/13/2014 10:34	10/13/2014 10:48	SEA00333	865.937	2nd Ave & Spring St	Occidental Park/ Occidental Ave S & S Washing...
435	10/13/2014 10:34	10/13/2014 10:49	SEA00202	923.923	2nd Ave & Spring St	Occidental Park/ Occidental Ave S & S Washing...

Table 1-3. *Print of Observations in the Last five Columns of Dataset*

from_station_id	to_station_id	usertype	gender	birthyear
CBD-06	PS-04	Member	Male	1960.0
CBD-06	PS-04	Member	Male	1970.0
CBD-06	PS-04	Member	Female	1988.0
CBD-06	PS-04	Member	Female	1977.0
CBD-06	PS-04	Member	Male	1971.0