The Data Science Process

The data science process involves components for data ingestion and serving of data models. However, we will discuss briefly on the steps for carrying out data analytics in lieu of data prediction modeling.

These steps consist of

1. Data summaries: The vital statistical summaries of the datasets' variables or features. This includes information such as the number of variables, their data types, the number of observations, and the count/percentage of missing data.
2. Data visualization: This involves employing univariate and multivariate data visualization methods to get a better intuition on the properties of the data variables and their relationship with each other. This includes metrics such as histograms, box and whisker plots, and correlation plots.
3. Data cleaning/preprocessing: This process involves sanitizing the data to make it amenable for modeling. Data rarely comes clean with each row representing an observation and each column an entity. In this phase of a data science effort, the tasks involved may include removing duplicate entries, choosing a strategy for dealing with missing data, as well as converting data features into numeric data types of encoded categories. This phase may also involve carrying out statistical transformation on the data features to normalize and/or standardize the data elements. Data features of wildly differing scales can lead to poor model results as they become more difficult for the learning algorithm to converge to the global minimum.
4. Feature engineering: This practice involves systematically pruning the data feature space to only select those features relevant to the modeling problem as part of the model task. Good feature engineering is often the difference between an average and high performant model.

5. Data modeling and evaluation: This phase involves passing the data through a learning algorithm to build a predictive model. This process is usually an iterative process that involves constant refinement in order to build a model that better minimizes the cost function on the hold-out validation set and the test set.

In this chapter, we provided a brief overview to the concept of data science, the challenge of big data, and its goal to unlock value from data. The next chapter will provide an introduction to programming with Python.

CHAPTER 9

Python

Python is one of the preferred languages for data science in the industry primarily because of its simple syntax and the number of reusable machine learning/deep learning packages. These packages make it easy to develop data science products without getting bogged down with the internals of a particular algorithm or method. They have been written, debugged, and tested by the best experts in the field, as well as by a large supporting community of developers that contribute their time and expertise to maintain and improve them.

In this section, we will go through the foundations of programming with Python 3. This section forms a framework for working with higher-level packages such as NumPy, Pandas, Matplotlib, TensorFlow, and Keras. The programming paradigm we will cover in this chapter can be easily adapted or applied to similar languages, such as R, which is also commonly used in the data science industry.

The best way to work through this chapter and the successive chapters in this part is to work through the code by executing them on Google Colab or GCP Deep Learning VMs.

Data and Operations

Fundamentally, programming involves storing data and operating on that data to generate information. Techniques for efficient data storage are studied in the field called data structures, while the techniques for operating on data are studied as algorithms.

Data is stored in a memory block on the computer. Think of a memory block as a container holding data (Figure 9-1). When data is operated upon, the newly processed data is also stored in memory. Data is operated by using arithmetic and boolean expressions and functions.

© Ekaba Bisong 2019
E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_9