Introduction

Machine learning and deep learning technologies have impacted the world in profound ways, from how we interact with technological products and with one another. These technologies are disrupting how we relate, how we work, and how we engage life in general. Today, and in the foreseeable future, intelligent machines increasingly form the core upon which sociocultural and socioeconomic relationships rest. We are indeed already in the "age of intelligence."

What Are Machine Learning and Deep Learning?

Machine learning can be described as an assortment of tools and techniques for predicting or classifying a future event based on a set of interactions between variables (also referred to as features or attributes) in a particular dataset. Deep learning, on the other hand, extends a machine learning algorithm called neural network for learning complex tasks which are incredibly difficult for a computer to perform. Examples of these tasks may include recognizing faces and understanding languages in their varied contextual meanings.

The Role of Big Data

A key ingredient that is critical to the rise and future improved performance of machine learning and deep learning is data. Since the turn of the twenty-first century, there has been a steady exponential increase in the amount of data generated and stored. The rise of humongous data is partly due to the emergence of the Internet and the miniaturization of processors that have spurned the "Internet of Things (IoT)" technologies. These vast amounts of data have made it possible to train the computer to learn complex tasks where an explicit instruction set is infeasible.

The Computing Challenge

The increase in data available for training learning models throws up another kind of problem, and that is the availability of computational or processing power. Empirically, as data increases, the performance of learning models also goes up. However, due to the increasingly enormous size of datasets today, it is inconceivable to train sophisticated, state-of-the-art learning models on commodity machines.

Cloud Computing to the Rescue

Cloud is a term that is used to describe large sets of computers that are networked together in groups called data centers. These data centers are often distributed across multiple geographical locations. Big companies like Google, Microsoft, Amazon, and IBM own massive data centers where they manage computing infrastructure that is provisioned to the public (i.e., both enterprise and personal users) for use at a very reasonable cost.

Cloud technology/infrastructure is allowing individuals to leverage the computing resources of big business for machine learning/deep learning experimentation, design, and development. For example, by making use of cloud resources such as Google Cloud Platform (GCP), Amazon Web Services (AWS), or Microsoft Azure, we can run a suite of algorithms with multiple test grids for a fraction of time that it will take on a local machine.

Enter Google Cloud Platform (GCP)

One of the big competitors in the cloud computing space is Google, with their cloud resource offering termed "Google Cloud Platform," popularly referred to as GCP for short. Google is also one of the top technology leaders in the Internet space with a range of leading web products such as Gmail, YouTube, and Google Maps. These products generate, store, and process tons of terabytes of data each day from Internet users around the world.

To deal with this significant data, Google over the years has invested heavily in processing and storage infrastructure. As of today, Google boasts some of the most impressive data center design and technology in the world to support their

computational demands and computing services. Through Google Cloud Platform, the public can leverage these powerful computational resources to design and develop cutting-edge machine learning and deep learning models.

The Aim of This Book

The goal of this book is to equip the reader from the ground up with the essential principles and tools for building learning models. Machine learning and deep learning are rapidly evolving, and often it is overwhelming and confusing for a beginner to engage the field. Many have no clue where to start. This book is a one-stop shop that takes the beginner on a journey to understanding the theoretical foundations and the practical steps for leveraging machine learning and deep learning techniques on problems of interest.

Book Organization

This book is divided into eight parts. Their breakdown is as follows:

- Part 1: Getting Started with Google Cloud Platform
- Part 2: Programming Foundations for Data Science
- Part 3: Introducing Machine Learning
- Part 4: Machine Learning in Practice
- Part 5: Introducing Deep Learning
- Part 6: Deep Learning in Practice
- Part 7: Advanced Analytics/Machine Learning on Google Cloud Platform
- Part 8: Productionalizing Machine Learning Solutions on GCP

It is best to go through the entire book in sequence. However, each part and its containing chapters are written in such a way that one can shop around and get out what is of primary interest. The code repository for this book is available at `https://github.com/Apress/building-ml-and-dl-models-on-gcp`. The reader can follow through the examples in this book by cloning the repository to Google Colab or GCP Deep Learning VM.

PART I

Getting Started with Google Cloud Platform