

Preface

My foray in machine learning started in 1992, while working on my Masters thesis titled *Predicting torsional vibration response of a marine power transmission shaft*. The model was based on an iterative procedure using the Newton–Raphson rule to optimize a continuum of state vectors defined by transfer matrices. The optimization algorithm was written using the C programming language and it introduced me to the power of machines in numerical computation and its vulnerability to floating point errors. Although the term “machine learning” came much later intuitively, I was using the power of an 8088 chip on my mathematical model to predict a response.

Much later, I started using different optimization techniques using computers both in the field of engineering and business. All through I kept making my own notes. At some point of time, I thought it was a good idea to organize my notes, put some thought on the subject, and write a book which covers the essentials of machine learning—linear algebra, statistics, and learning algorithms.

The Data-Driven Universe

Galileo in his *Discorsi* [1638] stated that data generated from natural phenomena can be suitably represented through mathematics. When the size of data was small, then, we could identify the obvious patterns. Today, a new era is emerging where we are “downloading the universe” to analyze data and identify more subtle patterns.

The Merriam Webster dictionary defines the word “cognitive”, as “*relating to, or involving conscious mental activities like learning*”. The American philosopher of technology and founding executive editor of *Wired*, Kevin Kelly, defines “cognitize” as injecting intelligence to everything we do, through machines and algorithms. The ability to do so depends on data, where intelligence is a stowaway in the data cloud. In the data-driven universe, therefore, we are not just using data but constantly seeking new data to extract knowledge.

Causality—The Cornerstone of Accountability

Smart learning technologies are better at accomplishing tasks but they do not think. They can tell us “what” is happening but they cannot tell us “why”. They may tell us that some stromal tissues are important in identifying breast cancer but they lack the cause behind why some tissues are playing the role. Causality, therefore, is the rub.

The Growth of Machines

For the most enthusiastic geek, the default mode just 30 years ago from today was offline. Moore’s law has changed that by making computers smaller and faster, and in the process, transforming them from room-filling hardware and cables to slender and elegant tablets. Today’s smartphone has the computing power, which was available at the MIT campus in 1950. As the demand continues to expand, an increasing proportion of computing is taking place in far-off warehouses thousands of miles away from the users, which is now called “cloud computing”—*de facto* if not *de jure*. The massive amount of cloud-computing power made available by Amazon and Google implies that the speed of the chip on a user’s desktop is becoming increasingly irrelevant in determining the kind of things a user can do.

Recently, AlphaGo, a powerful artificial intelligence system built by Google, defeated Lee Sedol, the world’s best player of Go. AlphaGo’s victory was made possible by clever machine intelligence, which processed a data cloud of 30 million moves and played thousands of games against itself, “learning” each time a bit more about how to improve its performance. A learning mechanism, therefore, can process enormous amounts of data and improve their performance by analyzing their own output as input for the next operation(s) through **machine learning**.

What is Machine Learning?

This book is about data mining and machine learning which helps us to discover previously unknown patterns and relationships in data. Machine learning is the process of automatically discovering patterns and trends in data that go beyond simple analysis. Needless to say, sophisticated mathematical algorithms are used to segment the data and to predict the likelihood of future events based on past events, which cannot be addressed through simple query and reporting techniques.

There is a great deal of overlap between learning algorithms and statistics and most of the techniques used in learning algorithms can be placed in a statistical framework. Statistical models usually make strong assumptions about the data and, based on those assumptions, they make strong statements about the results.

However, if the assumptions in the learning model are flawed, the validity of the model becomes questionable. Machine learning transforms a small amount of input knowledge into a large amount of output knowledge. And, the more knowledge from (*data*) we put in, we get back that much more knowledge out. Iteration is therefore at the core of machine learning, and because we have constraints, the driver is optimization.

If the knowledge and the data are not sufficiently complete to determine the output, we run the risk of having a model that is not “real”, and is a foible known as *overfitting* or *underfitting* in machine learning.

Machine learning is related to artificial intelligence and deep learning and can be segregated as follows:

- **Artificial Intelligence** (AI) is the broadest term applied to any technique that enables computers to mimic human intelligence using logic, if-then rules, decision trees, and machine learning (including deep learning).
- **Machine Learning** is the subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with the experience gained while executing the tasks. If we have input data x and want to find the response y , it can be represented by the function $y = f(x)$. Since it is impossible to find the function f , given the data and the response (due to a variety of reasons discussed in this book), we try to approximate f with a function g . The process of trying to arrive at the best approximation to f is through a process known as machine learning.
- **Deep Learning** is a scalable version of machine learning. It tries to expand the possible range of estimated functions. If machine learning can learn, say 1000 models, deep learning allows us to learn, say 10000 models. Although both have infinite spaces, deep learning has a larger viable space due to the math, by exposing multilayered neural networks to vast amounts of data.

Machine learning is used in web search, spam filters, recommender systems, credit scoring, fraud detection, stock trading, drug design, and many other applications. As per Gartner, AI and machine learning belong to the top 10 technology trends and will be the driver of the next big wave of innovation.¹

Intended Audience

This book is intended both for the newly initiated and the expert. If the reader is familiar with a little bit of code in R, it would help. R is an open-source statistical programming language with the objective to make the analysis of empirical and simulated data in science reproducible. The first three chapters lay the foundations of machine learning and the subsequent chapters delve into the mathematical

¹<http://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017/>.

interpretations of various algorithms in regression, classification, and clustering. These chapters go into the detail of supervised and unsupervised learning and discuss, from a mathematical framework, how the respective algorithms work. This book will require readers to read back and forth. Some of the difficult topics have been cross-referenced for better clarity. The book has been written as a first course in machine learning for the final-term undergraduate and the first-term graduate levels. This book is also ideal for self-study and can be used as a reference book for those who are interested in machine learning.

Kolkata, India
August 2017

Abhijit Ghatak



<http://www.springer.com/978-981-10-6807-2>

Machine Learning with R

Ghatak, A.

2017, XIX, 210 p. 56 illus., Hardcover

ISBN: 978-981-10-6807-2