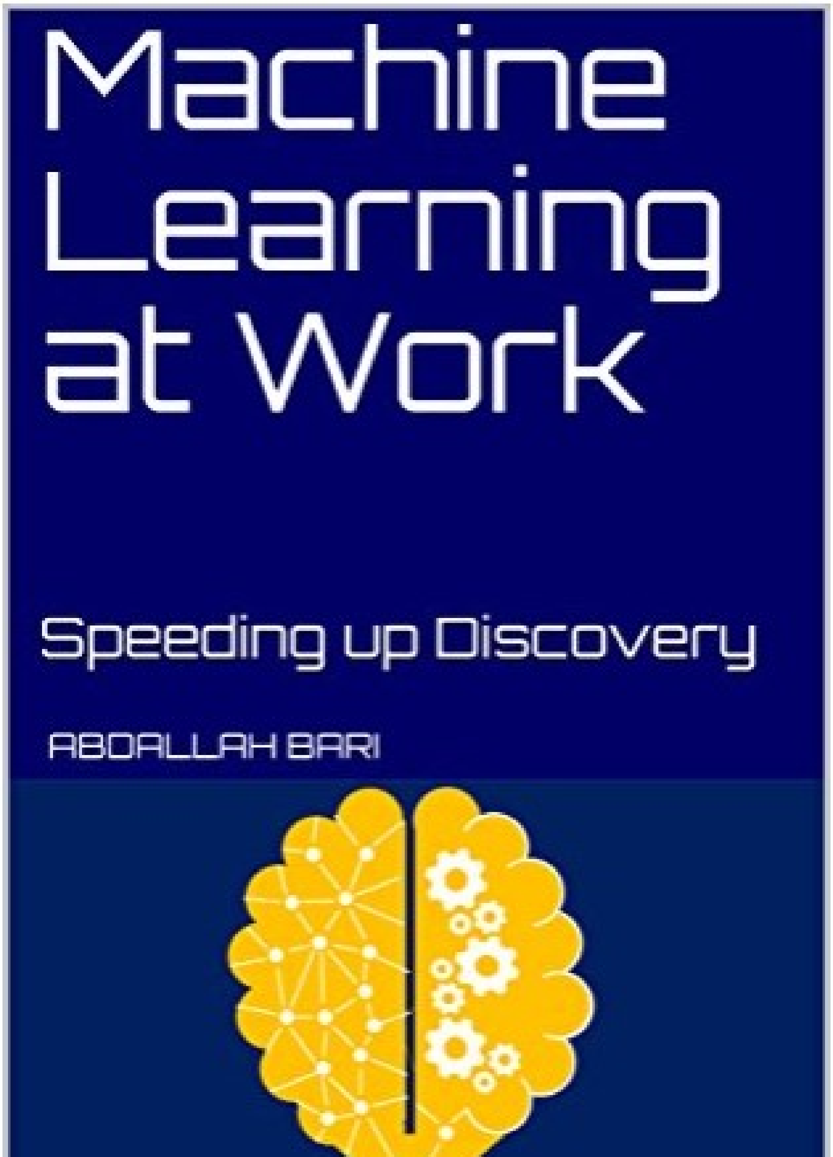

Machine Learning at Work

Speeding up Discovery

ABDALLAH BARI

Machine Learning

@

Work

Speeding
up
Discovery

Abdallah Bari

ASIN: B07BRYQWPL

To Leila, Tarek, Nour and Jad;
as well as John, Coral, Julie, Karen and Katherine
.

Product or corporate names may be trademarks and/or registered trademarks of their respective companies, and they are used here for identification and explanation purposes.

Quotes are mentioned to support some of the arguments put forward as well as for the purpose of commentary explanation. Some of the quotes were repeated in their original languages to keep the original meaning intended by their respective authors.

The sources of quotes, text and/or data are acknowledged and credited for throughout the book and listed in the foot notes. However, if any copyright material, including data, have not been acknowledged or credited for, grateful if you could notify to rectify them in the subsequent updates and editions.

Math Coding and Analytics
Westmount, Quebec, Canada
ASIN: B07BRYQWPL

ACKNOWLEDGMENTS

I am grateful to many great colleagues and friends, since I have started my career as a researcher, starting from Cyprus.

I am also grateful to my great teachers including my math teachers while at high school and my supervisors during my university studies, including my studies on machine learning applications at Cordoba University in Spain.

PREFACE

Today, machine learning (ML) and artificial intelligence (AI) are almost everywhere. They both existed before, even before the Internet, however they stayed dormant until they re-emerged, in recent years.

One of the reasons for their big comeback is because of Big Data, which has enabled their potential. ML and in particular its subfield Deep Learning performs well when feed with large datasets. The second reason is the dramatic increase in the processing computing power, with the emergence of graphic processing unit (GPU), which is allowing to process large amount of data in parallel.

As the processing power continue to progress, Machine Learning will also continue to increase the ability to handle the ever increasing amounts of Big Data and to help turn Big Data into actionable insights. Today, ML with its Deep Learning subset, are becoming priority for research and development (R&D), including research for speeding up discovery and time delivery processes.

Discovery and time delivery processes can be long and costly processes and in the case of discovery the outcome can even be uncertain. ML is poised to help by reducing dramatically these processes and by shortening the time and reducing dramatically the costs that may be incurred during such processes.

I have worked with colleagues over many years using ML in combination of imaging techniques to assess subtle variation among thousands of images. The variation captured and detected among these thousands of images was so subtle that the eye cannot capture all the hidden variation. Prior to the use of ML, thousands of images were processed and pertinent features were extracted using fractal geometry, then ML was used, leading to accurate identification of up to 90% of variation. The potential value of ML techniques combined with fractals also lies in their ability to mimic natural complexity.

Fractals have helped vastly in today's information technology with its use in mobile phone. Fractal-based models helped in the description and classification of scale-related phenomena in life sciences, from molecular to ecosystem levels of organization.

Many patterns can be complex and "non-accretive" components, although the eye has the capacity to discern features such as texture and shape in images, the values that are assigned to score these features can be subjective. Studies on scoring methodology, based on visual examination, have shown that the assignment to categories or classes can be problematic and prone to bias[1].

[1] Gift & Stevens 1997 - GIFT, N. & STEVENS, P. F. (1997). Vagaries in the delimitation of character states in quantitative variation – an experimental study. Systematic Biology 46, 112–125.

Based on the 2015 MIT's report, which discusses the eye's ability to discern patterns, the computer today has the ability to access large datasets in a timely manner. Elaborated computer programs are helping to unravel many relationships, some of them maybe concealed and some of them may not yet be fully comprehended.

In recent years, I have also worked with ML in light of Big Data to both speed up and also scale up discovery process, leading to results considered as breakthroughs in 2015. The ML based "in silico" approach developed and elaborated helped in the rapid identification of new features with high accuracy when compared to heuristic approaches. These ML based results were validated and the sought-for items were identified, some of them have been long sought for in vain. These ML based approaches are now used as standard procedure in R&D.

This book refers to ML based approaches used in the identification of subtle variation and the identification of rare traits or genes. ML is also used to discover drugs, including new drugs that can be only effective for a small number of people with a particular gene. Recently, researchers used also ML to discover largely unknown diversity of viruses amounting to thousands of previously unknown virus, based on a research presented mid-March 2017 at a meeting organized by the US Department of Energy (DOE)[2]. ML and AI are becoming the cutting edge of discovery.

Machine learning technology is constantly evolving and the current trends promise that every sector will be data driven and will have the capacity of using machine learning in the cloud to incorporate artificial intelligence apps. ML and AI are both pushing today the entire world by performing excellently across all sectors.

The book focuses on ML to speed up discovery as well as delivery processes of products. The book contains 10 chapters, the first chapter highlights ML quests, chapter 2 provides a detailed historical perspective, chapter 3 shows how ML works by introducing conceptual frameworks of ML, chapter 4 lists some of the metrics used to assess the performance of ML types, chapter 5, 6 and 7 focus on different types of ML including supervised, unsupervised and reinforced learning. Chapter 8 and chapter 9 introduces the implementation platforms of R and Spark with different libraries including Spark MLlib. Chapter 10 provides different walk-through ML examples.

Abdallah Bari
Math Coding and Analytics
Westmount, Quebec, Canada
21 May 2018

[2] Nature NEWS 19 MARCH 2018. Machine learning spots treasure trove of elusive viruses Artificial intelligence could speed up metagenomic studies that look for species unknown to science. By Amy Maxmen

CONTENTS

Chapter 1

INTRODUCTION

Machine learning at the cutting-edge technology

Machine learning (ML), which is a subset of artificial intelligence (AI), existed before but stayed nearly dormant until Big Data has enabled its remarkable potential us as cutting-edge technology, to optimize production and discovery processes[3,4]. ML works by learning from any data to detect patterns and then makes predictions, recommendations and prescriptions, without explicit programming instructions. As it learns from data, it acquires the ability to carry out cognitive functions, such as perceiving and reasoning as well as speeding up production and discovery processes.

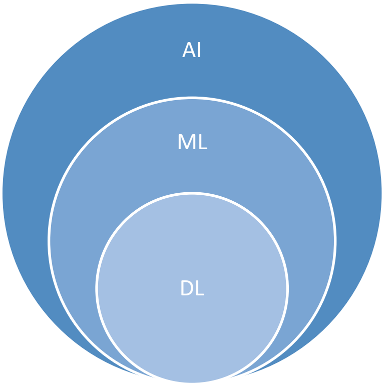

Figure 1.1 ML is a subset of AI and Deep Learning (DL) in turn is a subset of ML

[3] Bernard Marr (2018) Forbes.

[4] MIT

Most achievements in today's cutting-edge intelligent applications have been attributed to machine learning applied to Big Data[5]. Machine-learning algorithms are acquiring their ability from datasets to adapt in response to new datasets. They are evolving to have the ability to acquire functions that are normally associated with Human Intelligence (HI) such as reasoning. ML algorithms can anticipate what will happen and provide recommendations on what to do to achieve goals and improve efficacy over time.[6]

ML algorithms are speeding up processes including products delivery as well as gene or drug discovery. Product time delivery or discovery processes may take substantial amount of time and resources. Such processes may involve multi-step procedures with sometimes thousands or even hundreds of thousands of samples to be evaluated and to be tested and only a few may turn out to be effective. To shorten the time and reduce the costs of these long and costly procedures, organizations and companies are turning to ML and AI.

> "If AI or machine learning can improve that success percentage up just a few points to, say, 14 percent or 16 percent, it would be worth billions to the industry. A program that could, for example, better predict the likelihood of toxicity in the earliest stages before the company even tries to take a drug to clinical trials would directly save the company millions and importantly save it time.", wrote Jon Walker[7].

ML is helping to detect and predict non-obvious connections that could potentially lead to new treatments tailored to individuals with particular genes and thus help with personalized healthcare. The ML based predictions in healthcare are sought for to help to discover new drugs and their effectiveness, prior to their release. The Food and Drug Administration, in the USA, announced July 7, 2017 that its scaling Big Data analytics using high-performance computing (HPC) to make drug development and testing more

[5] Mckinsey & Company An Executive's Guide to AI, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai)

[6] Mckinsey & Company An Executive's Guide to AI, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai)

[7] Jon Walker (2018) Machine Learning Drug Discovery Applications – Pfizer, Roche, GSK, and More

effective[8]. Microsoft's Project Hanover aims to use ML and personalized medicine to discover cures for common and rare health issues.[9]

Deep Learning, which is a subset of ML, has helped pharmaceutical companies such as Merck in predicting useful drug candidates. The company made available to researchers a database of entries of more than 30,000 small molecules, each of which had thousands of numerical chemical-property descriptors. Researchers were able to improve Merck's predictions by 15% more[10].

Researchers have also used ML to discover recently largely unknown diversity of viruses, amounting to thousands of previously unknown viruses, based on a research presented mid-March of 2017 at a meeting organized by the US Department of Energy (DOE)[11].

The patterns identified, in recent years, using ML in combination with Big Data helped to discover new genes in crops. Some of the genes discovered have been searched for in vain in the past. ML based in-silico evaluation, as virtual screening approach, not only helped us to tap on Big Data's potential but also shortened the time to discovery, which is crucial today for business as well as for research and development to keep pace of rapid global changes including climate change[12].

[8] Avanade's survey (2017) - https://www.avanade.com/en-us/technologies/data-analytics

[9] University of New South Wales. "Quantum computers: 10-fold boost in stability achieved." ScienceDaily, 2016. /Nave, Karthryn. "Quantum computing is poised to transform our lives. Meet the

[10] Computer science: The learning machines by Nicola Jones - Nature 505 146–148 (09 January 2014).

[11] Amy Maxmen Nature (2018) Machine learning spots treasure trove of elusive viruses Artificial intelligence could speed up metagenomic studies that look for species unknown to science.

[12] Bari et al (2016). In silico evaluation of plant genetic resources to search for traits for adaptation to climate change. Climate Change 134: 667–680 - https://link.springer.com/article/10.1007/s10584-015-1541-9