

AN APPROACH TO MACHINE LEARNING

Dr. G Maria Jones

Ms. G.Subathra

Ms. Kalaivani A

Ms. D. Nancy Kirupanithi

Ms. Santhiya P



AN APPROACH TO MACHINE LEARNING

Authors:

Dr. G Maria Jones

Ms. G.Subathra

Ms. Kalaivani A

Ms. D. Nancy Kirupanithi

Ms. Santhiya P

@ All rights reserved with the publisher.

First Published: November 2022

ISBN 978-93-92090-08-0



9 789392 090080 >

ISBN: 978-93-92090-08-0

DOI: <https://doi.org/10.47716/MTS.B.978-93-92090-08-0>

Pages: 166 (Front pages 8 & Inner pages 158)

Price: 350/-

Publisher & Imprint:

Magestic Technology Solutions (P) Ltd

Chennai. Tamil Nadu, India.

www.magesticts.com

Title Verso

Title of the Book:

An Approach to Machine Learning

Author's Name:

Dr. G Maria Jones

Ms. G.Subathra

Ms. Kalaivani A

Ms. D. Nancy Kirupanithi

Ms. Santhiya P

Published By:

Magestic Technology Solutions (P) Ltd.

Publisher's Address:

544, Anna Street, Kathirvedu, Chennai 600 066

Tamil Nadu

Printer's Details:

Magestic Technology Solutions (P) Ltd.

Edition Details: First Edition

ISBN: 978-93-92090-08-0

Copyright: @ Magestic Technology Solutions (P) Ltd.

Magestic Technology Solutions (P) Ltd.
544, Anna Street, Kathirvedu, Chennai 600 066
Tamil Nadu, India.

@ 2022, Magestic Technology Solutions (P) Ltd.
Imprint Magestic Technology Solutions (P) Ltd

Printed on acid-free paper
International Standard Book Number (ISBN): 978-93-92090-08-0
(Paperback)
Digital Object Identifier (DOI): 10.47716/MTS.B.978-93-92090-08-0.

This book provides information obtained from reliable and authoritative sources. The author and publisher have made reasonable attempts to publish accurate facts and information, but they cannot be held accountable for any content's accuracy or usage. The writers and publishers have endeavoured to track down the copyright holders of every content copied in this book and regret if permission to publish in this format was not acquired. Please notify us through email if any copyright-protected work has not been recognised so that we may make the necessary corrections in future reprints. No portion of this book may be reprinted, reproduced, transmitted, or used in any form by any electronic, mechanical, or other means, now known or hereafter developed, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without the publisher's written permission.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

Visit the Magestic Technology Solutions (P) Ltd.
Web site at <http://www.magesticts.com>

Preface

The process of automatically recognising significant patterns within large amounts of data is referred to as "machine learning." Over the course of the last couple of decades, it has evolved into a tool that is used in almost every activity that requires the extraction of information from large data sets. We are surrounded by technology that is based on machine learning: Search engines are learning how to bring us the best results (while placing profitable ads), antispam software is learning how to filter our email messages, and credit card transactions are secured by software that learns how to detect frauds. Intelligent personal assistance software on smart phones are able to learn to recognise voice commands and digital cameras are able to train themselves to identify faces. Accident-prevention systems in vehicles are constructed with the help of machine learning algorithms. These systems are installed in modern automobiles. In addition, machine learning is extensively utilised in a variety of scientific applications including bioinformatics, medicine, and astronomy.

One aspect that is shared by all of these applications is the fact that, in contrast to more conventional applications of computers, in these situations, due to the complexity of the patterns that need to be detected, a human programmer is unable to provide an explicit, fine-detailed specification of how such tasks should be carried out. This is one of the characteristics that makes all of these applications unique. Taking cues from other intelligent beings, the majority of our capabilities have been obtained or improved via the process of learning from our experiences (rather than following explicit instructions given to us). Tools for machine learning are used to give computer programmes the capacity to "learn" and modify their behaviour on their own.

The first objective of this book is to provide the fundamental ideas that underpin machine learning in a manner that is comprehensive while still being simple to understand.

- Authors

Dr. G Maria Jones

Ms. G.Subathra

Ms. Kalaivani A

Ms. D. Nancy Kirupanithi

Ms. Santhiya P

This Page Intentionally Left Blank

Table of Contents

CHAPTER-I

Overview of Machine Learning.....1

CHAPTER-II

Machine Learning Categories.....13

CHAPTER-IIi

Machine Learning ToolBox.....23

CHAPTER-IV

Data Scrubbing.....39

CHAPTER-V

Setting up Our Data.....53

CHAPTER-VI

Regression Analysis.....61

CHAPTER-VII

Clustering.....79

CHAPTER-VIII

BIAS and Variance.....93

CHAPTER-IX

Artificial Neural Networks.....103

CHAPTER-X

Decision Trees.....117

CHAPTER-XI

Ensemble Modeling.....129

CHAPTER-XII

Building a Model in Python.....135

This Page Intentionally Left Blank

Chapter - 1
Overview of Machine Learning

This page Intentionally Left Blank

Chapter - 1

Overview of Machine Learning

IBM released a study with a vague and befuddling title in their IBM Research Development Journal in 1959. The purpose of the research paper, which IBM's Arthur Samuel authored, was "to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the programme." The study was carried out using machine learning techniques in the game of checkers.

It was not the first publication to use the term "machine learning" per se; however, Arthur Samuel is widely thought to be the first individual to coin and define machine learning in the form we know today. This was the case even though it was not the first publication to use the term "machine learning" per se. This was the case even though the publication did not use the term "machine learning" per se. The landmark article "Some Studies in Machine Learning Using the Game of Checkers" that Samuel submitted to the journal is early evidence of homo sapiens' determination to impart our technique of learning to machines that man has created.



Figure 1: Historical mentions of "machine learning" in published books.

Source: Google Ngram Viewer, 2017

Computers may be taught new skills via a field of computer science called machine learning. In his article, Arthur Samuel introduces

machine learning as a branch of IT. This definition has been accepted as valid for almost sixty years now.

Even though it is not explicitly addressed in Arthur Samuel's definition of machine learning, the idea of self-learning is a fundamental component of this field. This refers to using statistical modelling to identify trends and enhance performance based on data and empirical information; all of this is accomplished without direct programming instructions. As Arthur Samuel said, this is what he meant when he spoke about the ability to learn without being explicitly programmed. He does not show that robots can form opinions without being taught how to do so. However, machine learning relies heavily on predetermined computer code. Samuel realised that computers might do a task without being given explicit instructions but rather by being fed data.

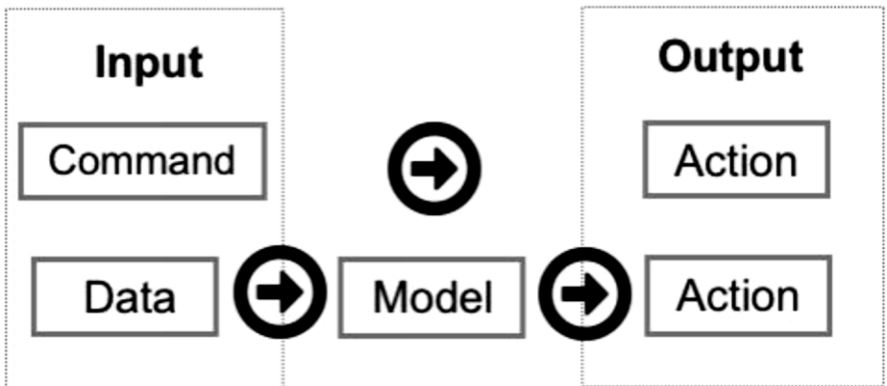


Figure 2: Comparing Input Command vs Input Data

For instance, in the computer language Python, entering "2+2" and then pressing the "Enter" button is an example of an input command.

```
>>> 2+2
4
>>>
```

This is a direct order, and we will respond with a straightforward question.

However, the data that is inputted is unique. After the information is entered into the machine, the algorithm will be selected, the hyperparameters (settings) will be adjusted, and the device will be given the command to do the analysis. Using a process of trial and error, the computer continues the process of deciphering patterns that have been detected in the data. The data model of the machine, which was developed by evaluating the practices of the data, may subsequently be used to forecast future values.

In conventional computer programming, there is a layer of separation between the programmer and the machine, even though there is some connection between the two. This is because the machine forms choices centred on experience and emulates the decision-making method dependent on human input.

Take, for instance, the scenario in which our computer, having analysed the YouTube viewing habits of data scientists, concludes that there is a significant correlation between this profession and movies featuring cats. Later on, our computer will determine whether or not there are any correlations between the physical characteristics of baseball players and the possibility that they will win the Most Valuable Player award (MVP) for the season. In the first possible outcome, the machine analyses to determine which YouTube video data scientists love watching the most based on the level of user interaction shown by the number of likes, subscribers, and subsequent views. In the second scenario, the computer evaluated various aspects of former Major League Baseball Most Valuable Players, including their age, education level, and other physical characteristics.

On the other hand, none of these possibilities included our equipment deliberately designed to generate an inevitable result. We were responsible for giving the machine the appropriate input data, and we were also responsible for establishing the algorithms; nonetheless, the machine itself arrived at the final prediction via the processes of self-

learning and data modelling. Constructing a data model may be compared to teaching a guide dog new command.

Guide dogs go through rigorous training in order to learn how to behave appropriately in a variety of settings. For instance, the dog will learn how to heel at a red light and correctly direct its master around obstacles. Additionally, the dog will learn how to retrieve dropped items. Another ability that the dog will learn is how to collect objects that have been dropped. If the dog has been well trained, the guide dog will ultimately be able to function without the assistance of its trainer and be able to put its training to use in various settings where it will not be watched. Similarly, machine learning models may be taught to develop conclusions based on prior experiences and can be used to make judgments.

One clear example would be creating a model that could recognise unsolicited commercial emails (spam). The model has been programmed to ignore emails that have dubious subject lines and text in the body that includes three or more of the following keywords: "dear friend," "free," "invoice," "PayPal," "Viagra," "casino," "payment," and "winner." However, we are not yet engaged in machine learning at this juncture in time. Let us go back to the graphic depiction of the input command against the input data. We can see that this procedure consists of two steps: Command > action.

The machine learning process consists of three steps: Data > Model > Action.

Therefore, to include machine learning in our spam detection system, we need to change the term "command" to "data." We will also need to add the phrase "model" to produce an action. Both of these changes are necessary in order for us to generate action (output). In this specific scenario, the model is composed of rules derived from statistics, and the data consists of emails. The terms on our first negative list have been included in the model's parameters. After that, training and validation of the model against the data follow.

After the data have been entered into the model, there is a reasonable likelihood that the model's assumptions will result in some incorrect forecasts because the model contains the assumptions. For instance, according to the principles of this model, the subject line of an email that reads "PayPal has accepted our money for Casino Royale bought on eBay" would be automatically labelled as spam.

The spam detection system is fooled into giving a false positive based on the negative list of keywords included in the model because this email is legitimate and was received through an auto-responder associated with PayPal. This causes the spam detection system to give a higher probability of a false positive. Traditional programming is very susceptible to the situations mentioned above since no technique is built to check the assumptions' validity and modify the model's governing principles. Traditional programming is especially sensitive to such cases because it lacks flexibility. On the other hand, machine learning can respond to mistakes and adjust its assumptions via its three-step procedure. This allows it to adapt to new situations.

Training & Test Data

When dealing with data for machine learning, it is common to practise separating the information into two distinct categories: training data and test data. The training data are derived from the first split, representing the initial reserve of data used in developing our model. In the case of the identification of spam emails, examples of false positives compared to the auto-response provided by PayPal could be found in the training data. The system will then need the addition of new rules or adjustments. For instance, email alerts sent from the address "payments@paypal.com" should be exempt from being filtered as spam.

When we were satisfied with the model's accuracy, we built using the training data. When we have successfully established a model using those data, we can test the model using the remaining data, referred to as the test data. When we have completed this step, we will know the

necessary to determine whether or not the model is correct. When we have reached a point where we are content with the outcomes of both the training data and the test data, the machine learning model will be ready to filter incoming emails and create judgments on how to classify the messages received. This will allow us to organise the emails received more effectively.

The difference between traditional programming and machine learning can seem to be a trivial one at first. However, it will become clear as we go through further examples and watch the extraordinary potential of self-learning in ever more complex situations.

The second key takeaway from this chapter is the context in which machine learning exists within the more significant data science and computer science fields.

This requires understanding how machine learning links to its sister disciplines and parent domains. This is crucial because we will encounter these related phrases while looking for suitable study resources, and we will hear them discussed to the point of exhaustion in classes that introduce

"Machine learning" and "data mining" are two examples of relevant fields that, at first appearance, are not always easy to differentiate from one another.

Let us get started with a high-level overview, shall we? Computer science studies anything having to do with the construction and use of computers. It is the foundation of various essential topics, including computer programming, data mining, machine learning, and almost every other study area (except traditional statistics). The following expansive discipline is data science, which may be found inside the all-encompassing realm of computer science. Data science is a computer science subfield that focuses on extracting information and insights from data via various methodologies and systems.

**Computer
Science**



**Data
Science**



AI



**Machine
Learning**



Figure 3: The progression of machine learning, shown here as a series of nesting matryoshka dolls from Russia

The field of artificial intelligence is the third Matryoshka doll to emerge from the intersection of computer science and data science. The term "artificial intelligence" (AI) refers to the capacity of robots to carry out activities that require intellect and cognitive abilities. In the same way that the Industrial Revolution ushered in an era of machines that could do tasks previously performed by humans, artificial intelligence (AI) is driving the development of computers capable of simulating cognitive abilities.

AI comprises a plethora of well-liked subfields. Although it is somewhat broad overall, it is far more specialised than computer science and data science. These subfields include perception, natural language processing (NLP), machine learning, search and planning, reasoning and knowledge representation, and search and planning. Because of the widespread usage of self-learning algorithms, machine learning has begun to seep into other artificial intelligence subfields, such as natural language processing and perception.

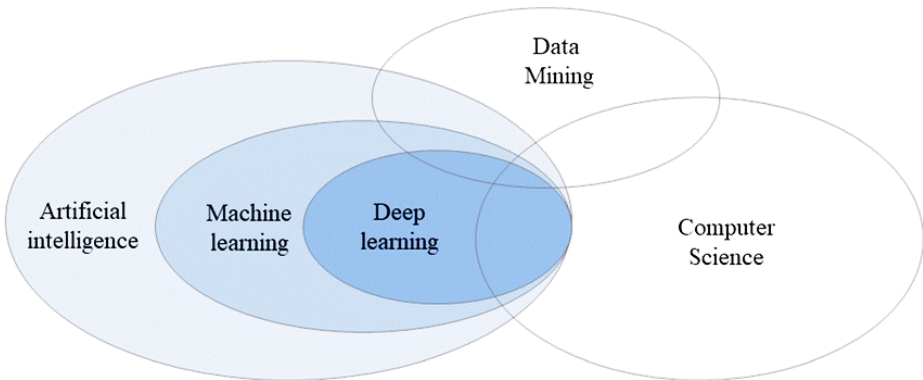


Figure 4: A pictorial representation of the relationship between the many data-related areas

Machine learning is an ideal starting point for students interested in artificial intelligence (AI), since it gives a more focused and practical lens through which to study, in contrast to the conceptual ambiguity associated with AI. The algorithms developed for machine learning may also be used in other areas of study, such as natural language processing and perception. It is possible to get a certain amount of expertise in machine learning with just a Master's degree. However, a Doctor of Philosophy degree is necessary to make significant advancements in artificial intelligence.

As was previously said, there is some overlap between machine learning and data mining. This related field focuses on uncovering and excavating patterns in massive datasets. To do data analysis, data mining and machine learning use well-known algorithms. K-means clustering, association analysis, and regression analysis are a few methods that fall under this category. On the other hand, data mining narrows down on cleaning up massive datasets to extract valuable information from the past. At the same time, machine learning concentrates on the progressive process of self-learning and data modelling to make predictions.

One way to prove the distinction between data mining and machine learning is by using an example using two different groups of

archaeologists. The first group comprises archaeologists, and their primary objective is to clear away any material that may be blocking their view of essential artefacts so that they may better examine the site. Their primary objectives are to dig the region, make new findings of great value, pack up their equipment and move on to the following location. They leave the previous day's excavation site and travel the next day to another foreign location to start a new project without any connection to the previous day's work site.

The second group of archaeologists is likewise excavating historical sites; however, these archaeologists use a distinct methodological approach. They stop digging the central pit on purpose for a few weeks while they rethink the situation. During this period, they go to many other significant archaeological sites in the region and investigate the excavation methods used at each site. Upon arriving back at the location of their project, they put this newfound information to use by excavating many smaller trenches around the giant hole.

After that, the archaeologists analyse the findings. They first dig one pit, think about what they learned from that experience, and then adjust how they excavate the next pit. This involves creating innovative ways to decrease error and enhance the precision of their job, as well as forecasting the amount of time it takes to dig a pit, recognising the variety and patterns found in the local topography, and understanding how much time it takes to excavate a pit. As a result of this experience, they can now optimise their approach to building a strategic model to dig the central pit.

The first team is committed to data mining if it is not apparent. In contrast, the second team is more interested in machine learning. Both data mining and machine learning seem to be identical on a microscopic level. They do make use of many of the same techniques. Both of these groups earn their income by excavating historical places to find objects of value. However, their technique differs in their application in the real world. The team working on machine learning

places a high priority on separating their dataset into training and test data to construct a model. They will increase their ability to make predictions based on previous experiences if they use this knowledge. Meanwhile, the data mining team is focusing their efforts on digging the target region in the most efficient manner possible—without using a self-learning model—to move on to the next cleaning operation.

Chapter - 2
Machine Learning Categories

This page Intentionally Left Blank

Chapter – 2

ML CATEGORIES

Any individual who works in this subject is constantly faced with difficulty selecting the appropriate algorithm or combination of algorithms for the task at hand. Machine learning is comprised of several hundred different statistically-based algorithms. However, before we get into the specifics of algorithms, it is necessary to understand the three main subfields that comprise machine learning. These three types include supervised learning, unsupervised learning and reinforcement.

Supervised Learning

Supervised learning is the initial branch of machine learning, and its primary focus is on learning patterns by establishing a connection between variables and known outcomes, as well as through working with datasets that have been labelled.

In supervised learning, one feeds the machine sample data that contains a variety of characteristics (expressed as "X") as well as the proper value output of the data (represented as "y"). This allows the computer to learn. The fact that both the output and the feature values are known makes the dataset a candidate for the labelling process. The algorithm then decodes patterns in the data. It builds a model capable of recreating the same basic rules using new data.

For instance, a supervised algorithm can formulate predictions by analysing the relationship between car attributes (such as the year of make, car brand, and mileage) and the selling price of other cars sold based on historical data in order to estimate the going rate for the purchase of a used car on the market. This can be done to forecast the market rate for purchasing a used car. Given that the supervised algorithm is aware of the total price of previous cards that have been sold, it can analyse the link between the automobile's attributes and its worth by working backwards.

Bibliography

- Agrawal, T. (2021).** *Hyperparameter optimization in machine learning: make your machine learning and deep learning models more efficient*. New York, NY, USA: Apress.
- Alpaydin, E. (2016).** *Machine learning: the new AI*. MIT press.
- Alpaydin, E. (2021).** *Machine learning*. MIT Press.
- Athey, S. (2018).** The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press.
- Bengio, Y., Lodi, A., & Prouvost, A. (2021).** Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2), 405-421.
- Bonaccorso, G. (2017).** *Machine learning algorithms*. Packt Publishing Ltd.
- Brunton, S. L., & Kutz, J. N. (2022).** *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.
- Burkov, A. (2019).** *The hundred-page machine learning book* (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.
- Burkov, A. (2020).** *Machine learning engineering* (Vol. 1). True Positive Incorporated.
- El Naqa, I., & Murphy, M. J. (2015).** What is machine learning? In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham.
- Géron, A. (2022).** *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."
- Harrington, P. (2012).** *Machine learning in action*. Simon and Schuster.
- Langlely, P. (1996).** *Elements of machine learning*. Morgan Kaufmann.
- Mitchell, T. M. (2006). *The discipline of machine learning* (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.
- Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021).** A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement*, 167, 108288.
- Mahesh, B. (2020).** Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], 9, 381-386.

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021).** A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.
- Mitchell, T. M. (2006).** *The discipline of machine learning* (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.
- Mueller, J. P., & Massaron, L. (2021).** *Machine learning for dummies*. John Wiley & Sons.
- Murphy, K. P. (2022).** *Probabilistic machine learning: an introduction*. MIT press.
- Provost, F., & Kohavi, R. (1998).** On applied research in machine learning. *MACHINE LEARNING-BOSTON-*, 30, 127-132.
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022).** Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 1-85.
- Sammut, C., & Webb, G. I. (Eds.). (2011).** *Encyclopedia of machine learning*. Springer Science & Business Media.
- Schölkopf, B. (2022).** Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (pp. 765-804).
- Shalev-Shwartz, S., & Ben-David, S. (2014).** *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- Shavlik, J. W., Dietterich, T., & Dietterich, T. G. (Eds.). (1990).** *Readings in machine learning*. Morgan Kaufmann.
- Sra, S., Nowozin, S., & Wright, S. J. (Eds.). (2012).** *Optimization for machine learning*. Mit Press.
- Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., ... & Müller, K. R. (2021).** Machine learning force fields. *Chemical Reviews*, 121(16), 10142-10186.
- Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2016).** Machine learning basics. *Deep Learn*, 98-164.
- Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2022).** Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3), 169-181.
- Zhang, Y. (Ed.). (2010).** *New advances in machine learning*. BoD—Books on Demand.