1. The Craft of Machine Learning

Learning from Data provides an introduction to machine learning using R. It assumes no prior experience with either machine learning or the R language. The book is designed for upper level undergraduate or postgraduate students in computer science as well as other disciplines, since machine learning has found application in a wide variety of fields from the social sciences, biological sciences, economics, and more. The handbook will also be a good resource for professionals wanting to get started with machine learning. Prior courses in linear algebra, probability and statistics are assumed; however, the book attempts to fill in as much detail as needed for the subject at hand. This is a handbook[1], not a textbook. A few practice problems are provided with the assumption that more problems will be provided by your instructor. In these days where solutions to end of chapter problems are found online, for a price, there is not much educational value to be had in end-of-chapter problems. Readers can find on the web many sample R notebooks and learn from those, as well. One particularly good resource is `https://www.kaggle.com`.

The aim of this handbook is to provide an introduction to the craft of machine learning through conceptual explanations of the algorithms and small examples of running the algorithms in R. Sample notebooks are provided on the author's github at: `https://github.com/kjmazidi/Learning_from_Data`.

Figure 1.1 shows fields related to machine learning. The arrows are pointing outward to denote dependency. Without these fields we would not have machine learning. Statistics and probability form the mathematical foundations of many of the algorithms we will learn in this book. AI and computer science pushed the frontiers of what computers could do which made machine learning possible.

1.1 Machine Learning

We will use *machine learning* as an umbrella term for many closely related terms. Some statisticians call machine learning *statistical learning*. The field of *data science* and the task of *data mining* often involve machine learning techniques, coupled with more data exploration and analysis.

[1] A reference work providing guidance for a technical application or art.

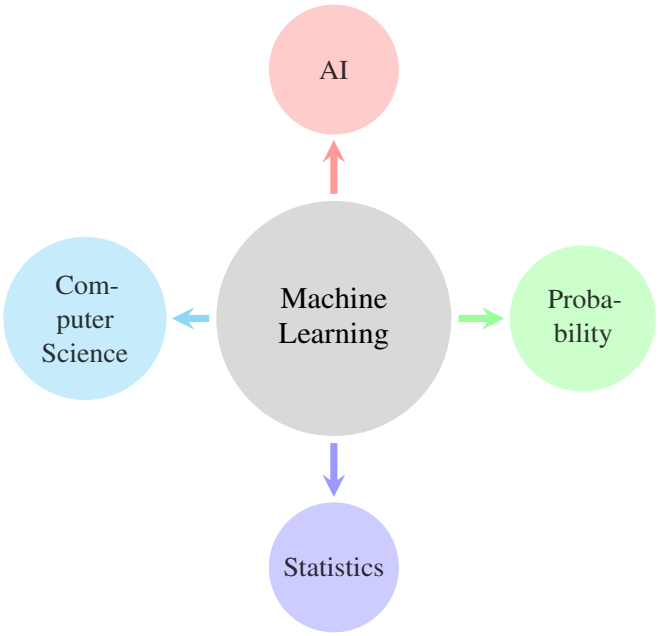

Figure 1.1: Origins of Machine Learning

Machine learning has received varied definitions as the field developed. This book proposes the following definition:

Definition 1.1.1 — Machine Learning. Machine learning trains computers to accurately recognize patterns in data for purposes of data analysis, prediction, and/or action selection by autonomous agents.

The key words in this definition are: data, patterns, predictions, and actions along with the caveat: accurate. Let's examine these in detail.

1.1.1 Data

Nothing can be learned without data. The data and what we wish to learn from the data go hand in hand. For many learning scenarios, data takes the form of a table of values where each row represents one data example, and columns represent attributes or features of the examples. One learning scenario, clustering, seeks to group like instances. Another learning scenario, supervised learning, seeks to learn about one feature based on combinations of the other features. Data can take other forms besides tables, such as a set of actions weighted by features representing the current environment.

The data we will use in this handbook is small and neat compared to data you will encounter in real-world machine learning problems. This is by design, so that focus can be placed on the algorithms themselves which is the main objective of the book. Just be aware that in real-world scenarios, more of your time will be spent in data gathering and data cleaning that in the actual machine learning.

When gathering data or using data from other sources, ethical considerations must always guide our actions. Who owns the data? Who are the subjects of the data? Has the data been anonymized? If not, did the subjects give consent for the use of the data? How will the analysis of this data be used? How might it impact the subjects in the data as well as the larger community?

1.1.2 Patterns

The best *general* pattern recognition machine is the human mind but computers can actually beat human performance on narrowly defined tasks. The ability to recognize patterns in data enables algorithms to learn things like whether someone is a good credit risk, whether two people might be compatible, whether the object outlined in sensors is a human or a dark spot on the pavement.

When beginning a real-world machine learning project, how do you know what to look for? From raw data, organized data must be built. Once the data is organized, decisions must be made about what could be learned and what is important to learn from the data. These decisions often need to be in concert with domain experts and/or the owners and users of the data who wish to learn from it.

1.1.3 Predictions from Data

Learning patterns in data enables us to predict outcomes on future data – data the algorithm has never before seen. Predictions may simply involve finding like instances in the data, or predicting a target value which may be a number or membership in a group.

In the examples in this book, generally we train algorithms on a portion of the data and use the remaining data to test and evaluate how well the trained model can perform on previously unseen data. This is a common situation in machine learning, sometimes called batch learning because the data is fed into the algorithm in one batch. There are other approaches, however, such as online learning in which the algorithm is continually learning from newly available data and being evaluated in real time. Online learning techniques can also be used when the available data is too large to be stored in memory. An alternate approach to handle big data is to do parallel distributed machine learning, often in the cloud with specialized software.

1.1.4 Accuracy

Predictions must be accurate or they are not predictions but random guesses. Machine learning makes use of many measurement techniques to gauge accuracy and evaluate performance of the algorithms. Many of these metrics are used to evaluate the training model itself and others will be used to evaluate performance on a held-out test set.

1.1.5 Actions

Every day more autonomous agents enter our lives, from smart thermostats, to automated assistants, to self-driving vehicles. These agents take actions based on what they have learned, and most continue to learn over time, usually by uploading data to a central learning repository. Some actions taken by autonomous agents will be controversial in the coming decades as ethical and legal issues evolve in response to humans co-existing with autonomous agents.

1.2 Machine Learning Scenarios

There are scores of machine learning algorithms with countless variations each. This book describes the most common algorithms, while providing a foundation for students to learn more algorithms on their own. There are many ways to classify machine learning algorithms, and not all algorithms fit neatly into our categories but these categories serve as a helpful overview of the field.

1.2.1 Informative v. Active

Most machine learning algorithms covered in this book are **informative**. They provide information to us in the form of data analysis or prediction. These informative algorithms input data observations and output a model of the data that can then be used to predict outcomes for new data fed into the

model. In contrast, the field of Reinforcement Learning teaches **active** agents to identify optimal actions given the current environment and what has been learned in past experience. The input to these algorithms for initial training comes in the form of data but some agents may continue to learn with sensors that let them learn from the environment.

1.2.2 Supervised v. unsupervised learning

Informative algorithms are of two main types. The term **supervised learning** refers to scenarios where each data instance has a label. This label is used to train the algorithm so that labels can be predicted for future data items. The term **unsupervised learning** refers to scenarios where data does not have labels and the goal is simply to learn more about the data.

1.2.3 Regression v. classification

Supervised learning algorithms fall into two major groups. Regression and classification occur in supervised learning scenarios with labeled data. In **regression** our target is a real numbered value, like trying to predict the market value of a home given its square footage and other data. In **classification** our target is a class, like predicting if a borrower is a good credit risk or not, given their income, outstanding credit balance, and other predictors.

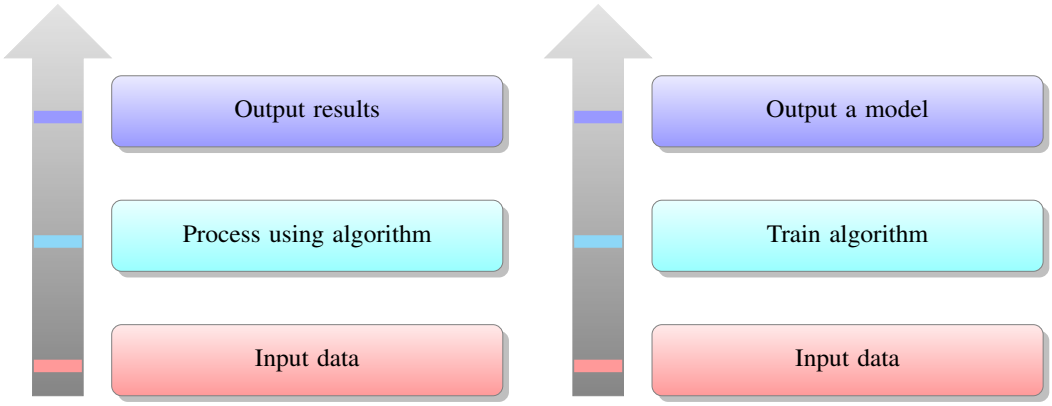

Figure 1.2: Traditional Programming (Left) v. Machine Learning (Right)

1.3 Machine learning v. traditional algorithms

Machine learning algorithms are different from traditional algorithms encountered in computer programming. In Figure 1.2 we see the traditional computer programming paradigm on the left: data is fed into code that processes it and outputs the results of the processing. In Figure 1.2 we see the machine learning paradigm on the right: we feed data into an algorithm which builds and outputs a model of the data. In traditional programming, all knowledge is explicitly encoded in the algorithm. In machine learning, knowledge is inferred from data.

Why do we need machine learning? Can't we just explicitly code algorithms for problems? There are two typical situations in which traditional programming cannot be used to solve problems. The first type is when it is not possible to encode all the rules needed to solve a problem. How would you encode rules for recognizing faces in photos? We don't even know the rules we use in our minds to recognize faces so it would not be possible to encode rules. However, we can train computers to recognize key edges and regions of photos that are likely to be faces. The second type of situation in which traditional programming cannot be used to solve a problem is when the scale

of the problem is too large. If a company has huge amounts of customer data it would take millions of human hours to find useful patterns in the data. Machine learning algorithms can find patterns quickly in large amounts of data.

As we go through the material in this handbook you will learn several machine learning algorithms. These algorithms typically have statistical and probabilistic foundations which we will explore. However, beyond the theory and technique, machine learning is also a craft as well as a science. Each major Part of the book devotes a chapter pointing out some innovations, ideas, and techniques from this evolving craft.

1.4 Terminology

Machine learning grew out of statistics, computer science, as well as other fields. For this reason there are often multiple terms for the same thing. Let's start with names for data. The table below contains a sample data set (with headings).

GPA	Hours	SAT	Class
3.2	15	1450	Junior
3.8	21	1420	Sophomore
2.5	9	1367	Freshman

Table 1.1: Student GPA, Average Hours Studied/Week, SAT, Class

We have only 3 rows of data. Each row is a sample data point, also called an **example**, **instance**, or an **observation**. Each column in the table is an **attribute**, also called a **feature** or a **predictor**. We have 4 features: GPA, average number of hours studied per week, SAT score, and classification. The first 3 are **quantitative**, or numeric, features while Class is a **qualitative** feature because it can only take on one of a finite set of values. Qualitative features are also called **factors** or **categorical data**.

If we want to learn GPA as a function of the other 3 features, we say that GPA is our **target**, or **response**, variable while the other 3 are **features** or **predictors**.

1.5 Notation

This book uses the following notation conventions for data:

- x_i subscript i indexes observations in a data set; i ranges from 1 to N, the number of observations (rows) in the data set
- $x_{i,j}$ subscript j indexes predictors in a data set; j ranges from 1 to P, the number of predictors (columns) in the data set. Here we are referencing predictor j from observation i.
- In matrix notation, lower case letters like x represent scalars, bold face lower case letters like $\mathbf{x}$ represent vectors, and upper case bold face letters like $\mathbf{X}$ represent matrices.