Part I

Preliminaries

1

Introduction

We are overwhelmed with data, but information is hard to extract. Humans have learned to navigate this complexity quite efficiently. Mathematics, statistics, and intelligent machines are still far less capable.

1.1 Modeling

Humans constantly use models to interpret reality and take decisions. We use models when we try to understand what is happening and when we try to predict what will happen next. We need them to manage, and perhaps reduce, the unpredictability of the world. We need them to survive.

Without models, reality will be an overwhelming amount of data with no information. Indeed, in my view, models are the instruments that transform data into information.

Humans construct models in various ways, consciously or unconsciously, rationally or instinctively. There is an increasing need to advance knowledge and understanding of the mechanisms and tools for reliable data-driven models. Nowadays, scientists and engineers are developing automated modeling tools. The purpose is to give machines the instruments to select and learn models, developing their ability to interact with the real world and make autonomous decisions. This is broadly called artificial intelligence (AI).

There are many different kinds of models; the ones we use, for instance, to choose our food at the market are different from the ones we use to place our satellites into the right orbit. As an example, let's think about the model we use when we cross a busy road. Let's make an effort to analyze what we normally, spontaneously, do in this situation. The process is surprisingly complex and sophisticated. Normally, we start looking if there is any vehicle and, if there isn't, then we cross. In the case there are vehicles, then we gauge their distance and, if they are far away enough, we cross. If the vehicles are not far away, then we estimate their speed and predict how long they will take to arrive, pondering if there is a sufficient amount of time for us to cross the road also accounting for the uncertainty of our prediction. There are many other variables that we evaluate. We distinguish cars from trucks and trucks from buses; they have different typical speeds and behaviors. We might consider the vehicle's trajectory and, in some cases, look at the driver and guess his or her intent to let us pass. We also consider, and normally dismiss, other variables that are in most cases marginal,

for instance, the color of the vehicle, its brand, or the age of the driver. This operation is an example of data-driven modeling. We have learned these operations from observations and, perhaps, from imitation. We do not have a general theory; we do all this spontaneously and instinctively without noticing the amount of data that we are processing and the way we make predictions; we ponder, compare, and, use them. We consider this to be *simple* but it is actually a very *complex* operation that we, modern humans, can perform quite easily but that is extremely hard to encode in machines or formulate precisely with our current mathematical and computational tools.

1.2 Models as Maps between Data Representations

Models can be deterministic, perhaps including some degree of uncertainty due to limited and noisy observations, or they can be probabilistic. Newton's gravitation law and the consequent modeling of the motion of celestial bodies is an example of deterministic modeling that can be written in the form $\boldsymbol{y} = g(\boldsymbol{x})$ and for which the true model can be *learned* with arbitrary precision given a large enough number of observations.[1] In contrast, the Maxwell–Boltzmann distribution, describing the speed of particles in a gas, is an example of probabilistic modeling where the law to be learned is the probability that a particle has a velocity smaller or equal to a given value v: $Probability(V \leq v)$.[2] This is a different kind of problem than the previous; however, also in this case, one can formulate the problem in the same form $\boldsymbol{y} = g(\boldsymbol{x})$ where, this relation represents a form of *probabilistic* dependence between the variables and one must add an uncertainty term, ϵ, to the relation.[3] Some problems cannot be represented in terms of an input and an output, and the modeling task becomes the discovery of mutual dependence relations between the variables. The goal becomes to map the structure of the interactions between the system's variables and uncover their similarities, their hierarchies, and their causal relations. In other contexts, one might aim instead to discover emerging behaviors from simple rules, and for this task simulations are often adopted. What I have just described are different forms of modeling, and they all serve the same purpose of helping us to navigate reality by describing what is happening and making a prediction about what will happen. In other terms, models help us to interpret existing observations and make predictions about future ones.

Models are tools to transform data into useful information that can be used for decisions and actions. Data, and observations, can be images, movies, journal articles, chats, financial prices, or electric signals, or – indeed – anything that is produced by some process and carries some information. Data can always be represented as points and shapes in a space. The space might not be ordinary; it is often high-dimensional and sometimes non-Euclidean. In some cases, such as in deep learning, one might use data representations across several spaces.

[1] Here, I use the notation $\boldsymbol{x}$ and $\boldsymbol{y}$ to generically indicate two sets of input and output variables, and $g(\cdot)$ a function mapping between the two sets.
[2] Where V is the random variable indicating the velocity of a particle and v is a value.
[3] Deterministic models could be seen as probabilistic models with zero uncertainty.

At this level of abstraction, models are maps between points in these spaces. These maps must both guide us as precisely as possible to the positions of the existing observations and to help us to find the most likely positions of future observations. The model provides a *description* of the system by locating the observations and their interrelation in these spaces; the model can also provide *predictions* by inferring the presence and location of unobserved points.

Such maps between different data representation spaces are functions that models learn to approximate. Scientists have developed efficient methods to approximate them with so-called universal approximators. There are several universal approximators that can be used for modeling, from polynomials to trees (Tikk et al., 2003). Among the most versatile and presently popular are deep neural networks. Whether deep learning architectures are the best-suited tools for modeling has still to be proved. However, they are certainly surprisingly good and efficient for many modeling tasks.

There are other kinds of modeling that might not be directly associated with function approximations. In some contexts, a meaningful modeling approach consists of starting from the elementary components of the system and modeling their behaviors and interactions from first principles. This microscopic modeling can produce precise explanations of the system behavior and can help us to understand the origin of macroscopic phenomena from fundamental microscopic laws. For instance, this approach is sometimes used to derive the properties of materials from the constitutive atoms in so-called *ab initio* simulations. However, in many real and complex systems, there is nothing comparable to the atoms. Indeed, in these systems, the elementary components are complex themselves and their laws of interaction are often unknown. In complex systems, microscopic modeling approaches often end up either being unrealistic oversimplifications or being too complex to be of use to explain the underlying system.

1.3 Models for Real and Complex Systems

Scientists are challenged to handle and solve increasingly complex systems, from markets to self-driving cars. Although data-driven modeling ultimately aims to automate the process of learning new models, human intuition and creativity is still central to this domain. This is why, to successfully build data-driven models, we must learn both the science and the art of modeling, the mathematical rigor and the intuition.

Complex systems are not abstract objects. They are very real, they are everywhere. Humans are complex systems and so are human societies. Animals – even the simplest – are complex systems and many human-created artificial systems – such as financial markets – are complex themselves. The defining characteristic of complexity is that simplification is not possible without losing crucial properties of the system. In complex systems, important properties *emerge* from the combination of many different elements. Although their elements might be known, the emergent property is the result of their combination, and prediction is hard to achieve from the analysis of the constituting elements in isolation (see Parisi

(2002); Boccara (2010)). An example is a living organism, even a simple one, which is made of parts with properties and functions that might be well known and understood. However, even the deepest knowledge of all its parts, will not reveal the most important property of the organism: the fact that it is alive. Being alive is the emergent property of the whole system, which is of greater importance than the sum of the parts. Understanding emergent properties is very important in complex systems modeling, it is – literally – a question of life and death, and it is what makes such modeling very challenging.

Henry Louis Mencken – known as the Sage of Baltimore – once noted that *"for every complex problem there is an answer that is clear, simple, and wrong"* (Perry, 2022). Nonetheless, despite complex systems being challenging, one can devise effective models to describe and predict, at some level of accuracy, their behavior. We do it all the time. Indeed, we are complex creatures who live in a complex world. The modeling of complex systems has the same nature and scope as the modeling of any other system. In complex systems, models are used to describe the system and to predict its behavior. However, the task can be more arduous. For instance, often in complex systems not only is the system non-deterministic but also the internal rules change and adapt. Nonetheless, the modeling of these systems is still based on the scientific method's circular approach (see Section 19.1), and the general principles remain the same and models can be formulated and tested using the general tools that are presented in this book.

Let me mention that one must distinguish between systems described by a theory and others described by a model trained from examples. A theory, like Kepler's laws governing planetary motion, involves equations and principles that accurately describe a natural phenomenon. These theories provide both understanding and predictability. However, for most real, complex systems, a clear theoretical framework might be lacking and sometimes impossible. In such cases, purely data-driven models might fit data points and might predict trends but do not provide any coherent theoretical foundation and their applicability outside the context where the data have been harvested is often impossible. In contrast, established scientific theories, like Kepler's laws, can lead to deeper comprehension of the phenomenon and can be applied to different systems from the ones originally used to formulate the theory.

1.4 Models as Multivariate Probabilities

Models are used to interpret and understand reality. Some models help us to distinguish between different scenarios such as distinguishing between friends and foes or isolating food from poison. These kinds of models can be defined as recognition and "classification" models. In this setting, one has an input dataset and an output information. For instance, one could have a picture of an animal as input and information about the kind of animal provided by the model as output.

Similarly, one could have a set of observations of a physical system, for instance, the positions of the planets in the heavens over a period of time, and the model

might output the mathematical law for the motion of the planets. These problems are often referred to as "regression" and they concern the discovery of relations between two sets of variables one called "dependent" (the output) and the other called "independent" (the input). In this framework, modeling consists of solving a so-called *inverse problem*: given the observations find the law that generates them.

Models are also used to infer the internal relations between a set of variables. In this case, one has no inputs or outputs and all variables are interdependent. The problem consists of uncovering the structures of similarities, hierarchies, dependencies, and causalities that are in the data structure.

Models must be capable of generalizing and predicting outcomes that have not been observed yet. We use them constantly in our everyday life – to survive – for instance, to avoid being crushed by a bus when we cross the road. The accuracy of the model prediction in various circumstances is a measure of the goodness of the model.

Classification, regression, data-structure investigation, and prediction tasks are not necessarily distinct and can be seen as different ways of addressing a problem and interpreting the model. Indeed, the distinction between input and output variables is just a useful convention and the laws that map inputs into outputs coincide with the dependency structure of the dataset. Further, prediction consists of inferring the dependence between variables at different times or across different settings.

In the real and artificial systems that are of interest, the general problem concerns a set of several variables and their relations. In all these cases, the knowledge of the multivariate probability of all the system's variables provides the full information about the system and it is, therefore, the instrument to model it. I indeed argue and demonstrate in this book that the vast majority of what we call modeling can be formulated in terms of the probability functions of the variables characterizing the system or, in other words, in terms of *multivariate probabilities* (see Chapter 6).

2

Fundamentals of Probability

2.1 Probability

Probability is a quantification of the extent to which something is likely to happen. Data-driven probabilistic modeling is about estimating probability from observations. The purpose of this chapter is to provide a common referential, self-contained background on fundamental concepts and definitions concerning probabilities. In this chapter, I present notation, basic definitions, and perspectives that are fundamental to this domain and essential for the rest of the book. Knowledgeable readers could skip this chapter and use it only as a reference for later chapters. Readers interested in further insights on this part can refer to textbooks on probability such as, for instance, Feller (1957), Ross (2014), Durrett (2019), and Grimmett and Stirzaker (2020).

In this book, I consider only real-valued random variables, both continuous and discrete, and they are denoted by X. The random variable X is, in general, the outcome of some process and phenomena that results in aleatoric output values. Following the literature, I indicate with upper case X the random variable itself and with lower case x its value for a given observation. These values can be considered as drawn from a given population, which is the collection of all possible observations and is associated with a probability distribution. Such a probability distribution of the population makes the draw of some observations more likely, others less likely, and some impossible.

Probabilities are real numbers between zero and one. A value of the probability equal to zero means that the event has no chance to occur. Conversly, a value of the probability equal to one represents certainty that the event will occur. The sum of the probabilities of all possible outcomes must be equal to one. Indeed, at least one of those outcomes is certain to occur. Probabilities cannot be negative.

2.2 Theory of Probability

The fundamental concepts associated with probability are quite intuitive; however, the mathematical foundation of probability theory is not elementary. Some definitions and fundamental concepts are necessary and they are provided in Appendix A. In this section, I report some intuitive concepts referring to the material in Appendix A for formal foundations. The interested reader can further delve

into this fascinating subject by, for instance, starting from the book by Jaynes (2003).

Definition 2.1 (Probability) Probability is formalized in terms of **probability space**, a triplet $(\Omega, \mathcal{F}, P)$ where Ω is the **sample space** which is the set of all possible outcomes that are the events in the **event space** $\mathcal{F}$; P is the **probability measure**, a function associating each event with a number between 0 and 1.

Definition 2.2 (Random variable) A **random variable** X is a measurable function (see Definition A.5) that maps from the sample space Ω to the real numbers $\mathbb{R}$:

$$X : \Omega \to \mathbb{R}. \tag{2.1}$$

When the image[1] of X is countable, then X is called a **discrete random variable**. When the image is an uncountable interval (possibly infinite) then X is called a **continuous random variable**.

The random variable X is defined on the σ-algebra (see Definition A.2) of the events, while its image is the probability defined on Borel's σ-algebra (see Definition A.6) of the real numbers $\mathbb{R}$.

In simple words, the probability of a random variable is a function that maps each possible value of the random variable to a number between zero and one, which is indeed the probability. This map captures the likelihood of different outcomes occurring when the random variable is observed or measured.

Example 2.1 (Probability and random variables) Let me provide a very simple example: the rolling of a fair standard six-sided dice. In this case, the **sample space** consists of all possible outcomes that can occur when rolling the dice, which, for this example is

$$\Omega = (1, 2, 3, 4, 5, 6). \tag{2.2}$$

The **event space** is a collection of subsets of the sample space. Each subset represents an event; that is, a set of outcomes. Specifically, for a single fair dice roll, one can consider as event space the set of all subsets of the sample space Ω:

$$\mathcal{F} = ((\emptyset), (1), (2), (3), (4), (5), (6), (1, 2), (1, 3)..., (1, 2, 3), ..., \Omega). \tag{2.3}$$

This is the largest event space, which is called the **power set**.

The **random variable** associated with the fair dice roll is a function

[1] The image of a function is the set of all possible output values that the function can produce when its entire domain is considered.

that maps each outcome in the sample space to a numerical value which, in this case, is simply the number on the dice's upper face. Therefore, the random variable X takes on the values 1, 2, 3, 4, 5, or 6.

The **probability measure** assigns a probability value to each event. For this fair dice roll problem, the probability measure assigns the probability $1/6$ to the draw of each value of the dice face, the probability $1/6^2$ for each combination of two values, etc.

2.2.1 Cumulative Distribution Function

Definition 2.3 (Cumulative distribution function) The **cumulative distribution function**[2] (CDF) is the probability that the random variable $X \in \mathbb{R}$ will take a value smaller than or equal to a value x. I denote it as

$$F(x) = P(X \leq x). \tag{2.4}$$

Appendix A for details and notation.

The CDF, $F(x)$, is nondecreasing and it starts from a value equal to zero at the extreme left of the support where no smaller observations are possible and it ends at the other extreme, on the right of the support, where it is equal to one and no larger observations are possible. In general, $F(-\infty) = 0$ and $F(+\infty) = 1$.

2.2.2 Complementary Cumulative Distribution Function

Definition 2.4 (Complementary cumulative distribution function) The **complementary cumulative distribution function** (CCDF) $P(X > x)$ is the probability that the random variable $X \in \mathbb{R}$ will take a value larger than x. One has

$$P(X > x) = 1 - F(x). \tag{2.5}$$

Example 2.2 (Cumulative Distribution Function) For instance, the likelihood that in a classroom someone picked at random is shorter than 160 cm can be expressed in terms of the CDF $F(x) = P(X \leq x)$ with $X =$ Height and $x = 160$ cm. It is clear that in this example $F(0) = 0$ (nobody can be shorter than 0 cm) and $F(500) = 1$ (no human has ever been reported to have been taller than five meters). Conversely, the probability that someone picked at random is taller than 160 cm is the CCDF $P(X > x) = 1 - F(x)$.

[2] Sometimes called cumulative probability distribution function.