Handbook
Statistical foundations of machine learning
Second edition

Gianluca Bontempi

Machine Learning Group
Computer Science Department
ULB, Université Libre de Bruxelles, Belgium
mlg.ulb.ac.be

July 15, 2021

And indeed all things that are known have number. For it is not possible that anything whatsoever be understood or known without this.

Philolaus, 400 BC

Not everything that can be counted counts, and not everything that counts can be counted.

W. B. Cameron, 1963

Preface to the 2021 edition

The book is dedicated to all students interested in machine learning who are not content with only running lines of (deep-learning) code but who are eager to learn about this discipline's assumptions, limitations, and perspectives. When I was a student, my dream was to become an AI researcher and save humankind with intelligent robots. For several reasons, I abandoned such ambitions (but you never know). In exchange, I discovered that machine learning is much more than a conventional research domain since it is intimately associated with the scientific process transforming observations into knowledge.

The first version of this book was made publicly available in 2004 with two objectives and one ambition. The first objective was to provide a handbook to ULB students since I was (and still am) strongly convinced that a decent course should come with a decent handbook. The second objective was to group together all the material that I consider fundamental (or at least essential) for a Ph.D. student to undertake a thesis in my lab. At that time, there were already plenty of excellent machine learning reference books. However, most of the existing work did not sufficiently acknowledge what machine learning owes to statistics and concealed (or did not make explicit enough, notably because of incomplete or implicit notation) important assumptions underlying the process of inferring models from data.

The ambition was to make a free academic reference on the foundations of machine learning available on the web. There are several reasons for providing free access to this work: I am a civil servant in an institution that already takes care of my salary; most of the material is not original (though its organisation, notation definition, exercises, code and structure represent the primary added value of the author); in many parts of the world access to expensive textbooks or reference material is still difficult for the majority of students; most of the knowledge underlying this book was obtained by the author thanks to free (or at least non charged) references and, last but not least, education seems to be the last societal domain where a communist approach may be as effective as rewarding. Personally, I would be delighted if this book could be used to facilitate the access of underfunded educational and research communities to state-of-the-art scientific notions.

Though machine learning was already a hot topic at the end of the 20th century, nowadays, it is definitely surrounded by a lot of hype and excitement. The number of publications describing or using a machine learning approach in the last decades is countless, making it impossible to address the heterogeneity of the domain in a single book. Therefore, it is interesting to check how much material from the first edition is still useful: reassuringly enough, the more the nature of the content is fundamental, the less it is prone to obsolescence. Nevertheless, a lot of new things (not only deep learning) happened in the domain, and, more specifically, I realised the importance of some fundamental concepts that were neglected in the first edition.

In particular, during those years, I realised the importance of exposing young researchers to notions of multivariate dependency and independence. These notions are brilliantly summarised in the topic of graphical models whose knowledge is es-

sential to grasp aspects of dimensionality reduction and feature selection. Secondly, I (re)discovered that the foundations of machine learning lie in epistemology, the branch of philosophy aiming to explain the meaning of knowledge and the process of discovering it. Third, I became convinced that a process of discovering knowledge from data should not be limited to modelling associations but aimed at discovering causal mechanisms. Finally, I added a number of exercises, R scripts, and Shiny dashboards to visualise and illustrate (sometimes too abstract) probabilistic and estimation notions. In this sense, I am convinced that the adoption of Monte Carlo simulation to introduce probabilistic concepts should be a more common habit in introductory statistics classes.

For sure, I am strongly indebted to a lot of authors and their publications. I hope I acknowledged them adequately in the bibliography. If I did not give enough credit to some of the existing works, please do not hesitate to contact me. Last but not least, the book is dedicated to all my ULB students and MLG researchers in whom I have tried for many years to inculcate complex concepts of statistical learning. Their eyes staring at my hand-waving, while I was trying to elucidate some abstruse notions, were the best indicators of how to adapt, select and improve the book's content.

To all those who want to send a note or continue to follow my machine learning journey, see you on my blog `https://datascience741.wordpress.com`.

Acknowledgements

Though the book is not peer-reviewed, the added value of writing a handbook for students and researchers is that they are typically very careful readers and willing to pinpoint mistakes, inconsistencies, bad English and (a lot of) typos. First, I would like to thank (in random order) the MLG researchers who sent me very useful comments: Abhilash Miranda, Yann-aël Le Borgne, Souhaib Ben Taieb, Jacopo De Stefani, Patrick Meyer, Olivier Caelen, Liran Lerman. Thanks as well to the following students and readers (in random order) for their comments and remarks: Robin de Haes, Mourad Akandouch, Zheng Liangliang, Olga Ibanez Solé, Maud Destree, Wolf De Wulf, Dieter Vandesande, Miro-Manuel Matagne, Henry Morgan, Pascal Tribel. A big thank to all of you! And do not hesitate to drop me an email if you have comments or remarks!

Contents

Chapter 1

Introduction

Over the last decades, a growing number of organisations have been allocating a vast amount of resources to construct and maintain databases and data warehouses. In scientific endeavours, data refers to carefully collected observations about some phenomenon under study. In business, data capture information about economic trends, critical markets, competitors, and customers. In manufacturing, data record machinery performances and production rates in different conditions. There are essentially two reasons why people gather increasing volumes of data. First, they think some valuable assets are implicitly coded within them, and, second, computer technology enables effective data storage and processing at reduced costs.

The idea of extracting useful knowledge from volumes of data is common to many disciplines, from statistics to physics, from econometrics to system identification and adaptive control. The procedure for finding useful patterns in data is known by different names in different communities, viz., knowledge extraction, pattern analysis, data processing. In the artificial intelligence community, the most common name is *machine learning* [70]. More recently, the set of computational techniques and tools to support the modelling of a large amount of data is grouped under the more general label of *data science*.

The need for programs that can learn was stressed by Alan Turing, who argued that it might be too ambitious to write from scratch programs for tasks that even humans must learn to perform. This handbook aims to present the statistical foundations of machine learning intended as the discipline which deals with the automatic design of models from data. In particular, we focus on *supervised learning* problems (Figure 1.1), where the goal is to model the relation between a set of *input* variables and one or more *output* variables, which are considered to be dependent on the inputs in some manner.

Since the handbook deals with artificial learning methods, we do not take into consideration any argument of biological or cognitive plausibility of the learning methods we present. Learning is postulated here as a problem of statistical estimation of the dependencies between variables on the basis of empirical data.

The relevance of statistical analysis arises as soon as there is a need to extract useful information from data records obtained by repeatedly measuring an observed phenomenon. Suppose we are interested in learning about the relationship[1] between two observed variables x (e.g. the height of a child) and y (e.g. the weight of a child), which are quantitative observations of some phenomenon of interest (e.g. obesity during childhood). Sometimes, the *a priori* knowledge that describes the relation between x and y is available. In other cases, no satisfactory theory exists, and all that we can use are repeated measurements of x and y.

[1]Note that the term relation simply denotes the statistical association (due to a probabilistic dependency) between the two variables and has no causal connotation.

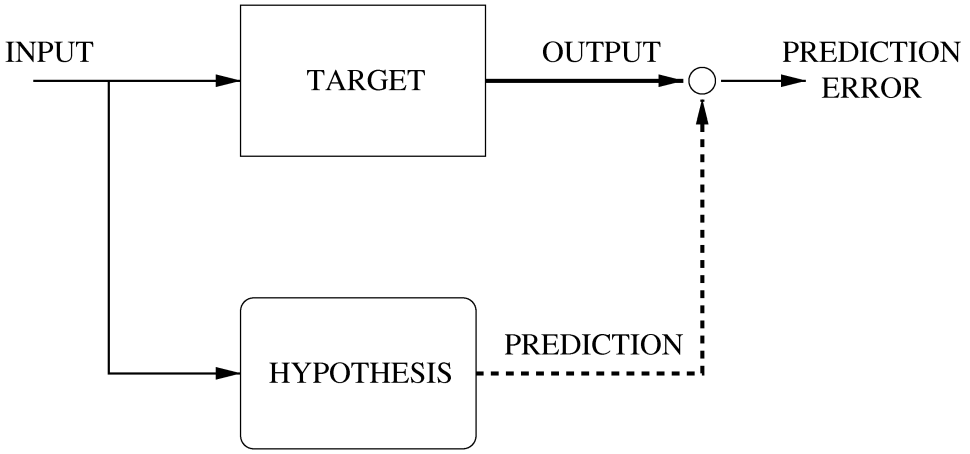


Figure 1.1: The supervised learning setting. Machine learning aims to infer from observed data the best model of the stochastic input/output dependency.

In this book, our focus is the second situation where we assume that only a set of observed data is available. The reasons for addressing this problem are essentially two. First, the more complex is the input/output relation, the less effective will be the contribution of a human expert in extracting a model of the relation. Second, data-driven modelling may be a valuable support for the designer also in modelling tasks where she can take advantage of existing knowledge.

Though machine learning is becoming a central component in many (so-called) intelligent applications, we deem that simply considering it as a powerful computational technology would be utterly reductive. The process of extracting knowledge from observations lies at the root of the modern scientific process, and the most challenging issues in machine learning relate to well-established philosophical and epistemological problems, notably induction or the notion of truth. This is the reason why we added in this new version of the handbook a preliminary chapter to situate the machine learning problem into the broader context of human knowledge acquisition.

Modelling from data

Modelling from data is often viewed as an art, mixing an expert's insight with the information contained in the observations. A typical modelling process cannot be considered as a sequential process but is better represented as a loop with many feedback paths and interactions with the model designer. Various steps are repeated several times aiming to reach, through continuous refinements, a good description of the phenomenon underlying the data.

The modelling process consists of a *preliminary* phase that brings the data from their original form to a structured configuration and a *learning* phase that aims to select the *model*, or *hypothesis*, that best approximates the data (Figure 1.2).

The preliminary phase can be decomposed in the following steps:

Problem formulation. Here the model designer chooses a particular application domain, a phenomenon to be studied, a number of descriptive variables and hypothesises the existence of a (stochastic) relation (or dependency) between the measurable variables. The definition of the input variables (and where necessary their transformations) is a very crucial step and is called *feature*

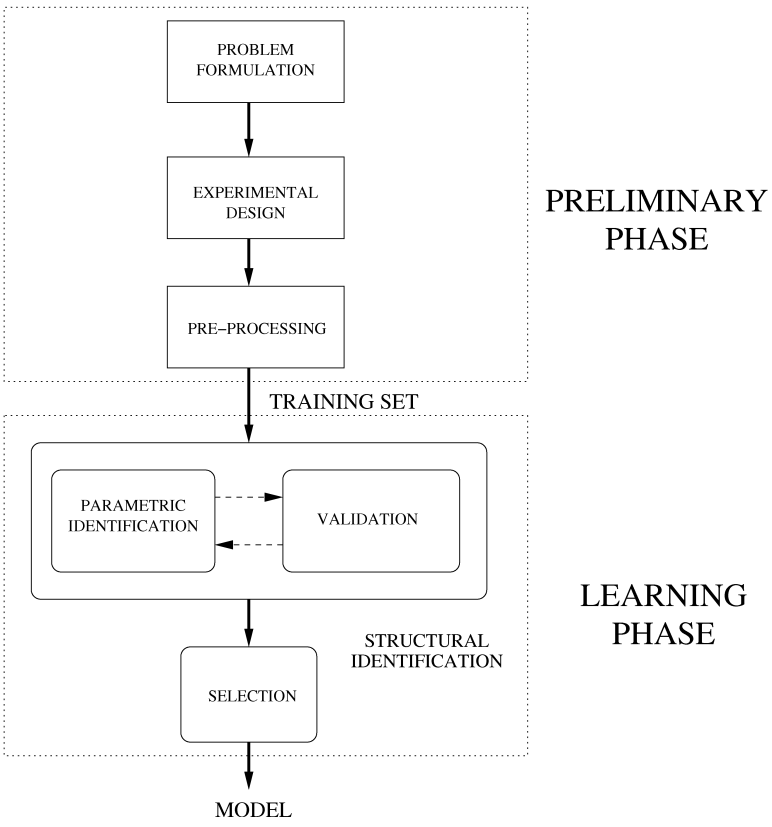


Figure 1.2: The modelling process and its decomposition in the preliminary phase and learning phase.

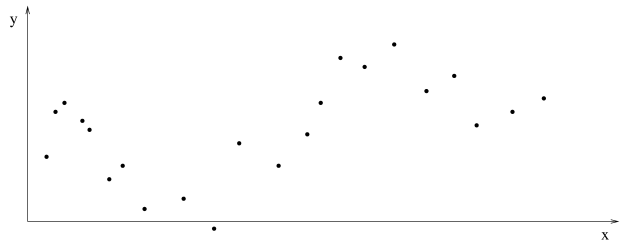


Figure 1.3: A training set for a simple supervised learning problem with one input variable x and one output variable y. The dots represent the observed samples.

engineering. It is important to stress here the proactive role played by the human (in contrast to a *tabula rasa* approach), and that this role is a necessary condition for any knowledge process.

Experimental design. This step aims to return a dataset which, ideally, should be made of observations that are well-representative of the phenomenon in order to maximise the performance of the modelling process [54].

Pre-processing. In this step, raw data are cleaned to make learning easier. Pre-processing includes a large set of actions on the observed data, such as noise filtering, outlier removal, missing data treatment [122], feature selection, and so on.

Once the preliminary phase has returned the dataset in a structured input/output form (e.g. a two-column table), called *training set*, the learning phase begins. A graphical representation of a training set for a simple learning problem with one input variable x and one output variable y is given in Figure 1.3. This manuscript will mostly focus on this second phase assuming that the preliminary steps have already been performed by the model designer.

Suppose that, on the basis of the collected data, we wish to learn the unknown dependency existing between the x variable and the y variable. The knowledge of this dependency could shed light on the observed phenomenon and let us predict the value of the output y for a given input (e.g. what is the expected weight of a child who is 120cm tall?). What is difficult and tricky in this task is the finiteness and the random nature of data. For instance, a second set of observations of the same pair of variables could produce a dataset (Figure 1.4) that is not identical to the one in Figure 1.3 though both originate from the same measurable phenomenon. This simple fact suggests that a simple interpolation of the observed data would not produce an accurate model of the data.

The goal of machine learning is to formalise and optimise the procedure which brings from data to model and consequently from data to predictions. A learning procedure can be concisely defined as a search, in a space of possible model configurations, of the model which best represents the phenomenon underlying the data. As a consequence, a learning procedure requires both a *search space* where possible solutions may be found and an *assessment criterion* that measures the quality of the solutions in order to select the best one.

The search space is defined by the designer using a set of nested classes with increasing capacity (or representation power). For our introductory purposes, it is

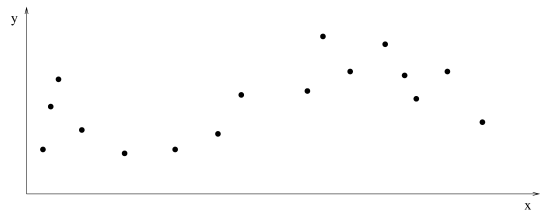


Figure 1.4: A second realisation of the training set for the same phenomenon observed in Figure 1.3. The dots represent the observed examples.

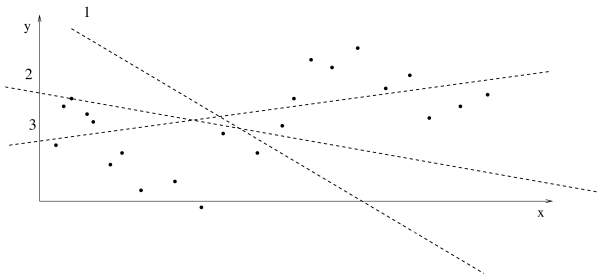


Figure 1.5: Training set and three parametric models which belong to the class of first order polynomials.

sufficient to consider here a *class* as a set of input/output models (e.g. the set of polynomial models) with the same *model structure* (e.g. second-order degree) and the *capacity* of the class as a measure of the set of input/output mappings which can be approximated by the models belonging to the class.

Figure 1.5 shows the training set of Figure 1.3 together with three parametric models which belong to the class of first-order polynomials. Figure 1.6 shows the same training set with three parametric models, which belong to the class of second-order polynomials.

The reader could visually decide whether the class of second-order models is more suitable or not than the first-order class to model the dataset. At the same time, she could guess which among the three plotted models is the one that produces the best fitting.

In real high-dimensional settings, however, a visual assessment of the quality of a model is neither possible nor sufficient. Data-driven quantitative criteria are therefore required. We will assume that the goal of learning is to achieve a good *statistical generalisation*. This means that the learned model is expected to return an accurate prediction of the dependent (output) variable for new (unseen) values of the independent (input) variables. By new values we intend values which are not part of the training set but are generated by the same stochastic process.

Once the classes of models and the assessment criteria are fixed, the goal of a learning algorithm is to search i) for the best class of models and ii) for the best parametric model within such a class. Any supervised learning algorithm is then made of two nested loops denoted as the *structural* identification loop and the *parametric* identification loop.

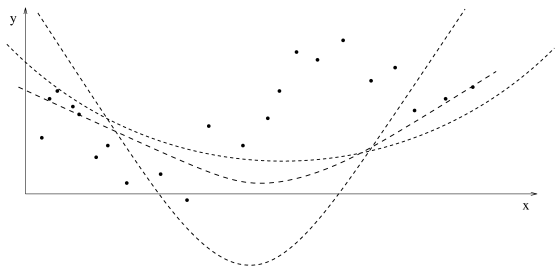


Figure 1.6: Training set and three parametric models which belong to the class of second-order polynomials.

Structural identification is the outer loop that seeks the model structure which is expected to have the best accuracy. It is composed of a *validation* phase, which assesses each model structure on the basis of the chosen assessment criterion, and a *selection* phase which returns the best model structure on the basis of the validation output. Parametric identification is the inner loop that returns the best model for a fixed model structure. We will show that the two procedures are intertwined since the structural identification requires the outcome of the parametric step in order to assess the goodness of a class.

Statistical machine learning

On the basis of the previous section, we could argue that learning is nothing more than a standard problem of optimisation. Unfortunately, the reality is far more complex. In fact, because of the finite amount of data and their random nature, there exists a strong correlation between parametric and structural identification steps, which makes non-trivial the problem of assessing and, finally, choosing the prediction model. In fact, the random nature of the data demands a definition of the problem in stochastic terms and the adoption of statistical procedures to choose and assess the quality of a prediction model. In this context, a challenging issue is how to determine the class of models more appropriate to our problem. Since the results of a learning procedure are found to be sensitive to the class of models chosen to fit the data, statisticians and machine learning researchers have proposed over the years a number of machine learning algorithms. Well-known examples are linear models, neural networks, local modelling techniques, support vector machines, and regression trees. The aim of such learning algorithms, many of which are presented in this book, is to combine high generalisation with an effective learning procedure.

However, the ambition of this handbook is to present machine learning as a scientific domain that goes beyond the mere collection of computational procedures. Since machine learning is deeply rooted in conventional statistics, any introduction to this topic must include some introductory chapters to the foundations of probability, statistics and estimation theory. At the same time, we intend to show that machine learning widens the scope of conventional statistics by focusing on a number of topics often overlooked by statistical literature, like nonlinearity, large dimensionality, adaptivity, optimisation and analysis of massive datasets.

It is important to remark, also, that the recent adoption of machine learning models is showing the limitation of pure black-box approaches, targeting accuracy at the cost of interpretability. This is made evident by the embedding of automatic approaches in decision-making processes with impact on ethical, social, political, or

juridical aspects. While we are personally skeptical about gaining any interpretability from a large number of parameters and hyperparameters underlying a supervised learner, we are confident that human insight can be obtained by techniques able to reduce or modularise large variate tasks. In this direction, feature selection and causal inference techniques are promising approaches to master the complexity of data-driven modelling and return human accessible descriptions (e.g. in the form of mechanisms).

This manuscript aims to find a good balance between theory and practice by situating most of the theoretical notions in a real context with the help of practical examples and real datasets. All the examples are implemented in the statistical programming language R [158] made available by the companion package `gbcode` (Appendix F). In this second edition, we provide as well a number of Shiny dashboards (Appendix G) to give the reader a more tangible idea of somewhat abstract concepts. For an introduction to R we refer the reader to [52, 185]. This practical connotation is particularly important since machine learning techniques are nowadays more and more embedded in plenty of technological domains, like bioinformatics, robotics, intelligent control, speech and image recognition, multimedia, web and data mining, computational finance, business intelligence.

Outline

The outline of the book is as follows. Chapter 2 is one of the novelties of the second edition. Its aim is to situate the process of modelling from data in a larger epistemological domain dealing with the problem of extracting knowledge from observations. We deem it interesting to show how some of the formal problems addressed in the book dates back to old philosophical disputes and works. Chapter 3 summarises the relevant background material in probability. Chapter 4 has been added to introduce graphical modelling, a flexible and interpretable way of representing large variate problems in probabilistic terms. In particular, this formalism puts into evidence the importance of conditional independence as a key notion to illustrate the properties of dependencies and simplify the modelling of large dimensional tasks. Chapter 5 introduces the parametric approach to parametric estimation and hypothesis testing. Chapter 6 presents some nonparametric alternatives to the parametric techniques discussed in Chapter 5. Chapter 7 introduces supervised learning as the statistical problem of assessing and selecting a hypothesis function on the basis of input/output observations. Chapter 8 reviews the steps which lead from raw observations to a final model. This is a methodological chapter that introduces some algorithmic procedures underlying most of the machine learning techniques. Chapter 9 presents conventional linear approaches to regression and classification. Chapter 10 introduces some machine learning techniques which deal with nonlinear regression and classification tasks. Chapter 11 presents the model averaging approach, a recent and powerful way for obtaining improved generalisation accuracy by combining several learning machines. Chapter 12 deals with the problem of dimensionality reduction and in particular with feature selection strategies. Chapter 13 has been added in the 2nd edition to make clear the limitations of associational approaches and to stress the risk of wrong extrapolation and biases if pure statistical results are interpreted in a causal manner. We believe that causal reasoning represents the ultimate step in the data analytics process going from data to knowledge.

Although the book focuses on supervised learning, some related notions of unsupervised learning and density estimation are presented in Appendix A.

1.1 Notations

Throughout this manuscript, boldface denotes random variables and normal font is used for instances (realisations) of random variables. Strictly speaking, one should always distinguish in notation between a random variable and its realisation. However, we will adopt this extra notational burden only when the meaning is not clear from the context. Then we will use $\text{Prob}\{z\}$ (or $(p(z))$ as a shorthand for $\text{Prob}\{\mathbf{z} = z\}$ $((p_{\mathbf{z}}(z))$ when the identity of the random variable is clear from the context.

As far as variables are concerned, lowercase letters denote scalars or vectors of observables, greek letters denote parameter vectors, and uppercase denotes matrices. Uppercase in italics denotes generic sets while uppercase in greek letters denotes sets of parameters.

Gender-neutral pronoun: computer sciences suffer from the gender issue and probably much more than other sciences. Of course, you won't find any solution in this book but the author (a man) felt odd in referring to a generic reader by using a masculine pronoun only. He then decided to use as much as possible a "(s)he" notation or, alternatively, a (balanced) random gender choice.

Generic notation

- θ: Parameter vector.
- $\boldsymbol{\theta}$: Random parameter vector.
- M: Matrix.
- $[N \times n]$ or $[N, n]$: Dimensionality of a matrix with N rows and n columns.
- M^T: Transpose of the matrix M.
- $\text{diag}[m_1, \ldots, m_N]$: Diagonal matrix with diagonal $[m_1, \ldots, m_N]$
- $\mathbf{M}$: Random matrix.
- $\hat{\theta}$: Estimate of θ.
- $\hat{\boldsymbol{\theta}}$: Estimator of θ.
- τ: Index in an iterative algorithm.

Probability Theory notation

- Ω : Set of possible outcomes.
- ω: Outcome (or elementary event).
- $\{\mathcal{E}\}$: Set of possible events.
- $\mathcal{E}$: Event.
- $\text{Prob}\{\mathcal{E}\}$: Probability of the event $\mathcal{E}$.
- $(\Omega, \{\mathcal{E}\}, \text{Prob}\{\cdot\})$: Probabilistic model of an experiment.
- $\mathcal{Z}$: Domain of the random variable $\mathbf{z}$.

- $P(z)$: Probability distribution of a discrete random variable $\mathbf{z}$. Also $P_{\mathbf{z}}(z)$.
- $F(z) = \text{Prob}\{\mathbf{z} \leq z\}$: Distribution function of a continuous random variable $\mathbf{z}$. Also $F_{\mathbf{z}}(z)$.
- $p(z)$: Probability density of a continuous r.v.. Also $p_{\mathbf{z}}(z)$.
- $E[\mathbf{z}]$: Expected value of the random variable $\mathbf{z}$.
- $E_{\mathbf{x}}[\mathbf{z}] = \int_{\mathcal{X}} z(x,y)p(x)dx$: Expected value of the random variable $\mathbf{z}$ averaged over $\mathbf{x}$.
- $\text{Var}[\mathbf{z}]$: Variance of the random variable $\mathbf{z}$.
- $L_N(\theta)$: Likelihood of a parameter θ given the dataset D_N.
- $l_N(\theta)$: Log-Likelihood of a parameter θ given the dataset D_N.
- $\mathcal{U}(a,b)$: univariate uniform probability density between a and $b \geq a$.
- $\mathcal{N}(\mu,\sigma^2)$: univariate Normal probability density with mean μ and variance σ^2 (Section 3.4.2).
- $\mathbf{z} \sim p_{\mathbf{z}}(z)$: random variable $\mathbf{z}$ with probability density $p_{\mathbf{z}}(z)$.
- $\mathbf{z} \sim \mathcal{N}(\mu,\sigma^2)$: random variable $\mathbf{z}$ with Normal density with mean μ and variance σ^2.

Learning Theory notation

- $\mathbf{x}$: Multidimensional random input variable.
- $\mathbf{x}_j$: jth component of the multidimensional input variable.
- $\mathcal{X} \subset \mathbb{R}^n$: Input space.
- $\mathbf{y}$: Multidimensional output variable.
- $\mathcal{Y} \subset \mathbb{R}$: Output space.
- x_i: ith observation of the random vector $\mathbf{x}$.
- x_{ij}: ith observation of the jth component of the random vector $\mathbf{x}$.
- $f(x)$: Target regression function.
- $\mathbf{w}$: Random noise variable.
- $z_i = \langle x_i, y_i \rangle$: Input-output example (also observation or data point): i^{th} case in training set.
- N: Number of observed examples in the training set.
- $D_N = \{z_1, z_2, \ldots, z_N\}$: Training set.
- Λ: Class of hypothesis.
- α: Hypothesis parameter vector.
- $h(x,\alpha)$: Hypothesis function.

- Λ_s: Hypothesis class of capacity (or complexity) s.
- $L(y, f(x, \alpha))$: Loss function.
- $R(\alpha)$: Functional risk.
- α_0: $\arg\min_{\alpha \in \Lambda} R(\alpha)$.
- $R_{\text{emp}}(\alpha)$: Empirical functional risk.

- α_N: Parameter which minimises the empirical risk of D_N

- G_N: Mean integrated squared error (MISE).
- l: Number of folds in cross-validation.
- $\hat{G}_{\text{cv}}$: Cross-validation estimate of G_N.
- $\hat{G}_{\text{loo}}$: Leave-one-out estimate of G_N.
- N_{tr}: Number of examples used for training in cross-validation.
- N_{ts}: Number of examples used for test in cross-validation.
- $D_{(i)}$: Training set with the ith example set aside.
- $\alpha_{N(i)}$: Parameter which minimises the empirical risk of $D_{(i)}$.
- $\hat{G}_{\text{bs}}$: Bootstrap estimate of G_N.
- $D_{(b)}$: Bootstrap training set of size N generated by D_N with replacement.
- $\alpha_{(b)}$: Parameter which minimises the empirical risk of the bootstrap set $D_{(b)}$.
- B: Number of bootstrap examples.

Data analysis notation

- x_i: ith row of matrix X.
- $x_{\cdot j}$: jth column of matrix X.
- x_{ij}: jth element of vector x_i.
- X_{ij}: ijth element of matrix X.
- q: Query point (point in the input space where a prediction is required).
- $\hat{y}_q$: Prediction in the query point.
- $\hat{y}_i^{-j}$: Leave-one-out prediction in x_i with the j^{th} s example set aside.
- $e_j^{\text{loo}} = y_j - \hat{y}_j^{-j}$: Leave-one-out error with the j^{th} example set aside.
- $K(\cdot)$: Kernel function.
- B: Bandwidth.

- β: Linear coefficients vector.
- $\hat{\beta}$: Least-squares parameters vector.
- $\hat{\beta}^{-j}$: Least-squares parameters vector with the j^{th} example set aside.
- $h_j(x, \alpha)$: j^{th}, $j = 1, \ldots, m$, local model in a modular architecture.
- ρ_j: Activation or basis function.
- η_j: Set of parameters of the activation function.

Chapter 2

Setting the foundations: machine learning and epistemology

Machine learning is a relatively new discipline, but its foundations rest on much older notions like modelling, reasoning, information, truth, knowledge, uncertainty, induction. Nowadays, much of those notions have a mathematical and/or computational interpretation, also thanks to machine learning. Nevertheless, before reaching a mathematical formalisation, they have been the object of an extensive philosophical inquiry and discussion. The aim of this chapter (primarily inspired by the book [80]) is to provide a rapid historical journey over the most important contributions of philosophy to epistemology, the branch of philosophy of science that investigates how humans extract and attain knowledge in the scientific process.

The two main phases of human reasoning are the acquisition of true knowledge and its manipulation in a truth-preserving manner. Induction is concerned with the first part, while deductive logic addresses the second one. In ancient times, logic was the only aspect of knowledge that deserved the attention of philosophers and epistemologists. A possible reason was that, until the scientific revolution, it was a common belief that either truth was inaccessible (e.g. the allegory of Plato's cave) or could be attained only by an initiatory process of inspiration, made possible by the benevolence of God.

2.1 Deductive logic

The most ancient discipline formalising the notions of truth, reasoning, and knowledge is logic, whose origin dates back to Aristotle. Logic is concerned with defining the properties that reasoning mechanisms should have in order to transform consistently true statements into other true statements. The objects of reasoning are *arguments*, i.e. groups of propositions where a proposition is a *statement* that can be either true or false. According to [104] an argument (or inference) is made of two groups of statements, one of which (*premises*) is claimed to provide support for the other (*conclusions*). For instance

If A, then C
A.

C $\quad \therefore$

is an argument where the groups of premises is made of the two propositions ("If A, then C" and "A") and the conclusion is the proposition "C". Premises are the statements that define the evidence while the conclusion is the statement that the evidence is supposed to imply. An argument consisting of exactly two premises and one conclusion, like the one above, is called a *syllogism*. If one of the two premises is in the conditional form (as the example above), it is called a hypothetical syllogism.

Logic cannot, in general, tell whether premises are true or false (factual claim). It is instead concerned with the quality of the reasoning process, which links premises to conclusions (inferential claim). Its purpose is to develop methods and techniques that allow us to distinguish good arguments (where the premises do support the conclusion) from bad ones. In particular, logic distinguishes between validity and sound arguments. An argument is *valid* if

- it is logically impossible for the conclusion to be false when the premises are true,
- conclusion is a logical consequence of (it follows from) the premises,
- it is truth-preserving, i.e. the conclusion is implicitly contained in the premises.

Two examples of valid arguments are

1. Premises: "If A, then C." and "A is true". Conclusion:"C is true".
2. Premises: "Every F is G." and "b is F". Conclusion: "b is G".

Validity is something that is determined by the relationship between premises and conclusion ("does the premises support the conclusion?") and not by the actual truth of premises and/or conclusions[1]. It follows that valid arguments are risk-free arguments. Note also that the validity of an argument depends only on its form (or pattern) and not on the content (i.e. no matter what are the substitutes for A and C in the first argument). A valid argument is also called a *deductive* argument. Examples of deductive arguments are arguments in which the conclusion depends on some arithmetic or geometric computations or mathematical demonstrations. All arguments in pure mathematics are deductive.

An argument is *sound* if it is valid and its premises are true. Soundness for deductive logic has to do with both the validity and truth of the premises. Every sound argument, by definition, will have a true conclusion as well. For instance, the argument

All Italians play pretty good football
Gianluca is Italian.

Gianluca plays pretty good football $\quad \therefore$

is valid since the conclusion follows necessarily from the premises but not sound (otherwise, Gianluca would have been playing for Fiorentina AC).

2.2 Formal and empirical science

Epistemologists are used to distinguishing between formal and empirical sciences. Deductive arguments are the workhorse of formal sciences like geometry and mathematics. Those disciplines are built on a number of axioms, taken for true, and

[1] with the exception that a deductive argument with true premises and a false conclusion is necessarily invalid

on an effective truth-preserving mechanism. As such, they reason about a conceptual world, not necessarily in relation to the material world, where it is possible to define notions of truth, correctness, and soundness. On the empirical side, we find disciplines like physics, biology, and economics, whose statements are supposed to have a strong relationship with (some aspects of) sensible human experiences. Though empirical sciences often rely on formal sciences to define notions, concepts, and models, the validity of an empirical science proposition does not derive exclusively from its formal truth but essentially from the fact that its predictions are *in accordance with experimental observations*. Empirical sciences make then use of inductive arguments where the content of the conclusion is in some way intended to "go beyond" the content of the premises: a typical example is a prediction about a future event based on the observation of some events, i.e. the supervised learning scenario illustrated in Figure 1.1.

Modern empirical science, and the critical analysis of its inductive basis, began around the 16th and 17th centuries when the demand for new technologies (e.g. for military or exploration reasons) stimulated the inquiry into the origins of knowledge. In 1620 Francis Bacon, an English philosopher (1561-1626), published the *Novum Organum*, which presented an inductivist view of science. According to Bacon, scientific reasoning consists of making generalisations, or *inductions*, from observations to general *laws of nature* (e.g. moving to the conclusion that all swans are white after a number of historical observations). In other terms, the observations are supposed to induce the formulation of natural laws in the mind of the scientist.

2.3 Induction, projection, and abduction

Induction is defined as an inference in which one takes the past as grounds for beliefs about the future or the observed as grounds for beliefs about the unobserved. In other words, an inference is *ampliative*, i.e. it has more content in the conclusion than in the premises, unlike logical reasoning, which is deductive and non-ampliative. In inductive inference, the premises or departure points are called data or observations, and the conclusions are referred to as hypotheses[2]. A probabilistic language is usually adapted to express a hypothesis derived from induction. Induction has the following properties that contrast with the deductive pattern of inference [19]

1. The conclusion (e.g. hypothesis $h(D)$) follows non-monotonically from the premises (e.g. the dataset D). The addition of an extra premise (i.e. more data) might change the conclusion even when the extra premise does not contradict any of the other premises. In other terms, $D_1 \subset D_2 \not\Rightarrow h(D_1) \subset h(D_2)$, where $h(D)$ is the inductive consequence of the set of observations D.

2. The truth of the premises is not enough to guarantee the truth of the conclusion as there is no correspondence to the notion of deductive validity.

3. There is an information gain in induction since a hypothesis asserts more than data alone.

Another substantial difference is that, while logical arguments derive their validity from their form, this does not apply to inductive arguments: two inductive arguments may have the same form, but one may be good and the other not. So inductive inference is both useful and unsafe: no conclusion is a guaranteed truth, and it can dissolve even if no premise is removed.

[2]Note that this should not be confused with what mathematicians call *mathematical induction* which is a kind of deduction

There are several forms of inductive arguments:

1. Statement about a sample drawn from a population ⇒ Statement about the population as a whole
2. Statement about a population ⇒ Statement about a sample
3. Statement about a sample ⇒ Statement about a new sample
4. Observation of facts ⇒ Hypothesis

The third form of inference is also called *projection* [80] and is implemented in statistical learning by memory-based (e.g. lazy learning in Section 10.1.11) or transduction algorithms. The fourth form is also known as *abduction*, *explanatory inference*, *deduction in reverse* or *inference to the best explanation*. Abduction is a less ambitious form of induction since it does not infer to a generalisation but to a hypothesis that explains the data. In abduction, given $h \rightarrow D$ and the observation of D we infer the condition h. The rationale is that explanatory considerations are a guide to inference: in other words, the hypothesis that would (if correct) best explains the evidence is the hypothesis that is most likely to be correct. Note that this is the mechanism typically used in statistical hypothesis testing (Section 5.11).

An example of abduction is the Darwin theory. At his time, Darwin inferred the hypothesis of natural selection because, though not entailed by biological evidence, natural selection would provide the best explanation of that evidence. Darwin did not witness specific cases of evolution but formulated his hypothesis as an explanation of the available observations.

2.4 Hume and the induction problem

The downside of the induction success is its problematic and unsafe aspect, i.e. the projection of regularity onto unseen cases. The main problem of induction is how to justify the inference from the observed (data) to the unobserved (laws of nature), from the past (historical time series) to the future (e.g. prediction).

David Hume (1711-1776) was a Scottish philosopher who studied the problem of induction from a philosophic perspective. In 1739 he published *A treatise of human nature*, one of the most influential books of Western philosophy. According to Hume, *all reasonings concerning nature are founded on experience, and all reasonings from experience are founded on the supposition that the course of nature will continue uniformly the same* or in other terms that the future will be like the past. Any attempt to show, based on experience, that a regularity that has held in the past will hold in the future too will be circular (since based on the principle of regularity itself).

So empirical sciences rely on a supposition that, as shown by Hume, has *no logical necessity*. In other words, there is no contradiction in supposing that the future could be totally unlike the past (Figure 2.1) since we have no logical reason to expect that the past resembles the future.

So why do humans expect the future to be like the past? According to Hume, this is part of human nature: we have inductive habits, but we cannot justify them. The *principle of uniformity of nature* is not a priori true, nor it can be proved empirically. There is no reason beyond induction to justify inductive reasoning. Thus, Hume offers a naturalistic explanation of the psychological mechanism by which empirical predictions are made but not any rational justification for this practice. Our inductive practices rest on habit and *custom* and cannot be justified by rational argument. Induction is psychologically natural to us [80].

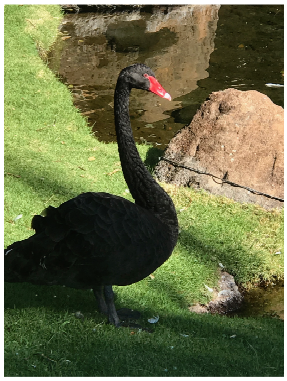


Figure 2.1: Falsification of the inductive hypothesis "Are all swans white?"

2.5 Logical positivism and verificationism

Logical positivism is a philosophical movement belonging to the wider family of empiricism, which developed in Europe after World War I and was established by a group of people (including Schlick, Neurath, and Carnap), also known as the Vienna Circle. They were inspired by the developments in sciences at the beginning of the XXth century, notably the work of Einstein. Two are the central ideas (or dogmas) of logical positivism: the *distinction between analytic and synthetic sentences* and the *verifiability theory of meaning* [80].

Analytic sentences are true or false, whatever is the world state. Analytical truths (e.g. in mathematics and logic) are necessary but somewhat empty. Mathematics does not describe the world and is independent of experience: they are a convention to use symbols in a particular way.

A synthetic sentence is true or false according to the actual state of the world. The value of synthetic sentences resides then in their method of verification. In other words, knowing the meaning of a sentence boils down to know how to verify it through observation. *Verificationism* is a strong empiricist principle: the only source of knowledge and the only source of meaning is observation. There are two categories of verifiable statements: i) observation statements (e.g. the temperature is below zero) which are directly verifiable, and ii) theoretical statements (indirectly verifiable) from which we can deduce observation statements.

Verificationists reject as "meaningless" statements specific to entire fields such as metaphysics, theology, ethics since they do not imply verifiable observations. Such statements may be meaningful in influencing emotions or human behaviour but provide no truth value, information, or factual content.

Science consists then of verifiable and then meaningful claims. According to the philosophy of logical positivism, a general statement or theory can be arrived at by inductive reasoning. Moreover, if such a theory is verified by observation or experiment, it can be promoted to a law. It follows that *verifiability* is the criterion of what is and what is not science (demarcation criterion).

Logical positivists stress that almost none of the evidence in everyday life and science may have the same degree of necessity as deductive logic. No evidence for a scientific theory is ultimately decisive since there is always the possibility of error, but this does not prevent science from being supported by evidence. The great aim of science is to discover and establish generalisations since there is no alternative to knowledge besides experience.

However, the verificationist ambition of grounding scientific truth in experience encountered some major problems related to the real possibility of verifying hypothesis in practice:

1. pure observations do not exist: observations are always *theory-laden*, i.e. they are inevitably affected by the theoretical beliefs (or expectations) of the investigator. Observations are neither neutral nor exhaustive, even in a big data world. To observe means to select what seems to be pertinent for the analysis, and this demands a specific and voluntary action from the experimenter (e.g. selection of the instrumentation or the language to communicate the results). Unfortunately, the analyst is often unaware of such selection, inducing then dangerous bias in the possible conclusions (Section 13.7.4).

2. no scientific assumption is testable in complete isolation (also known as the problem of *holism about testing*): the dogma of verificationism is naive since, in practice, only whole complex structured hypotheses may be submitted to empirical tests. Our ideas and hypothesis have contact with the experience only as a whole. Whenever we assess a theory by comparing it with observations, we need many additional assumptions to put a theoretical statement at the same level of observations.

3. unobservable entities escape from verification: one of the basic claims of logical positivists is that all aspects of science can be reduced to observational statements and submitted to verification (*in science, there are no depths, there is surface everywhere*). However, many successful and universally accepted scientific formulations rely on hidden structures and notions that are not directly observable (or mapped to observations in a univocal manner). Consider, for instance, the notions of gene or electron and the significant impact they have on the human understanding of reality.

Such criticisms contributed to the decline of the positivist program and opened the way to alternative interpretations of the knowledge discovery process.

2.6 Popper and the problem of induction

Karl Popper (1902-1994) is generally regarded as one of the greatest philosophers of science of the 20th century. His first achievement was an original definition of science based on the distinction between scientific and pseudo-scientific statements (also known as the *demarcation* problem). The solution he proposes is called *falsificationism* in opposition to the verificationism of positivists. Falsificationism claims that a hypothesis is scientific if and only if it has the potential to be refuted by some possible observation. To be scientific, a hypothesis has to entail testable prediction; in other words, it has to be bold, to take a risk. For instance, *All F is G* is a scientific statement while *Some F is G* is not. All scientific theories are universal in nature, and no finite collection of observation statements, however great, is logically equivalent to or can justify an unrestricted universal proposition. At the same time, we are never completely sure that a theory is true (aka *fallibilism*). A well-known example is Newton's physics which was considered for a long time as a gold standard of scientific theory until it was shown to be false in several respects.

Popper was sceptical about all forms of confirmation and notably about the theory of confirmation proposed by empiricists. According to him, the only good reasoning is deductively valid reasoning. According to Popper, humans or scientists do not make inductions; they make conjectures (or hypotheses) and test them (or their logical consequences obtained by deduction). If the test is successful, the conjecture is corroborated but never verified or proven. *Confirmation is thus a*

myth: no theory or belief about the world can be proven. Though no number of positive experimental outcomes can demonstrate the truth of a scientific theory, a single genuine counter instance can refute it (*modus tollens*). It follows that we learn something by deduction and not by induction. If the empirical test of the conjecture is not successful, the conjecture is refuted. The refutation of a hypothesis leads us (or the scientific community) to revise it or devise a more robust one. The final result is that scientific laws are falsifiable yet strictly unverifiable.

Scientific knowledge evolves via a two-step cycle that repeats endlessly: the first stage is made of *conjecture* making. The second stage is attempted *refutation*, when the hypothesis is submitted to critical testing. The most important qualities of a scientist are then imaginative (almost artistic) creativity, and rigorous testing.

Also, according to Popper, there are no "pure" or theory-free observations. Observation is always selective: it needs a chosen object, a definite task, an interest, a point of view, a problem. Observation is theory-laden and involves applying theoretical terms, descriptive language, and a conceptual scheme to particular experimental situations.

2.7 The hypothetico-deductive method and instrumentalism

Nowadays, the most commonly agreed vision of science (*hypothetico-deductivism*) merges the main ideas of logic and induction, realism and empiricism, of verificationism and falsificationism. According to this vision, science is a process where scientists formulate hypotheses (e.g. inductive step after a preliminary stage where observations were collected) and then deduce observational predictions from them. If predictions are accurate, then the theory is supported (in agreement with logical positivists) or (e.g. in Bayesian terms) its degree of truth increases. If predictions are not accurate, the theory is disconfirmed (this is coherent with Popper). The more tests a theory passes, the more confidence we can have in its truth[3].

If the value of scientific models is intimately related to the quality of prediction, they should be seen more as useful tools (or instruments) than a faithful representation of reality. The notion of *instrumentalism* was introduced by Van Fraassen [180]. An instrumentalist does not worry about whether a theory is a true description of the world (e.g. if electrons really exist). The role of a theory is to establish a good prediction. The question of whether our theory has some deeper match in the real world will never have an answer so we should stop asking it.

Van Fraassen thinks that the only aim of theories is to accurately describe the observable parts of the world. If this happens, they are empirically adequate. Trying to address the hidden nature of reality is of no interest to science.

2.8 Epistemology and machine learning: the cross-fertilisation

This chapter sketched some contributions of epistemology to the understanding of how humans extract and attain knowledge from observations, in particular during the scientific endeavour.

Machine learning, the topic of this book, is a computationally based approach aiming to produce knowledge from observed data. If we make the basic assumption

[3]Note also that the predominating use of probabilistic hypotheses to take into account noisy observation is in contradiction with the restrictive vision of Popper on logical deduction.

that both epistemology and machine learning refers to the same notion of knowledge (i.e. knowledge useful for human beings), an epistemological approach can be useful to understand both limits and potential of machine learning. The author is convinced that a fruitful cross-fertilisation can derive from a stronger synergy between epistemology and machine learning. In particular, he expects the following contributions from a machine learning approach to the study of knowledge discovery:

- machine learning deals intimately with induction or how observations can induce and/or confirm a theory, one of the most fundamental problems of philosophy of science. Also, it implements in a reproducible and testable way the mechanism of learning, generation of hypothesis, and testing.
- machine learning is today unavoidable in supporting discovery in scientific domains where human experts would be overwhelmed by complexity and dimensionality.
- machine learning is a key factor of the revolution transforming all empirical sciences into data sciences, i.e. inductive disciplines where the quality and the accuracy of the discoveries are strictly dependent on the capacity of extracting accurate information, predictions, or models from large amounts of observed data.
- machine learning generalises and democratises the notion of observed evidence by making it converging with the notion of data. Every instrument (or tool or simulator) producing data can be taken as the starting point of a knowledge discovery process. This extends the common notion of experimental evidence adopted in conventional sciences, like physics. A financial transaction, a tweet, or a GPS trace may be for some domains as informative as a CERN multimillion experience in physics.
- machine learning is the ultimate step in the scientific process moving from the optimistic objective of finding true descriptions of reality to the more realistic goal of attaining accurate models of observations.

At the same time, there a number of lessons that young data scientists could learn from ancient and recent philosophers of science:

- A critical analysis of the role of observations and data: all empirical sciences derive their justification from the fact of being firmly founded on experiments. The distinctive nature of machine learning, and a reason for its success, is the automatic process of extracting knowledge from data. Observations and data are then necessary conditions for triggering any knowledge discovery procedure. There is, however, the risk to sanctify the role of data (or facts) as an unquestionable and objective foundation of truth. This excess has been several times discussed and criticised by epistemologists (notably the critics of logical positivism). Pure facts and theory-neutral observations do not exist, not even in a big-data world. Observations (and more specifically experiments) are never passive or beyond any suspicion: they are the results of a specific human initiative (or intervention) that can be dictated by specific objectives, constraints, and motivation. The presupposition that the truth of empirical statements can be securely established only by observation is a naive attitude that could lead to disastrous consequences (e.g. sexist or racist AI applications due to sampling bias) [40].
- Skepticism about induction: the Hume analysis, confirmed by theoretical analysis in machine learning (notable the no-free-lunch theorem) reminds us that

a *tabula rasa* approach going from data to knowledge is not possible. There is no univocal (or optimal) way of proceeding from observations to models since every learning process relies on (explicit or more often implicit) assumptions. This is also related to the notion of *undetermination of theory by evidence*, which means that there will always be a range of alternative theories compatible with observations.

- Importance of hypothesis generation and validation: this important lesson comes straight from Popper and associates the scientific character of a knowledge discovery process to the possibility of falsification. In that sense, machine learning complies with the Popper interpretation of science and goes further by proposing a set of strategies for automatically generating hypotheses and validating them by empirical evidence. In more actual terms, the best way to ensure falsifiability to computation sciences is reproducibility and interpretability. These two aspects are essential to guarantee the respect of high standards of quality and rigour in computational approaches to knowledge discovery. Forgetting the assumptions underlying any data-driven effort may lead to accepting biased conclusions and misinterpretations (e.g. from a causal perspective), which are dangerously endorsed by the size of the dataset or the complexity of the algorithmic approach.

- Model as tools: the adoption of complex representation of reality (though characterised by high-level notions and principles) makes difficult, if not unrealistic, the validation of all the components of a model. As a consequence, a model should not be considered as a faithful copy of reality but as a convenient abstraction, which, if confirmed by experimental validation, becomes a useful instrument for prediction and decision making.

- The confirmation of a hypothesis requires taking into account the procedures involved in generating data: confirmation of a hypothesis with observations is not a go-nogo process. Since new evidence changes degrees of validity (or degree of belief), a probabilistic approach is necessary. This is why any introduction to machine learning needs first an introduction to probability, probabilistic reasoning, and then statistics.

Chapter 3

Foundations of probability

Uncertainty is inescapable in the real world. Even without resort to indeterminism, its pervasiveness is due to the complexity of reality and the limitations of human observational skills and modelling capabilities. According to [117] *uncertainty arises because of limitations in our ability to observe the world, limitations in our ability to model it, and possibly even because of innate nondeterminism.* Probability theory is one of many disciplines [141] concerned with the study of uncertain (or random) phenomena. It is also, according to the author, one of the most successful ones in terms of formalisation, theoretical and algorithmic developments and practical applications. For this reason, in this book, we will adopt probability as the mathematical language to describe and quantify uncertainty. Uncertain phenomena, although not predictable in a deterministic fashion, may present some regularities and consequently be described mathematically by idealised probabilistic models. These models consist of a list of all possible outcomes together with the respective probabilities. The theory of probability makes it possible to infer from these models the patterns of future behaviour.

This chapter presents the basic notions of probability which serve as a necessary background to understand the statistical aspects of machine learning. We ask the reader to become acquainted with two aspects: the notion of a random variable as a compact representation of uncertain knowledge and the use of probability as an effective formal tool to manipulate and process such uncertain information. In particular, we suggest the reader give special attention to the notions of conditional and joint probability. As we will see in the following, these two related notions are extensively used by statistical modelling and machine learning to define the dependence and the relationships between random variables.

3.1 The random model of uncertainty

We define a *random experiment* as any action or process which generates results or observations which cannot be predicted with certainty. *Uncertainty stems from the existence of alternatives.* In other words, each uncertain phenomenon is characterised by a multiplicity of possible configurations or outcomes. Weather is uncertain since it can take multiple forms (e.g. sunny, rainy, cloudy,...). Other examples of random experiments are tossing a coin, rolling dice, passing an exam or measuring the time to reach home.

A random experiment is then characterised by a *sample space* Ω that is a (finite or infinite) set of all the possible outcomes (or configurations) ω of the experiment. The elements of the set Ω are called *experimental outcomes* or *realisations.* For example, in the die experiment, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}$ and ω_i stands for the outcome

corresponding to getting the face with the number i. If ω is the outcome of a measurement of some physical quantity, e.g. pressure, then we could have $\Omega = \mathbb{R}^+$.

The representation of an uncertain phenomenon is the result of a modelling activity and, as such, it is not necessarily unique. In other terms different representations of a random experiment are possible. In the die experiment, we could define an alternative sample space made of two sole outcomes: numbers equal to and different from 1. Also, we could be interested in representing the uncertainty of two consecutive tosses. In that case, the outcome would be the pair $(\omega_{(t)}, \omega_{(t+1)})$ where $\omega_{(t)}$ is the outcome at time t.

Uncertainty stems from variability. Each time we observe a random phenomenon, we may observe different outcomes. In probabilistic jargon, observing a random phenomenon is interpreted as the realisation of a random experiment. A single performance of a random experiment is called a *trial*. This means that after each trial, we observe one outcome $\omega_i \in \Omega$.

A subset of experimental outcomes is called an *event*. Consider a trial that generated the outcome ω_i: we say that an event $\mathcal{E}$ occurred during the trial if the set $\mathcal{E}$ contains the element ω_i. For example, in the die experiment, an event (denoted *odd number*) is the set of odd values $\mathcal{E} = \{\omega_1, \omega_3, \omega_5\}$. This means that when we observe the outcome ω_5 the event *odd number* takes place.

An event composed of a single outcome, e.g. $\mathcal{E} = \{\omega_1\}$ is called an *elementary event*.

Note that since events $\mathcal{E}$ are subsets, we can apply to them the terminology of the set theory:

- Ω refers to the certain event i.e. the event that occurs in every trial.
- the notation
$$\mathcal{E}^c = \{\omega : \omega \notin \mathcal{E}\}$$
denotes the complement of $\mathcal{E}$.
- the notation
$$\mathcal{E}_1 \cup \mathcal{E}_2 = \{\omega \in \Omega : \omega \in \mathcal{E}_1 \text{ OR } \omega \in \mathcal{E}_2\}$$
refers to the event that occurs when $\mathcal{E}_1$ or $\mathcal{E}_2$ or both occur.
- the notation
$$\mathcal{E}_1 \cap \mathcal{E}_2 = \{\omega \in \Omega : \omega \in \mathcal{E}_1 \text{ AND } \omega \in \mathcal{E}_2\}$$
refers to the event that occurs when both $\mathcal{E}_1$ and $\mathcal{E}_2$ occur.
- two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are *mutually exclusive* or *disjoint* if
$$\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset \tag{3.1.1}$$
that is each time that $\mathcal{E}_1$ occurs, $\mathcal{E}_2$ does not occur as well.
- a partition of Ω is a set of disjoint sets $\mathcal{E}_j$, $j = 1, \ldots, J$ (i.e. $\mathcal{E}_{j_1} \cap \mathcal{E}_{j_2} = \emptyset$ $\forall j_i, j_2 \in J$) such that
$$\cup_{j=1}^{J} \mathcal{E}_j = \Omega$$
- given an event $\mathcal{E}$ we define the *indicator function* of $\mathcal{E}$ by
$$I_{\mathcal{E}}(\omega) = \begin{cases} 1 & \text{if } \omega \in \mathcal{E} \\ 0 & \text{if } \omega \notin \mathcal{E} \end{cases} \tag{3.1.2}$$

Let us consider now the notion of *class of events*. An arbitrary collection of subsets of Ω is not a class of events. We require that if $\mathcal{E}_1$ and $\mathcal{E}_2$ are events, the same also holds for the intersection $\mathcal{E}_1 \cap \mathcal{E}_2$ and the union $\mathcal{E}_1 \cup \mathcal{E}_2$. A set of events that satisfies these conditions is called, in mathematical terms, a *Borel field* [140]. We will consider only Borel fields since we want to deal not only with the probabilities of single events but also with the probabilities of their unions and intersections.

3.1.1 Axiomatic definition of probability

Probability is a measure of uncertainty. Once a random experiment is defined, this measure associates to each possible outcome ω a number between 0 and 1. It follows that we can assign to each event $\mathcal{E}$ a real number $\text{Prob}\{\mathcal{E}\} \in [0, 1]$ which denotes the *probability of the event* $\mathcal{E}$. The measure associated with the event including all possibilities is 1. The function $\text{Prob}\{\cdot\} : 2^{\Omega} \to [0, 1]$ is called *probability measure* or *probability distribution* and must satisfy the following three axioms:

1. $\text{Prob}\{\mathcal{E}\} \geq 0$ for any $\mathcal{E}$.
2. $\text{Prob}\{\Omega\} = 1$
3. $\text{Prob}\{\mathcal{E}_1 \cup \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_2\}$ if $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive (Equation (3.1.1)).

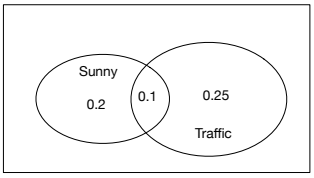
These conditions are known as the *axioms of the theory of probability* [118]. The first axiom states that all the probabilities are nonnegative real numbers. The second axiom attributes a probability of unity to the universal event Ω, thus providing a normalisation of the probability measure. The third axiom states that the probability function must be additive for disjoint events, consistently with the intuitive idea of how probabilities behave.

So from a mathematician perspective, probability is easy to define: it is a countably additive set function defined on a *Borel field*, with a total mass of one. Every probabilistic property, for instance $\mathcal{E}_1 \subset \mathcal{E}_2 \Rightarrow \text{Prob}\{\mathcal{E}_1\} \leq \text{Prob}\{\mathcal{E}_2\}$ or $\text{Prob}\{\mathcal{E}^c\} = 1 - \text{Prob}\{\mathcal{E}\}$, can be derived directly or indirectly from the axioms (and only the axioms).

There are many interpretations and justifications of these axioms, and we discuss the frequentist and the Bayesian interpretation in Section 3.1.4 briefly. What is relevant here is that the probability function is a formalisation of uncertainty and that most of its properties and results appear to be coherent with the human perception of uncertainty [108].

3.1.2 Visualisation of probability measures

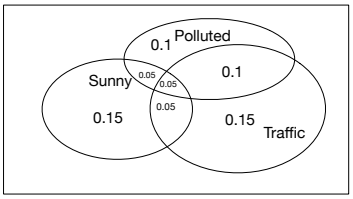
Since probabilistic events are sets of outcomes, Venn diagrams are a convenient manner to illustrate the relations between events and the notion of probability measure. Suppose that you are a biker and you are interested in representing the variability of weather and traffic conditions in your town in the morning. In particular, you are interested in the probability that the morning will be sunny (or not) and the road busy (or not). In order to formalise your practical issue, you could define the uncertainty about the morning state by defining a sample space which is the set of all possible morning conditions. Two events are of interest here: sunny mornings and traffic conditions. What is the relationship and probability of such two events? Figure 3.1 illustrates the sample space, the two events, and the (hypothetical) probability measures by means of a Venn diagram and two different tabular representations. The three representations in Figure 3.1 convey the same information in different manners. Notwithstanding, they do not necessarily scale-up in the same manner if we take into consideration a larger number of events.



	Traffic	No Traffic
Sunny	0.1	0.2
Not Sunny	0.25	0.45

Sunny	Traffic	P
0	0	0.45
0	1	0.25
1	0	0.2
1	1	0.1

Figure 3.1: Visualisation of two events and probability measures: Venn diagram (left), two-way table (center), probability distribution table (right)



Sunny	Traffic	Polluted	P
0	0	0	0.35
0	0	1	0.1
0	1	0	0.15
0	1	1	0.1
1	0	0	0.15
1	0	1	0.05
1	1	0	0.05
1	1	1	0.05

Figure 3.2: Visualisation of three events and related probability measures: Venn diagram (left), probability distribution table (right)

For instance, for n events the Venn diagram should contain all 2^n hypothetically possible zones[1].

Suppose that you are also interested in another type of event, i.e. the air quality. Adding such an event to your probability representation would make your Venn representation more complicated and the two-way table inadequate (Figure 3.2). The visualisation will still be more difficult to handle and interpret if we deal with more than three events.

Given their difficulty of encoding information in realistic probabilistic settings, Venn diagrams are a pedagogical yet very limited tool for representing uncertainty.

Once introduced the notion of probability, a major question remains still open: how to compute the probability value $\text{Prob}\{\mathcal{E}\}$ for a generic event $\mathcal{E}$? The assignment of probabilities is perhaps the most difficult aspect of constructing probabilistic models. Although the theory of probability is neutral, that is it can make inferences regardless of the actual probability values, its results will be strongly affected by the choice of a particular assignment. This means that if the assignments are inaccurate, the predictions of the model will be misleading and will not reflect the real behaviour of the modelled phenomenon. In the following sections, we are going to present some procedures which are typically adopted in practice.

3.1.3 Symmetrical definition of probability

Consider a random experiment where the sample space is made of a finite number M of symmetric outcomes (i.e., they are equally likely to occur). Let the number of outcomes that are favourable to the event $\mathcal{E}$ (i.e. the event $\mathcal{E}$ takes place if one of them occurs) be $M_{\mathcal{E}}$.

An intuitive definition of probability (also known as the *classical definition*) of the event $\mathcal{E}$, that adheres to the axioms, is

$$\text{Prob}\{\mathcal{E}\} = \frac{M_{\mathcal{E}}}{M} \tag{3.1.3}$$

[1]see Wikipedia `https://en.wikipedia.org/wiki/Venn_diagram`

In other words, according to the principle of indifference (a term popularised by J.M. Keynes in 1921), we have that *the probability of an event equals the ratio of its favourable outcomes to the total number of outcomes provided that all outcomes are equally likely* [140]. The computation of this quantity requires combinatorial methods for counting the favourable outcomes. This is typically the approach adopted for a fair die. Also, in most cases, the symmetric hypothesis is accepted as self-evident: *if a ball is selected at random from a bowl containing W white balls and B black balls, the probability that we select a white one is $W/(W+B)$.*

Note that this number is determined without any experimentation and is based on symmetrical and finite space assumptions. But how to be sure that the symmetrical hypothesis holds? and that is invariant? Think, for instance, to the probability that a newborn be a boy. Is this a symmetric case? More generally, how would one define the probability of an event if the symmetrical hypothesis does not necessarily hold or the space is not finite?

3.1.4 Frequentist definition of probability

Let us consider a random experiment and an event $\mathcal{E}$. Suppose we repeat the experiment N times and that we record the number of times $N_{\mathcal{E}}$ that the event $\mathcal{E}$ occurs. The quantity

$$\frac{N_{\mathcal{E}}}{N} \tag{3.1.4}$$

comprised between 0 and 1 is known as the *relative frequency* of $\mathcal{E}$. It can be observed that if the experiment is carried out a large number of times *under exactly the same conditions*, the frequency converges to a fixed value for increasing N. This observation led von Mises to use the notion of frequency as a foundation for the notion of probability.

Definition 1.1 (von Mises)**.** The probability $\text{Prob}\{\mathcal{E}\}$ of an event $\mathcal{E}$ is the limit

$$\text{Prob}\{\mathcal{E}\} = \lim_{N\to\infty} \frac{N_{\mathcal{E}}}{N} \tag{3.1.5}$$

where N is the number of observations and $N_{\mathcal{E}}$ is the number of times that $\mathcal{E}$ occurred.

This definition appears reasonable, and it is compatible with the axioms in Section 3.1.1. However, in practice, in any physical experience, the number N is finite[2], and the limit has to be accepted as a hypothesis, not as a number that can be determined experimentally [140].

Moreover, the assumption *under exactly the same conditions* is not as innocuous as it seems. How could you ensure that two experiments occur under exactly the same conditions? And what do those conditions refer to? Temperature, humidity, obsolescence of the equipment? Are humans really able to control exactly all of them? Would you be able to reproduce the exact same conditions of an experiment?

Notwithstanding, the frequentist interpretation is very important to show the links between theory and application. At the same time, it appears inadequate to represent probability when it is used to model a subjective degree of belief. Think, for instance, to the probability that your professor wins a Nobel Prize: how to define in such case a number N of repetitions?

An important alternative interpretation of the probability measure comes then from the Bayesian approach. This approach proposes a *degree-of-belief* interpretation of probability according to which $\text{Prob}\{\mathcal{E}\}$ measures an observer's strength of belief that $\mathcal{E}$ is or will be true [188]. This manuscript will not cover the Bayesian

[2] As Keynes said "In the long run we are all dead".

approach to statistics and data analysis for the sake of compactness, though the author is well aware that Bayesian machine learning approaches are more and more common and successful. Readers interested in the foundations of the Bayesian interpretation of probability are referred to [108]. Readers interested in introductions to Bayesian machine learning are referred to [77, 13].

3.1.5 The Law of Large Numbers

A well-known justification of the frequentist approach is provided by the Weak Law of Large Numbers, proposed by Bernoulli.

Theorem 1.2. *Let* $Prob\{\mathcal{E}\} = p$ *and suppose that the event* $\mathcal{E}$ *occurs* $N_{\mathcal{E}}$ *times in* N *trials. Then,* $\frac{N_{\mathcal{E}}}{N}$ *converges to* p in probability*, that is, for any* $\epsilon > 0$*,*

$$Prob\left\{\left|\frac{N_{\mathcal{E}}}{N} - p\right| \leq \epsilon\right\} \to 1 \quad as \quad N \to \infty$$

According to this theorem, the ratio $N_{\mathcal{E}}/N$ is close to p in the sense that, for any $\epsilon > 0$, the probability that $|N_{\mathcal{E}}/N - p| \leq \epsilon$ tends to 1 as $N \to \infty$. This result justifies the widespread use of the frequentist approach (e.g. in Monte Carlo simulation) to illustrate or numerically solve probability problems. The relation between frequency and probability is illustrated by the Shiny dashboard `lawlarge.R` (package `gbcode`).

Note that such a result does not imply that the number $N_{\mathcal{E}}$ will be close to Np as one could naively infer from (3.1.5). In fact,

$$\text{Prob}\{N_{\mathcal{E}} = Np\} \approx \frac{1}{\sqrt{2\pi Np(1-p)}} \to 0, \qquad \text{as } N \to \infty \tag{3.1.6}$$

For instance, in a fair coin-tossing game, this law does not imply that the absolute difference between the number of heads and tails should oscillate close to zero [176] (Figure 3.3). On the contrary, it could happen that the absolute difference keeps growing (though at a slower rate than the number of tosses) as shown in the R script `freq.R` and the Shiny dashboard `lawlarge.R`.

3.1.6 Independence and conditional probability

Let us consider two different events. We have already introduced the notions of complementary and disjoint events. Another important definition is the definition of independent events and the related notion of conditional probability. This notion is essential in machine learning since supervised learning aims to detect and model (in)dependencies by estimating conditional probabilities.

Definition 1.3 (Independent events)**.** Two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are *independent* if and only if

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2\} \tag{3.1.7}$$

and we write $\mathcal{E}_1 \perp\!\!\!\perp \mathcal{E}_2$.

The probability $\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\}$ of seeing two events occurring together is also known as *joint probability* and often noted as $\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2\}$. If two events are independent the joint probability depends only on the two individual probabilities. As an example of two independent events, think of two outcomes of a roulette wheel or of two coins tossed simultaneously.

From an uncertain reasoning perspective, independence is a very simplistic assumption since the occurrence (or the observation) of one event has no influence on the occurrence of the other, or similarly that the second event has no memory

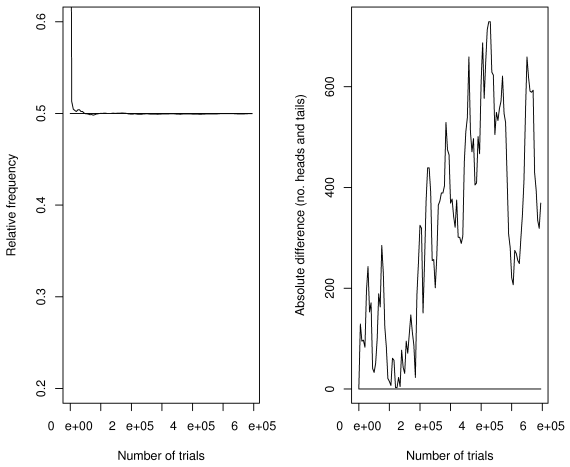


Figure 3.3: Fair coin-tossing random experiment: evolution of the relative frequency (left) and of the absolute difference (right) between the number of heads and tails (R script `freq.R` in `gbcode`).

of the first. In other words, independence considers the uncertainty of a complex joint event as a function of the uncertainties of its components[3]. This makes the reasoning much simpler but, at the same time, too rough.

Exercise

Suppose that a fair die is rolled and that the number ω appears. Let $\mathcal{E}_1$ be the event that the number ω is even, $\mathcal{E}_2$ be the event that the number ω is greater than or equal to 3, $\mathcal{E}_3$ be the event that the number ω is a 4,5 or 6.

Are the events $\mathcal{E}_1$ and $\mathcal{E}_2$ independent? Are the events $\mathcal{E}_1$ and $\mathcal{E}_3$ independent?

•

Let $\mathcal{E}_1$ be an event such that $\text{Prob}\{\mathcal{E}_1\} > 0$ and $\mathcal{E}_2$ a second event. We define the conditional probability of $\mathcal{E}_2$, given that $\mathcal{E}_1$ has occurred, the revised probability of $\mathcal{E}_2$ after we learn about $\mathcal{E}_1$ occurrence:

Definition 1.4 (Conditional probability)**.** If $\text{Prob}\{\mathcal{E}_1\} > 0$ then the conditional probability of $\mathcal{E}_2$ given $\mathcal{E}_1$ is

$$\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} = \frac{\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\}}{\text{Prob}\{\mathcal{E}_1\}} \tag{3.1.8}$$

The following result derives from the definition of conditional probability.

Lemma 1. If $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent events, then

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\} \tag{3.1.9}$$

In qualitative terms, the independence of two events means that the fact of observing (or knowing) that one of these events (e.g. $\mathcal{E}_1$) occurred does not change the probability that the other (e.g. $\mathcal{E}_2$) will occur.

[3]We refer the interested reader to the distinction between extensional and intensional reasoning in [145]. Extensional reasoning (e.g. logics) always makes an assumption of independence, while intensional reasoning (e.g. probability) consider independence as an exception.

Example

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ two disjoint events with positive probability. Can they be independent? The answer is no since

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\emptyset\} = 0 \neq \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2\} > 0$$

or equivalently $\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = 0$. We can interpret this result by noting that if two events are disjoint, the realisation of one of them is highly informative about the realisation of the other. For instance, though it is very probable that Italy will win the next football World Cup ($\text{Prob}\{\mathcal{E}_1\} >> 0$) , this probability goes to zero if the (rare yet possible) event $\mathcal{E}_2$ ("World cup won by Belgium") occurs ($\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = 0$). The two events are then dependent.

•

Exercise

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two independent events, and $\mathcal{E}_1^c$ the complement of $\mathcal{E}_1$. Are $\mathcal{E}_1^c$ and $\mathcal{E}_2$ independent?

•

Exercise

Consider the sample space Ω and the two events $\mathcal{E}_1$ and $\mathcal{E}_2$ in Figure 3.4. Suppose that the probability of the two events is proportional to the surface of the regions. From the Figure we compute

$$\text{Prob}\{\mathcal{E}_1\} = \frac{9}{100} = 0.09 \tag{3.1.10}$$

$$\text{Prob}\{\mathcal{E}_2\} = \frac{20}{100} = 0.2 \tag{3.1.11}$$

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \frac{1}{100} = 0.01 \neq \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2\} \tag{3.1.12}$$

$$\text{Prob}\{\mathcal{E}_1 \cup \mathcal{E}_2\} = 0.28 = \text{Prob}\{\mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_2\} - \text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} \tag{3.1.13}$$

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = \frac{1}{20} = 0.05 \neq \text{Prob}\{\mathcal{E}_1\} \tag{3.1.14}$$

$$\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} = \frac{1}{9} \neq \text{Prob}\{\mathcal{E}_2\} \tag{3.1.15}$$

and then derive the following conclusions: the events $\mathcal{E}_1$ and $\mathcal{E}_2$ are neither disjoint nor independent. Also, it is more probable that $\mathcal{E}_2$ occurs given that $\mathcal{E}_1$ occurred rather than the opposite.

•

From (3.1.8) we derive

$$\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} \tag{3.1.16}$$

If we replace the event $\mathcal{E}_2$ with the intersection of two events $\mathcal{E}_2$ and $\mathcal{E}_3$, from (3.1.16) we obtain

$$\begin{aligned}\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\} &= \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2, \mathcal{E}_3|\mathcal{E}_1\} = \\ &\text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2|\mathcal{E}_3, \mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_3|\mathcal{E}_1\} = \text{Prob}\{\mathcal{E}_1, \mathcal{E}_3\}\,\text{Prob}\{\mathcal{E}_2|\mathcal{E}_3, \mathcal{E}_1\}\end{aligned}$$

If we divide both terms by $\text{Prob}\{\mathcal{E}_3\}$ we obtain

$$\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2|\mathcal{E}_3\} = \text{Prob}\{\mathcal{E}_1|\mathcal{E}_3\}\,\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1, \mathcal{E}_3\} \tag{3.1.17}$$

which is the conditioned version of (3.1.16).

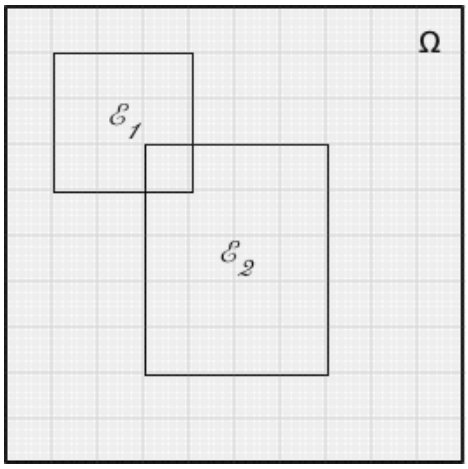


Figure 3.4: Events in a sample space.

3.1.7 The chain rule

The equation (3.1.16) shows that a joint probability can be *factorised* as the product of a conditional and an unconditional probability. In more general terms, the following rule holds.

Definition 1.5 (Chain rule)**.** For any sequence of events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$,

$$\begin{aligned}&\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n\} = \\ &\qquad \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_3|\mathcal{E}_1, \mathcal{E}_2\}\ldots\text{Prob}\{\mathcal{E}_n|\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{n-1}\}\end{aligned}$$

We will see in Chapter 4 that the chain rule factorisation and the notion of conditional independence play a major role in the adoption of graphical models to represent probability distributions.

3.1.8 The law of total probability and the Bayes' theorem

Let us consider an indeterminate practical situation where a set of events $\mathcal{E}_1$, $\mathcal{E}_2$,..., $\mathcal{E}_k$ may occur. Suppose that no two such events may occur simultaneously, but at least one of them must occur. This means that $\mathcal{E}_1$, $\mathcal{E}_2$,..., $\mathcal{E}_k$ are mutually exclusive and exhaustive or, in other terms, that they form a partition of Ω. The following two theorems can be proven.

Theorem 1.6 (Law of total probability)**.** *Let* $\text{Prob}\{\mathcal{E}_i\}$*,* $i = 1, \ldots, k$ *denote the probability of the* i*th event* $\mathcal{E}_i$ *and* $\text{Prob}\{\mathcal{E}|\mathcal{E}_i\}$*,* $i = 1, \ldots, k$ *the conditional probability of a generic event* $\mathcal{E}$ *given that* $\mathcal{E}_i$ *has occurred. It can be shown that*

$$\text{Prob}\{\mathcal{E}\} = \sum_{i=1}^{k} \text{Prob}\{\mathcal{E}|\mathcal{E}_i\}\,\text{Prob}\{\mathcal{E}_i\} = \sum_{i=1}^{k} \text{Prob}\{\mathcal{E} \cap \mathcal{E}_i\} \tag{3.1.18}$$

The quantity $\text{Prob}\{\mathcal{E}\}$ is referred to as *marginal probability* and denotes the probability of the event $\mathcal{E}$ irrespective of the occurrence of other events. A commonsense interpretation of this theorem is that if an event $\mathcal{E}$ (e.g. an effect) depends on the realisation of k disjoint events (e.g. causes), the probability of observing $\mathcal{E}$, is a weighted average of each single conditional probability $\text{Prob}\{\mathcal{E}|\mathcal{E}_i\}$ where the weights are given by the marginal probabilities of each event $\mathcal{E}_i, i = 1, \ldots, k$. For instance, we can compute the probability that the highway is busy once we

know the probability that an accident occurred or not (two disjoint events) and the conditional probabilities of traffic given the occurrence (or not) of an accident.

Theorem 1.7 (Bayes' theorem)**.** *The conditional ("inverse") probability of any $\mathcal{E}_i$, $i = 1, \ldots, k$ given that $\mathcal{E}$ has occurred is given by*

$$Prob\{\mathcal{E}_i|\mathcal{E}\} = \frac{Prob\{\mathcal{E}|\mathcal{E}_i\}\, Prob\{\mathcal{E}_i\}}{\sum_{j=1}^{k} Prob\{\mathcal{E}|\mathcal{E}_j\}\, Prob\{\mathcal{E}_j\}} = \frac{Prob\{\mathcal{E}, \mathcal{E}_i\}}{Prob\{\mathcal{E}\}} \qquad i = 1, \ldots, k \tag{3.1.19}$$

It follows that the Bayes theorem is the only sound way to derive from a conditional probability $\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}$ its inverse

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = \frac{\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_1\}}{\text{Prob}\{\mathcal{E}_2\}} \tag{3.1.20}$$

Any alternative derivation (or shortcut) will lead inevitably to fallacious reasoning and inconsistent results (see the Prosecutor fallacy discussion in Section 3.1.9).

It may be useful also to write a conditioning version of the total probability. Given an event $\mathcal{E}'$ and the set $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k$ of mutually exclusive events:

$$\text{Prob}\{\mathcal{E}|\mathcal{E}'\} = \sum_{i=1}^{k} \text{Prob}\{\mathcal{E}|\mathcal{E}_i, \mathcal{E}'\}\,\text{Prob}\{\mathcal{E}_i|\mathcal{E}'\} \tag{3.1.21}$$

From (3.1.20) and by conditioning on a third event $\mathcal{E}_3$, we obtain a conditioning version of the Bayes theorem

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2, \mathcal{E}_3\} = \frac{\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1, \mathcal{E}_3\}\,\text{Prob}\{\mathcal{E}_1|\mathcal{E}_3\}}{\text{Prob}\{\mathcal{E}_2|\mathcal{E}_3\}} \tag{3.1.22}$$

as long as $\text{Prob}\{\mathcal{E}_2|\mathcal{E}_3\} > 0$

Example

Suppose that $k = 2$ and

- $\mathcal{E}_1$ is the event: "Tomorrow is going to rain".
- $\mathcal{E}_2$ is the event: "Tomorrow is not going to rain".
- $\mathcal{E}$ is the event: "Tonight is chilly and windy".

The knowledge of $\text{Prob}\{\mathcal{E}_1\}$, $\text{Prob}\{\mathcal{E}_2\}$ and $\text{Prob}\{\mathcal{E}|\mathcal{E}_k\}$, $k = 1, 2$ makes possible the computation of $\text{Prob}\{\mathcal{E}_k|\mathcal{E}\}$.

•

Exercise

Verify the validity of the law of total probability and of the Bayes theorem for the problem in Figure 3.5.

•

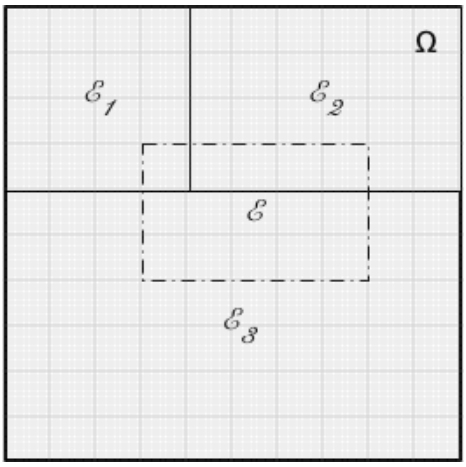


Figure 3.5: Events in a sample space

3.1.9 Direct and inverse conditional probability

The notion of conditional probability is central in probability and machine learning, but it is often prone to dangerous misunderstanding, for instance, when inappropriately used in domains like medical sciences or law. The most common error consists of taking a conditional probability $\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\}$ for its inverse $\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}$. This is also known as the *prosecutor fallacy*, as discussed in an example later.

The first important element to keep in mind is that for any fixed $\mathcal{E}_1$, the quantity $\text{Prob}\{\cdot|\mathcal{E}_1\}$ still satisfies the axioms of probability, i.e. the function $\text{Prob}\{\cdot|\mathcal{E}_1\}$ is itself a probability measure. *Conditional probabilities are probabilities* [27]. However, this does not generally hold for $\text{Prob}\{\mathcal{E}_1|\cdot\}$, which corresponds to fix the term $\mathcal{E}_1$ on the left of the conditional bar. For instance if $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_4$ are disjoint events we have

$$\text{Prob}\{\mathcal{E}_2 \cup \mathcal{E}_3 \cup \mathcal{E}_4|\mathcal{E}_1\} = \text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_3|\mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_4|\mathcal{E}_1\}$$

in agreement with the third axiom (Section 3.1.1) but

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2 \cup \mathcal{E}_3 \cup \mathcal{E}_4\} \neq \text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} + \text{Prob}\{\mathcal{E}_1|\mathcal{E}_3\} + \text{Prob}\{\mathcal{E}_1|\mathcal{E}_4\}$$

Also it is generally not the case that $\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} = \text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\}$. As a consequence if $\mathcal{E}_1$ and $\mathcal{E}_2$ are not independent then

$$\text{Prob}\{\mathcal{E}_1^c|\mathcal{E}_2\} = 1 - \text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\}$$

but

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2^c\} \neq 1 - \text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} \tag{3.1.23}$$

where $\mathcal{E}^c$ denotes the complement of $\mathcal{E}$.

Another remarkable property of conditional probability, which is also a distinctive aspect of probabilistic reasoning, is its non-monotonic property. Given a non conditional probability $\text{Prob}\{\mathcal{E}_1\} > 0$ a priori, we cannot say anything about the conditional term $\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\}$. This term can be larger, equal or smaller than $\text{Prob}\{\mathcal{E}_1\}$. For instance if observing the event $\mathcal{E}_2$ makes the event more (less) probable then $\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} > \text{Prob}\{\mathcal{E}_1\}$ ($\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} < \text{Prob}\{\mathcal{E}_1\}$). If the two events are independent, then the probability of $\mathcal{E}_1$ does not change by conditioning. It follows that the degree of belief of an event (or statement) depends on the context. Note that this does not apply to conventional logical reasoning where the validity of a statement is context-independent.

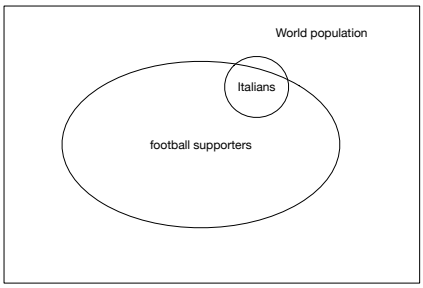


Figure 3.6: Italians and football fans.

In more general terms, it is possible to say that any probability statement is conditional since it has been formulated on the basis of an often implicit background knowledge $\mathcal{K}$. For instance, if we say that the probability of the event $\mathcal{E}$ ="rain tomorrow" is $\text{Prob}\{\mathcal{E}\} = 0.9$, we are implicitly taking into consideration the season, our location and probably the weather today. So we should better note it as $\text{Prob}\{\mathcal{E}|\mathcal{K}\} = 0.9$. As succinctly stated in [27] *all probabilities are conditional, and conditional probabilities are probabilities.*

Exercise

Consider as sample space Ω the set of all human beings. Let us define two events: the set $\mathcal{E}_1$ of Italians and the set $\mathcal{E}_2$ of football supporters. Suppose that the probability of the two events is proportional to the surface of the regions in Figure 3.6. Are these events disjoint? Are they independent? What about $\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\}$ and $\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}$? Are they equal? If not, which one is the largest?

•

The prosecutor fallacy

Consider the following story: *A crime occurs in a big city (1M of inhabitants), and a deteriorated DNA trace of the murderer is collected. The DNA profile matches the profile of a person in a police database. A geneticist is contacted, and she states that the probability of finding a person with the same DNA profile is one out of 100 thousand (i.e.* $1e-5$*). The prosecution lawyer asks for condemnation with the following argument: "since the chance of finding an innocent man with such characteristics is so tiny, then the probability that he is innocent will be tiny as well". The jury is impressed and ready to proceed with a life sentence. Then the defendant replies: "Do you know that the population of the city is 1M? So the average number of persons matching such DNA profile is 10. His chance of being innocent is not so tiny since it is 9/10 and not one in 100000" Lacking any additional evidence, the suspect is acquitted.*

This short story is inspired by a number of real cases in court that were confronted with the serious error of confounding direct and inverse conditional probability [167]. The impact of such false reasoning is so relevant in law that it is known as the *Prosecutor's fallacy*, a common default in reasoning when the collected evidence is tiny if the accused was innocent.

Let us analyse in probabilistic terms the fallacious reasoning that occurred in the example above. Let consider a criminal case for which we have 10 suspects, i.e. the responsible and 9 innocent persons (out of a 1 million population) matching the DNA profile. The probability of matching evidence (M) given that someone is innocent (I) is very low

$$\text{Prob}\{M|I\} = \frac{9}{999999} \approx 1e-5$$

However, what is relevant here is not the probability of the evidence given that he is innocent ($\text{Prob}\{M|I\}$) but the probability that is innocent given the evidence

$$\text{Prob}\{I|M\} = \frac{\text{Prob}\{M|I\}\,\text{Prob}\{I\}}{\text{Prob}\{M\}} = \frac{9/999999 \times 999999/1000000}{10/1000000} = 9/10.$$

We can rephrase the issue in the following frequentist terms. Given N inhabitants, m persons with DNA matching profiles and a single murderer, the following table shows the distribution of persons

	Match	No match
Innocent	$m-1$	$N-m$
Guilty	1	0

From the table above, it is easy to derive the inconsistency of the prosecutor fallacy reasoning since

$$\text{Prob}\{M|I\} = \frac{m-1}{N-1} \approx \text{Prob}\{M\} = \frac{m}{N}$$
$$\text{Prob}\{I|M\} = \frac{m-1}{m} >> \text{Prob}\{M|I\}$$

•

3.1.10 Logics and probabilistic reasoning

This section aims to present some interesting relationships between logic deduction and probabilistic reasoning.

First, we show that we can write down a probabilistic version of the deductive *modus ponens* rule of propositional logic (Section 2.1):

If $\mathcal{E}_1 \Rightarrow \mathcal{E}_2$ and $\mathcal{E}_1$ is true, then $\mathcal{E}_2$ is true as well.

Since $\mathcal{E}_1 \Rightarrow \mathcal{E}_2$ is equivalent in set terms to $\mathcal{E}_1 \subset \mathcal{E}_2$ we obtain

$$\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} = \frac{\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2\}}{\text{Prob}\{\mathcal{E}_1\}} = \frac{\text{Prob}\{\mathcal{E}_1\}}{\text{Prob}\{\mathcal{E}_1\}} = 1$$

i.e. a translation of the *modus ponens* argument in the probabilistic language. Interestingly enough, the probability theory provides us with a result also in the case of true $\mathcal{E}_2$. It is well-known that in propositional logic if $\mathcal{E}_1 \Rightarrow \mathcal{E}_2$ and $\mathcal{E}_2$ is true, then nothing can be inferred about $\mathcal{E}_1$. Probability theory is more informative since in this case we may derive from $\mathcal{E}_2 \subset \mathcal{E}_1$ that

$$\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = \frac{\text{Prob}\{\mathcal{E}_1\}}{\text{Prob}\{\mathcal{E}_2\}} \geq \text{Prob}\{\mathcal{E}_1\}$$

Note that this is a probabilistic formulation of the abduction principle (Section 2.3). In other words, probability supports the following common-sense reasoning: if both $\mathcal{E}_1 \Rightarrow \mathcal{E}_2$ and $\mathcal{E}_2$ apply, then the conditional probability of $\mathcal{E}_1$ (i.e. the probability of $\mathcal{E}_1$ once we know that $\mathcal{E}_2$ occurred) cannot be smaller than the unconditional probability (i.e. the probability of $\mathcal{E}_1$ if we knew nothing about $\mathcal{E}_2$).

Also the properties of transitivity and inverse *modus ponens* hold in probability. Let us consider three events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$. The transitivity principle in logics states that

$$\textit{If } \mathcal{E}_1 \Rightarrow \mathcal{E}_2 \textit{ and } \mathcal{E}_2 \Rightarrow \mathcal{E}_3 \textit{ then } \mathcal{E}_1 \Rightarrow \mathcal{E}_3$$

In probabilistic terms we can rewrite $\mathcal{E}_1 \Rightarrow \mathcal{E}_2$ as

$$\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} = 1$$

and $\mathcal{E}_2 \Rightarrow \mathcal{E}_3$ as

$$\text{Prob}\{\mathcal{E}_3|\mathcal{E}_2\} = 1$$

respectively. From the law of total probability (Equation (3.1.18)) we obtain

$$\text{Prob}\{\mathcal{E}_3|\mathcal{E}_1\} = \text{Prob}\{\mathcal{E}_3|\mathcal{E}_2,\mathcal{E}_1\}\underbrace{\text{Prob}\{\mathcal{E}_2^c|\mathcal{E}_1\}}_{0} + \underbrace{\text{Prob}\{\mathcal{E}_3|\mathcal{E}_2,\mathcal{E}_1\}}_{1}\underbrace{\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}}_{1} = 1$$

Inverse *modus ponens* in logics states that

$$\textit{If } \mathcal{E}_1 \Rightarrow \mathcal{E}_2 \textit{ then } \neg\mathcal{E}_2 \Rightarrow \neg\mathcal{E}_1$$

In probabilistic terms from $\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\} = 1$ it follows

$$\text{Prob}\{\mathcal{E}_1^c|\mathcal{E}_2^c\} = 1 - \text{Prob}\{\mathcal{E}_1|\mathcal{E}_2^c\} = 1 - \frac{\underbrace{\text{Prob}\{\mathcal{E}_2^c|\mathcal{E}_1\}}_{0}\text{Prob}\{\mathcal{E}_1\}}{\text{Prob}\{\mathcal{E}_2^c\}} = 1$$

Those results show that deductive logic rules can be seen as limiting cases of probabilistic reasoning and confirm the compatibility of probability reasoning with human common sense.

3.1.11 Combined experiments

So far we assumed that all the events belong to the same sample space. However, the most interesting use of probability concerns *combined* (or *multivariate*) random experiments whose sample space

$$\Omega = \Omega_1 \times \Omega_2 \times \ldots \Omega_n$$

is the Cartesian product of the spaces Ω_i, $i = 1, \ldots, n$. For instance, if we want to study the probabilistic dependence between the height and the weight of a child we define a joint sample space

$$\Omega = \{(w, h) : w \in \Omega^w, h \in \Omega^h\}$$

made of all pairs (w, h) where Ω^w is the sample space of the random experiment describing the weight and Ω^h is the sample space of the random experiment describing the height.

Note that all the properties studied so far also holds for events that do not belong to the same univariate sample space. For instance, given a combined experiment $\Omega = \Omega_1 \times \Omega_2$ two events $\mathcal{E}_1 \in \Omega_1$ and $\mathcal{E}_2 \in \Omega_2$ are independent iff $\text{Prob}\{\mathcal{E}_1|\mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\}$.

Some examples of real problems modelled by random combined experiments are presented in the following.

Gambler's fallacy

Consider a fair coin-tossing game. The outcome of two consecutive tosses can be considered independent. Now, suppose that we observe a sequence of 10 consecutive tails. We could be tempted to think that the chances that the next toss will be head are now very large. This is known as the *gambler's fallacy* [176]. In fact, to witness a very rare event (like 10 consecutive tails) does not imply that the probability of the next event will change or rather that it will become suddenly dependent on the past.

•

Example [188]

Let us consider a medical study about the relationship between the outcome of a medical test and the presence of a disease. We model this study as a combination of two random experiments:

1. the random experiment which models the state of the patient. Its sample space is $\Omega^s = \{H, S\}$ where H and S stand for a healthy and a sick patient, respectively.
2. the random experiment which models the outcome of the medical test. Its sample space is $\Omega^o = \{+, -\}$ where $+$ and $-$ stand for a positive and a negative outcome of the test, respectively.

The dependency between the state of the patient and the outcome of the test can be studied in terms of conditional probability.

Suppose that out of 1000 patients, 108 respond positively to the test and that among them, 9 result to be affected by the disease. Also, among the 892 patients who responded negatively to the test, only 1 is sick. According to the frequentist interpretation, the probabilities of the joint events Prob $\{\mathcal{E}^s, \mathcal{E}^o\}$ can be approximated according to expression (3.1.5) by

	$\mathcal{E}^s = S$	$\mathcal{E}^s = H$
$\mathcal{E}^o = +$	$\frac{9}{1000} = .009$	$\frac{108-9}{1000} = .099$
$\mathcal{E}^o = -$	$\frac{1}{1000} = .001$	$\frac{892-1}{1000} = .891$

Doctors are interested in answering the following questions. What is the probability of having a positive (negative) test outcome when the patient is sick (healthy)? What is the probability of being in front of a sick (healthy) patient when a positive (negative) outcome is obtained? From the definition of conditional probability we derive

$$\text{Prob}\{\mathcal{E}^o = +|\mathcal{E}^s = S\} = \frac{\text{Prob}\{\mathcal{E}^o = +, \mathcal{E}^s = S\}}{\text{Prob}\{\mathcal{E}^s = S\}} = \frac{.009}{.009 + .001} = .9$$

$$\text{Prob}\{\mathcal{E}^o = -|\mathcal{E}^s = H\} = \frac{\text{Prob}\{\mathcal{E}^o = -, \mathcal{E}^s = H\}}{\text{Prob}\{\mathcal{E}^s = H\}} = \frac{.891}{.891 + .099} = .9$$

According to these figures, the test appears to be accurate. Does this mean that we should be scared if we test positive? Though the test is accurate, the answer is negative, as shown by the quantity

$$\text{Prob}\{\mathcal{E}^s = S|\mathcal{E}^o = +\} = \frac{\text{Prob}\{\mathcal{E}^o = +, \mathcal{E}^s = S\}}{\text{Prob}\{\mathcal{E}^o = +\}} = \frac{.009}{.009 + .099} \approx .08$$

This example confirms that sometimes humans tend to confound Prob $\{\mathcal{E}^s|\mathcal{E}^o\}$ with Prob $\{\mathcal{E}^o|\mathcal{E}^s\}$ and that the most intuitive response is not always the right one (see example in Section 3.1.9).

•

3.1.12 Array of joint/marginal probabilities

Let us consider the combination of two random experiments whose sample spaces are $\Omega^A = \{A_1, \cdots, A_n\}$ and $\Omega^B = \{B_1, \cdots, B_m\}$, respectively. Assume that for each pair of events (A_i, B_j), $i = 1, \ldots, n$, $j = 1, \ldots, m$ we know the joint probability value $\text{Prob}\{A_i, B_j\}$. The joint probability array contains all the necessary information for computing *all* marginal and conditional probabilities by means of (3.1.18) and (3.1.8).

	B_1	B_2	$\cdots$	B_m	Marginal
A_1	$\text{Prob}\{A_1, B_1\}$	$\text{Prob}\{A_1, B_2\}$	$\cdots$	$\text{Prob}\{A_1, B_m\}$	$\text{Prob}\{A_1\}$
A_2	$\text{Prob}\{A_2, B_1\}$	$\text{Prob}\{A_2, B_2\}$	$\cdots$	$\text{Prob}\{A_1, B_m\}$	$\text{Prob}\{A_2\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_n	$\text{Prob}\{A_n, B_1\}$	$\text{Prob}\{A_n, B_2\}$	$\cdots$	$\text{Prob}\{A_n, B_m\}$	$\text{Prob}\{A_n\}$
Marginal	$\text{Prob}\{B_1\}$	$\text{Prob}\{B_2\}$	$\cdots$	$\text{Prob}\{B_m\}$	Sum=1

where $\text{Prob}\{A_i\} = \sum_{j=1,\ldots,m} \text{Prob}\{A_i, B_j\}$ and $\text{Prob}\{B_j\} = \sum_{i=1,\ldots,n} \text{Prob}\{A_i, B_j\}$. Using an entry of the joint probability matrix and the sum of the corresponding row/column, we may use (3.1.8) to compute the conditional probability as shown in the following example.

Example: dependent/independent scenarios

Let us model the commute time to go back home for a ULB student living in St. Gilles as a random experiment. Suppose that its sample space is Ω^t={LOW, MEDIUM, HIGH}. Consider also an (extremely:-) random experiment representing the weather in Brussels, whose sample space is Ω^w={G=GOOD, B=BAD}. Suppose that the array of joint probabilities is

	G (in Bxl)	B (in Bxl)	Marginal
LOW	0.15	0.05	$\text{Prob}\{\text{LOW}\} = 0.2$
MEDIUM	0.1	0.4	$\text{Prob}\{\text{MEDIUM}\} = 0.5$
HIGH	0.05	0.25	$\text{Prob}\{\text{HIGH}\} = 0.3$
	$\text{Prob}\{\text{G}\} = 0.3$	$\text{Prob}\{\text{B}\} = 0.7$	Sum=1

According to the above probability function, is the commute time dependent on the weather in Bxl? Note that if weather is good

	LOW	MEDIUM	HIGH
$\text{Prob}\{\cdot\|\text{G}\}$	0.15/0.3=0.5	0.1/0.3=0.33	0.05/0.3=0.16

Else if weather is bad

	LOW	MEDIUM	HIGH
$\text{Prob}\{\cdot\|\text{B}\}$	0.05/0.7=0.07	0.4/0.7=0.57	0.25/0.7=0.35

Since $\text{Prob}\{\cdot|\text{G}\} \neq \text{Prob}\{\cdot|\text{B}\}$, i.e. the probability of having a certain commute time changes according to the value of the weather, the relation (3.1.9) is not satisfied.

Consider now the dependency between an event representing the commute time and an event describing the weather in Rome.

	G (in Rome)	B (in Rome)	Marginal
LOW	0.18	0.02	$\text{Prob}\{\text{LOW}\} = 0.2$
MEDIUM	0.45	0.05	$\text{Prob}\{\text{MEDIUM}\} = 0.5$
HIGH	0.27	0.03	$\text{Prob}\{\text{HIGH}\} = 0.3$
	$\text{Prob}\{\text{G}\} = 0.9$	$\text{Prob}\{\text{B}\} = 0.1$	Sum=1

Our question now is: is the commute time dependent on the weather in Rome?

If the weather in Rome is good we obtain

	LOW	MEDIUM	HIGH
$\text{Prob}\{\cdot\|\text{G}\}$	0.18/0.9=0.2	0.45/0.9=0.5	0.27/0.9=0.3

$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	P($\mathcal{E}_1$,$\mathcal{E}_2$,$\mathcal{E}_3$)
CLEAR	RISING	DRY	0.4
CLEAR	RISING	WET	0.07
CLEAR	FALLING	DRY	0.08
CLEAR	FALLING	WET	0.10
CLOUDY	RISING	DRY	0.09
CLOUDY	RISING	WET	0.11
CLOUDY	FALLING	DRY	0.03
CLOUDY	FALLING	WET	0.12

Table 3.1: Joint probability distribution of the three-variable probabilistic model of the weather

while if the weather in Rome is bad

	LOW	MEDIUM	HIGH
Prob $\{\cdot\|B\}$	0.02/0.1=0.2	0.05/0.1=0.5	0.03/0.1=0.3

Note that the probability of a commute time event does NOT change according to the value of the weather in Rome, e.g. Prob $\{$LOW$|$B$\}$ = Prob $\{$LOW$\}$. Try to answer now the following question. If you would like to predict the commute time in Brussels, which event would return more information on it: the weather in Rome or in Brussels?

•

Example: three sample spaces

Consider a probabilistic model of the day's weather based on the combination of the following random descriptors where

1. the first represents the sky condition and its sample space is Ω^s =$\{$CLEAR, CLOUDY$\}$.

2. the second represents the barometer trend and its sample space is Ω^b =$\{$RISING, FALLING$\}$.

3. the third represents the humidity in the afternoon and its sample space is Ω^h =$\{$DRY,WET$\}$.

Let the joint probability values be given by Table 3.1. From the joint values we can calculate the probabilities $P(CLEAR, RISING) = 0.47$ and $P(CLOUDY) = 0.35$ and the conditional probability value

$$P(DRY|CLEAR, RISING) = \frac{P(DRY, CLEAR, RISING)}{P(CLEAR, RISING)} = \frac{0.40}{0.47} \approx 0.85$$

Take the time now to compute yourself other probabilities: for instance what is the probability of having a cloudy sky in wet conditions? Does a rising barometer increase or not this probability? Is the event "clear sky and falling barometer" independent from the event "dry weather"?

•

3.2 Random variables

Machine learning and statistics are concerned with numeric data and measurements while so far we have mainly been dealing with categories. What is then the link between the notion of random experiment and data? The answer is provided by the concept of random variable.

Consider a random experiment and the associated triple $(\Omega, \{\mathcal{E}\}, \text{Prob}\{\cdot\})$. Suppose that we have a mapping rule $\mathbf{z} : \Omega \to \mathcal{Z} \subset \mathbb{R}$ such that we can associate with each experimental outcome ω a real value $z = \mathbf{z}(\omega)$ in the domain $\mathcal{Z}$. We say that z is the value taken by the random variable $\mathbf{z}$ when the outcome of the random experiment is ω. Henceforth, in order to clarify the distinction between a random variable and its value, we will use the boldface notation for denoting a random variable (as in $\mathbf{z}$) and the normal face notation for the eventually observed value (as in $z = 11$).

Since there is a probability associated with each event $\mathcal{E}$ and we have a mapping from events to real values, a probability distribution can be associated with $\mathbf{z}$.

Definition 2.1 (Random variable)**.** Given a random experiment $(\Omega, \{\mathcal{E}\}, \text{Prob}\{\cdot\})$, a random variable $\mathbf{z}$ is the result of a mapping $\mathbf{z} : \Omega \to \mathcal{Z}$ that assigns a number z to every outcome ω. This mapping must satisfy the following two conditions:

- the set $\{\mathbf{z} \leq z\}$ is an event for every z.
- the probabilities

$$\text{Prob}\{\mathbf{z} = \infty\} = 0 \qquad \text{Prob}\{\mathbf{z} = -\infty\} = 0$$

Given a random variable $\mathbf{z} \in \mathcal{Z}$ and a subset $I \subset \mathcal{Z}$ we define the inverse mapping

$$\mathbf{z}^{-1}(I) = \{\omega \in \Omega | \mathbf{z}(\omega) \in I\} \tag{3.2.24}$$

where $\mathbf{z}^{-1}(I) \in \{\mathcal{E}\}$ is an event. On the basis of the above relation we can associate a probability measure to $\mathbf{z}$ according to

$$\text{Prob}\{\mathbf{z} \in I\} = \text{Prob}\left\{\mathbf{z}^{-1}(I)\right\} = \text{Prob}\{\omega \in \Omega | \mathbf{z}(\omega) \in I\} \tag{3.2.25}$$

$$\text{Prob}\{\mathbf{z} = z\} = \text{Prob}\left\{\mathbf{z}^{-1}(z)\right\} = \text{Prob}\{\omega \in \Omega | \mathbf{z}(\omega) = z\} \tag{3.2.26}$$

In other words, a *random variable* is a numerical quantity, linked to some experiment involving some degree of randomness, which takes its value from some set $\mathcal{Z}$ of possible real values. The notion of r.v. formalizes the notion of numeric measurements, which is indeed a mapping between an event (e.g. your body temperature) and a number (e.g. in the range $\mathcal{Z} = \{35, \ldots, 41\}$ returned by the thermometer). Another experiment might be the rolling of two six-sided dice and the r.v. $\mathbf{z}$ might be the sum (or the maximum) of the two numbers showing in the dice. In this case, the set of possible values is $\mathcal{Z} = \{2, \ldots, 12\}$ (or $\mathcal{Z} = \{1, \ldots, 6\}$).

Example

Suppose that we have to decide when to go home and watch Fiorentina AC playing the Champion's League final match against Anderlecht. In order to make such a decision, a quantity of interest is the (random) commute time $\mathbf{z}$ for getting from ULB to home. Our personal experience is that this time is a positive number that is not constant: for example, $z_1 = 10$ minutes, $z_2 = 23$ minutes, $z_3 = 17$ minutes, where z_i is the time taken on the ith day of the week. The variability of this quantity is related to a complex random process with a large sample space Ω (depending, for example, on the weather condition, the weekday, the sports events in town, and so on). The probabilistic approach uses a random variable to represent this uncertainty

and considers each measure z_i as the consequence of a random outcome ω_i. The use of a random variable $\mathbf{z}$ to represent the commute time becomes then a *compact (and approximate)* way of modelling the disparate set of causes underlying the uncertainty of this phenomenon. Whatever its limits, the probabilistic representation provides us with a computational way to decide when to leave if we want to bound the probability of missing the start of the game.

•

3.3 Discrete random variables

The *probability (mass) function* of a discrete r.v. $\mathbf{z}$ is the combination of

1. the countable set $\mathcal{Z}$ of values that the r.v. can take (also called *range*),
2. the set of probabilities associated to each value of $\mathcal{Z}$.

This means that we can attach to the random variable some specific mathematical function $P_{\mathbf{z}}(z)$ that gives for each $z \in \mathcal{Z}$ the probability that $\mathbf{z}$ assumes the value z

$$P_{\mathbf{z}}(z) = \text{Prob}\{\mathbf{z} = z\} \tag{3.3.27}$$

This function is called *probability function* or *probability mass function*. Note that henceafter will use $P(z)$ as a shorthand for $\text{Prob}\{\mathbf{z} = z\}$ when the identity of the random variable is clear from the context.

As depicted in the following example, the probability function can be tabulated for a few sample values of $\mathbf{z}$. If we toss a fair coin twice, and the random variable $\mathbf{z}$ is the number of heads that eventually turn up, the probability function can be tabulated as follows

Values of the random variable $\mathbf{z}$	0	1	2
Associated probabilities	0.25	0.50	0.25

3.3.1 Parametric probability function

Sometimes the probability function is not precisely known but can be expressed as a function of z and a quantity θ. An example is the discrete r.v. $\mathbf{z}$ that takes its value from $\mathcal{Z} = \{1, 2, 3\}$ and whose probability function is

$$P_{\mathbf{z}}(z, \theta) = \frac{\theta^{2z}}{\theta^2 + \theta^4 + \theta^6}$$

where θ is some fixed nonzero real number.

Whatever the value of θ, $P_{\mathbf{z}}(z) > 0$ for $z = 1, 2, 3$ and $P_{\mathbf{z}}(1) + P_{\mathbf{z}}(2) + P_{\mathbf{z}}(3) = 1$. Therefore $\mathbf{z}$ is a well-defined random variable, even if the value of θ is unknown. We call θ a *parameter*, that is some constant, usually unknown, involved in the analytical expression of a probability function. We will see in the following that the parametric form is a convenient way to formalise a family of probabilistic models and that the problem of estimation can be seen as a parameter identification task.

3.3.2 Expected value, variance and standard deviation of a discrete r.v.

Though the probability function $P_{\mathbf{z}}$ provides a complete description of the uncertainty of $\mathbf{z}$, it is often not practical to use since this requires to keep in mind (or in memory) as many values as the size of $\mathcal{Z}$. Therefore, it is more convenient to deal with some compact representation of $P_{\mathbf{z}}$ obtained by computing a functional (i.e.

a function of a function) of $P_{\mathbf{z}}$. The most common single-number summary of the distribution $P_{\mathbf{z}}$ is the expected value which is a measure of central tendency[4].

Definition 3.1 (Expected value)**.** The *expected value* of a discrete random variable $\mathbf{z}$ is

$$E[\mathbf{z}] = \mu = \sum_{z \in \mathcal{Z}} z P_{\mathbf{z}}(z) \tag{3.3.28}$$

assuming that the sum is well-defined.

An interesting property of the expected value is that it is the value which minimizes the squared deviation

$$\mu = \arg\min_{m} E[(\mathbf{z} - m)^2] \tag{3.3.29}$$

Note that the expected value is not necessarily a value that belongs to the domain $\mathcal{Z}$ of the random variable. It is important also to remark that while the term *mean* is used as a synonym of *expected value*, this is not the case for the term *average*. We will discuss in detail the difference between mean and sample average in Section 5.3.2.

Example [176]

Let us consider a European roulette with numbers $0, 1, \ldots, 36$ and where the number 0 is considered as winning for the house. The gain of a player who places a 1$ bet on a single number is a random variable $\mathbf{z}$ whose sample space is $\mathcal{Z} = \{-1, 35\}$. In other words, only two outcomes are possible: either she wins $z_1 = -1\$$ (or better he loses 1$) with probability $p_1 = 36/37$ or he wins $z_2 = 35\$$ with probability $p_2 = 1/37$. The expected gain is then

$$E[\mathbf{z}] = p_1 z_1 + p_2 z_2 = p_1 * (-1) + p_2 * 35 = -36/37 + 35/37 = -1/37 = -0.027$$

This means that while casinos gain on average 2.7 cents for every staked dollar, players on average are giving away 2.7 cents (whatever sophisticated their betting strategy is).

•

A common way to summarise the spread of a distribution is provided by the variance.

Definition 3.2 (Variance)**.** The *variance* of a discrete random variable $\mathbf{z}$ is

$$\mathrm{Var}[\mathbf{z}] = \sigma^2 = E[(\mathbf{z} - E[\mathbf{z}])^2] = \sum_{z \in \mathcal{Z}} (z - E[\mathbf{z}])^2 P_{\mathbf{z}}(z)$$

The variance is a measure of the dispersion of the probability function of the random variable around its mean μ. Note that the following relation holds

$$\sigma^2 = E[(\mathbf{z} - E[\mathbf{z}])^2] = E[\mathbf{z}^2 - 2\mathbf{z}E[\mathbf{z}] + (E[\mathbf{z}])^2] \tag{3.3.30}$$

$$= E[\mathbf{z}^2] - (E[\mathbf{z}])^2 = E[\mathbf{z}^2] - \mu^2 \tag{3.3.31}$$

whatever is the probability function of $\mathbf{z}$. Figure 3.7 illustrate two example discrete r.v. probability functions that have the same mean but different variance. Note that the variance $\mathrm{Var}[\mathbf{z}]$ does not have the same dimension as the values of $\mathbf{z}$. For instance, if $\mathbf{z}$ is measured in the unit $[m]$, $\mathrm{Var}[\mathbf{z}]$ is expressed in the unit $[m]^2$. Standard deviation is a measure for the spread that has the same dimension as $\mathbf{z}$. An alternative measure of spread is $E[|\mathbf{z} - \mu|]$ but this quantity is less used since more difficult to be analytically manipulated than the variance.

[4]This concept was first introduced in the 17th century by C. Huygens in order to study the games of chance

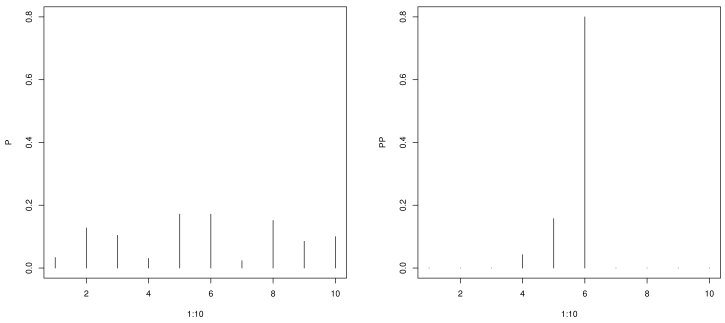


Figure 3.7: Two discrete probability functions with the same mean and different variance

Definition 3.3 (Standard deviation)**.** The *standard deviation* of a discrete random variable $\mathbf{z}$ is the positive square root of the variance.

$$\mathrm{Std}\left[\mathbf{z}\right] = \sqrt{\mathrm{Var}\left[\mathbf{z}\right]} = \sigma$$

Example

Let us consider a binary random variable $\mathbf{z} \in \mathcal{Z} = \{0, 1\}$ where $P_{\mathbf{z}}(1) = p$, $0 \leq p \leq 1$ and $P_{\mathbf{z}}(0) = 1 - p$. In this case

$$E[\mathbf{z}] = p * 1 + 0 * (1 - p) = p \tag{3.3.32}$$

$$E[\mathbf{z}^2] = p * 1 + 0 * (1 - p) = p \tag{3.3.33}$$

$$\mathrm{Var}\left[\mathbf{z}\right] = E[\mathbf{z}^2] - (E[\mathbf{z}])^2 = p - p^2 = p(1 - p) \tag{3.3.34}$$

•

Definition 3.4 (Moment)**.** For any positive integer r, the rth moment of the probability function is

$$\mu_r = E[\mathbf{z}^r] = \sum_{z \in \mathcal{Z}} z^r P_{\mathbf{z}}(z) \tag{3.3.35}$$

Note that the first moment coincides with the mean μ, while the second moment is related to the variance according to Equation (3.3.30). Higher-order moments provide additional information, other than the mean and the spread, about the shape of the probability function.

Definition 3.5 (Skewness)**.** The skewness of a discrete random variable $\mathbf{z}$ is defined as

$$\gamma = \frac{E[(\mathbf{z} - \mu)^3]}{\sigma^3} \tag{3.3.36}$$

Skewness is a parameter that describes asymmetry in a random variable's probability function. Probability functions with positive skewness have long tails to the right, and functions with negative skewness have long tails to the left (Figure 3.8).

Definition 3.6 (Kurtosis)**.** The kurtosis of a discrete random variable $\mathbf{z}$ is defined as

$$\gamma = \frac{E[(\mathbf{z} - \mu)^4]}{\sigma^4} \tag{3.3.37}$$

Kurtosis is always positive. Its interpretation is that the probability function of a distribution with large kurtosis has fatter tails, compared with the probability function of a distribution with smaller kurtosis.

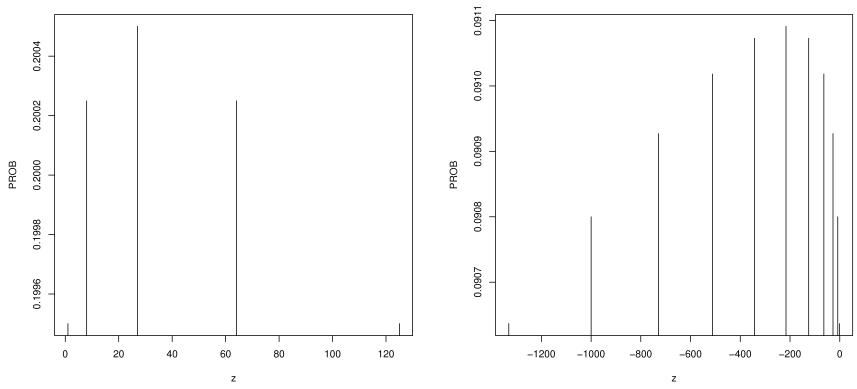


Figure 3.8: A discrete probability function with positive skewness (left) and one with a negative skewness (right).

3.3.3 Entropy and relative entropy

Definition 3.7 (Entropy)**.** Given a discrete r.v. $\mathbf{z}$, the *entropy* of the probability function $P_{\mathbf{z}}(z)$ is defined by

$$H(\mathbf{z}) = -\sum_{z \in \mathcal{Z}} P_{\mathbf{z}}(z) \log P_{\mathbf{z}}(z)$$

$H(\mathbf{z})$ is a measure of the unpredictability of the r.v. $\mathbf{z}$. Suppose that there are M possible values for the r.v. $\mathbf{z}$. The entropy is maximized (and takes the value $\log M$) if $P_{\mathbf{z}}(z) = 1/M$ for all z. It is minimized iff $P(z) = 1$ for a single value of $\mathbf{z}$ (i.e. all others probability values are null).

Although entropy measures as well as variance the uncertainty of a r.v., it differs from the variance since it depends only on the probabilities of the different values and not on the values themselves. In other terms, H can be seen as a function of the probability function $P_{\mathbf{z}}$ rather than of $\mathbf{z}$.

Let us now consider two different discrete probability functions on the same set of values

$$P_0 = P_{\mathbf{z}_0}(z), \qquad P_1 = P_{\mathbf{z}_1}(z)$$

where $P_0(z) > 0$ if and only if $P_1(z) > 0$. The *relative entropies* (or the *Kullback-Leibler divergences*) associated with these two functions are

$$H(P_0||P_1) = \sum_z P_0(z) \log \frac{P_0(z)}{P_1(z)} = \sum_z P_0(z) \log P_0(z) - \sum_z P_0(z) \log P_1(z) \tag{3.3.38}$$

$$H(P_1||P_0) = \sum_z P_1(z) \log \frac{P_1(z)}{P_0(z)} = \sum_z P_1(z) \log P_1(z) - \sum_z P_1(z) \log P_0(z) \tag{3.3.39}$$

where the term

$$-\sum_z P_0(z) \log P_1(z) = -E_{\mathbf{z}}[\log P_1] \tag{3.3.40}$$

is also called the *cross-entropy*. These asymmetric quantities measure the dissimilarity between the two probability functions. A symmetric formulation of the dissimilarity is provided by the *divergence* quantity

$$J(P_0, P_1) = H(P_0||P_1) + H(P_1||P_0).$$

3.4 Continuous random variable

An r.v. $\mathbf{z}$ is said to be a *continuous random variable* if it can assume any of the infinite values within a range of real numbers. The following quantities can be defined:

Definition 4.1 (Cumulative distribution function)**.** The *(cumulative) distribution function* of $\mathbf{z}$ is the function $F_{\mathbf{z}} : \mathbb{R} \to [0, 1]$

$$F_{\mathbf{z}}(z) = \text{Prob}\{\mathbf{z} \leq z\} \tag{3.4.41}$$

This function satisfies the following two conditions:

1. it is right-continuous: $F_{\mathbf{z}}(z) = \lim_{y \to z} F_{\mathbf{z}}(y)$,
2. it is non-decreasing: $z_1 < z_2$ implies $F_{\mathbf{z}}(z_1) \leq F_{\mathbf{z}}(z_2)$,
3. it is normalized, i.e

$$\lim_{z \to -\infty} F_{\mathbf{z}}(z) = 0, \qquad \lim_{z \to \infty} F_{\mathbf{z}}(z) = 1$$

Definition 4.2 (Density function)**.** The *density function* of a real random variable $\mathbf{z}$ is the derivative of the distribution function

$$p_{\mathbf{z}}(z) = \frac{dF_{\mathbf{z}}(z)}{dz} \tag{3.4.42}$$

at all points z where $F_{\mathbf{z}}(\cdot)$ is differentiable.

Probabilities of continuous r.v. are not allocated to specific values but rather to interval of values. Specifically

$$\text{Prob}\{a \leq z \leq b\} = \int_a^b p_{\mathbf{z}}(z)dz, \qquad \int_{\mathcal{Z}} p_{\mathbf{z}}(z)dz = 1$$

Some considerations about continuous r.v. are worthy to be mentioned:

- the quantity $\text{Prob}\{\mathbf{z} = z\} = 0$ for all z,
- the quantity $p_{\mathbf{z}}(z)$ can be bigger than one (since it is a density and not a probability) and even unbounded,
- two r.v.s $\mathbf{z}_1$ and $\mathbf{z}_2$ with the same domain $\mathcal{Z}$ are equal in distribution if $F_{\mathbf{z}_1}(z) = F_{\mathbf{z}_1}(z)$ for all $z \in \mathcal{Z}$.

Note that hence-after we will use $p(z)$ as a shorthand for $p_{\mathbf{z}}(z)$ when the identity of the random variable is clear from the context.

3.4.1 Mean, variance, moments of a continuous r.v.

Consider a continuous scalar r.v. with range $\mathcal{Z} = (l, h)$ and density function $p(z)$. We may define the following quantities.

Definition 4.3 (Expectation or mean)**.** The mean of a continuous scalar r.v. $\mathbf{z}$ is the scalar value

$$\mu = E[\mathbf{z}] = \int_l^h zp(z)dz \tag{3.4.43}$$

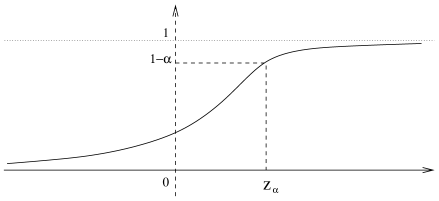


Figure 3.9: Cumulative distribution function and upper critical point.

Definition 4.4 (Variance)**.** The variance of a continuous scalar r.v. $\mathbf{z}$ is the scalar value

$$\sigma^2 = E[(\mathbf{z} - \mu)^2] = \int_l^h (z - \mu)^2 p(z) dz \tag{3.4.44}$$

Definition 4.5 (Moments)**.** The r-th moment of a continuous scalar r.v. $\mathbf{z}$ is the scalar value

$$\mu_r = E[\mathbf{z}^r] = \int_l^h z^r p(z) dz \tag{3.4.45}$$

Note that the moment of order $r = 1$ coincides with the *mean* of $\mathbf{z}$.

Definition 4.6 (Quantile function)**.** Given the cumulative function $F_{\mathbf{z}}$, the quantile (or inverse cumulative) function is the function $F_{\mathbf{z}}^{-1} : [0, 1] \to \mathbb{R}$ such that

$$F_{\mathbf{z}}^{-1}(q) = \inf\{z : F_{\mathbf{z}}(z) > q\}$$

The quantities $F_{\mathbf{z}}(1/4)$, $F_{\mathbf{z}}(1/2)$, $F_{\mathbf{z}}(3/4)$ are called the first quartile, the median and the third quartile, respectively.

Definition 4.7 (Upper critical point)**.** For a given $0 \leq \alpha \leq 1$ the *upper critical point* of a continuous r.v. $\mathbf{z}$ is the value z_α such that

$$1 - \alpha = \text{Prob}\{\mathbf{z} \leq z_\alpha\} = F(z_\alpha) \Leftrightarrow z_\alpha = F^{-1}(1 - \alpha)$$

Figure 3.9 shows an example of cumulative distribution together with the upper critical point. A compact review of univariate discrete and continuous distributions is available in Appendix C.1. In what follows we will detail only the univariate normal case.

3.4.2 Univariate Normal (or Gaussian) distribution

A continuous scalar random variable $\mathbf{x}$ is said to be *normally distributed* with parameters μ and σ^2 (also $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$) if its probability density function is Normal (or Gaussian). The analytical form of a Normal probability density function is

$$p_{\mathbf{x}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.4.46}$$

where the coefficient before the exponential ensures that $\int p_{\mathbf{x}}(x)dx = 1$. The mean of the Normal random variable $\mathbf{x}$ is μ and its variance is σ^2. An interesting property of a normal r.v.is that the probability that an observation x is within 1 (2) standard deviations from the mean is 0.68 (0.95). You may find more probabilistic relationships in Table 3.2. When $\mu = 0$ and $\sigma^2 = 1$ the distribution is called *standard normal* (Figure 3.10) and its distribution function is denoted $F_{\mathbf{z}}(z) = \Phi(z)$. All

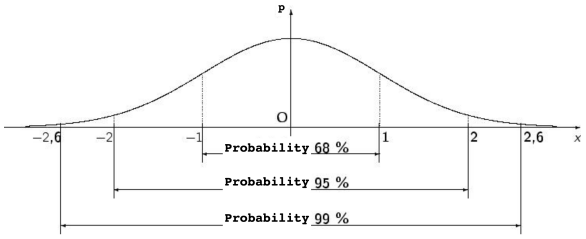


Figure 3.10: Density of a standard r.v. $\mathcal{N}(0,1)$

$\text{Prob}\{\mu - \sigma \leq \mathbf{x} \leq \mu + \sigma\} \approx 0.683$
$\text{Prob}\{\mu - 1.282\sigma \leq \mathbf{x} \leq \mu + 1.282\sigma\} \approx 0.8$
$\text{Prob}\{\mu - 1.645\sigma \leq \mathbf{x} \leq \mu + 1.645\sigma\} \approx 0.9$
$\text{Prob}\{\mu - 1.96\sigma \leq \mathbf{x} \leq \mu + 1.96\sigma\} \approx 0.95$
$\text{Prob}\{\mu - 2\sigma \leq \mathbf{x} \leq \mu + 2\sigma\} \approx 0.954$
$\text{Prob}\{\mu - 2.57\sigma \leq \mathbf{x} \leq \mu + 2.57\sigma\} \approx 0.99$
$\text{Prob}\{\mu - 3\sigma \leq \mathbf{x} \leq \mu + 3\sigma\} \approx 0.997$

Table 3.2: Some probabilistic relations holding for $\mathbf{x} \in \mathcal{N}(\mu, \sigma^2)$

random variables $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$ are linked to a standard variable $\mathbf{z}$ by the following relation

$$\mathbf{z} = (\mathbf{x} - \mu)/\sigma. \tag{3.4.47}$$

It follows that $\mathbf{z} \sim \mathcal{N}(0,1) \Rightarrow \mathbf{x} = \mu + \sigma\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$.

The practitioner might now wonder why the Normal distribution is so ubiquitous in statistics books and literature. There are plenty of reasons both from the theoretical and the practical side. From a theoretical perspective, the adoption of a Normal distribution is justified by the Central Limit theorem (Appendix C.7) which states that, under conditions almost always satisfied in practice, a linear combination of random variables converges to a Normal distribution. This is particularly useful if we wish to represent in a compact lumped form the variability that escapes to a modelling effort (e.g. the regression plus noise form in Section 10.1). Another relevant property of Gaussian distributions is that they are invariant to linear transformations, i.e. a linear transformation of a Gaussian r.v. is still Gaussian, and its mean (variance) depends on the mean (variance) of the original r.v.. From a more pragmatic perspective, an evident asset of a Gaussian representation is that only a finite number of parameters (two in the univariate case) are sufficient to characterise the entire distribution.

Exercise

Test yourself the relations in Table 3.2 by random sampling and simulation using the script `norm.R`.

•

3.5 Joint probability

So far, we considered scalar random variables only. However, the most interesting probabilistic (and machine learning) applications are multivariate, i.e. concerning a number of variables larger than one. Let us consider a probabilistic model described by n discrete random variables. A fully-specified probabilistic model gives the *joint*

probability for every combination of the values of the n r.v.s. In other terms, the joint probability contains all the information about the random variables.

In the discrete case, the model is specified by the values of the probabilities

$$\text{Prob}\{\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2, \ldots, \mathbf{z}_n = z_n\} = P(z_1, z_2, \ldots, z_n) \tag{3.5.48}$$

for every possible assignment of values $z_1, \ldots, z_n$ to the variables.

Spam mail example

Let us consider a bivariate probabilistic model describing the relation between the validity of a received email and the presence of the word *Viagra* in the text. Let $\mathbf{z}_1$ be the random variable describing the validity of the email ($\mathbf{z}_1 = 0$ for no-spam and $\mathbf{z}_1 = 1$ for spam) and $\mathbf{z}_2$ the r.v. describing the presence ($\mathbf{z}_2 = 1$) or the absence ($\mathbf{z}_2 = 0$) of the word *Viagra*. The stochastic relationship between these two variables can be defined by the joint probability distribution given by the table

	$\mathbf{z}_1 = 0$	$\mathbf{z}_1 = 1$	$P_{\mathbf{z}_2}$
$\mathbf{z}_2 = 0$	0.8	0.08	0.88
$\mathbf{z}_2 = 1$	0.01	0.11	0.12
$P_{\mathbf{z}_1}$	0.81	0.19	1

•

In the case of n continuous random variables, the model is specified by the joint distribution function

$$\text{Prob}\{\mathbf{z}_1 \leq z_1, \mathbf{z}_2 \leq z_2, \ldots, \mathbf{z}_n \leq z_n\} = F(z_1, z_2, \ldots, z_n)$$

which returns a value for every possible assignment of values $z_1, \ldots, z_n$ to the variables.

3.5.1 Marginal and conditional probability

Let $\{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be a subset of size m of the n discrete r.v.s for which a joint probability function (3.5.48) is defined. The *marginal probabilities* for the subset can be derived from expression (3.5.48) by summing over all possible combinations of values for the remaining variables.

$$P(z_1, \ldots, z_m) = \sum_{\tilde{z}_{m+1}} \cdots \sum_{\tilde{z}_n} P(z_1, \ldots, z_m, \tilde{z}_{m+1}, \ldots, \tilde{z}_n) \tag{3.5.49}$$

Exercise

Compute the marginal probabilities $P(\mathbf{z}_1 = 0)$ and $P(\mathbf{z}_1 = 1)$ from the joint probability of the spam mail example.

•

For continuous random variables the marginal density is

$$p(z_1, \ldots, z_m) = \int p(z_1, \ldots, z_m, z_{m+1}, \ldots, z_n) dz_{m+1} \ldots dz_n \tag{3.5.50}$$

This is also known as the *sum rule* or the *marginalisation property*.
The following definition for r.v. derives directly from Equation (3.1.8).

Definition 5.1 (Conditional probability function)**.** The *conditional probability function* for one subset of discrete variables $\{\mathbf{z}_i : i \in S_1\}$ given values for another disjoint subset $\{\mathbf{z}_j : j \in S_2\}$ where $S_1 \cap S_2 = \emptyset$, is defined as the ratio

$$P(\{z_i : i \in S_1\}|\{z_j : j \in S_2\}) = \frac{P(\{z_i : i \in S_1\}, \{z_j : j \in S_2\})}{P(\{z_j : j \in S_2\})}$$

Definition 5.2 (Conditional density function)**.** The *conditional density function* for one subset of continuous variables $\{\mathbf{z}_i : i \in S_1\}$ given values for another disjoint subset $\{\mathbf{z}_j : j \in S_2\}$ where $S_1 \cap S_2 = \emptyset$, is defined as the ratio

$$p(\{z_i : i \in S_1\}|\{z_j : j \in S_2\}) = \frac{p(\{z_i : i \in S_1\}, \{z_j : j \in S_2\})}{p(\{z_j : j \in S_2\})} \tag{3.5.51}$$

where $p(\{z_j : j \in S_2\})$ is the marginal density of the set S_2 of variables. When $p(\{z_j : j \in S_2\}) = 0$ this quantity is not defined.

The simplified version of (3.5.51) for two r.v.s $\mathbf{z}_1$ and $\mathbf{z}_2$ is

$$\begin{aligned} &p(\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2) = \\ &\quad = p(\mathbf{z}_2 = z_2|\mathbf{z}_1 = z_1)p(\mathbf{z}_1 = z_1) = p(\mathbf{z}_1 = z_1|\mathbf{z}_2 = z_2)p(\mathbf{z}_2 = z_2) \end{aligned} \tag{3.5.52}$$

which is also known as the *product* rule.

By combining (3.4.43), the sum rule (3.5.50) and the product rule (3.5.52) we obtain

$$p(z_1) = \int p(z_1, z_2)dz_2 = \int p(z_1|z_2)p(z_2)dz_2 = E_{\mathbf{z}_2}[p(z_1|z_2)]$$

where the subscript $\mathbf{z}_2$ makes clear that the expectation is computed with respect to the distribution of $\mathbf{z}_2$ only (while z_1 is fixed).

3.5.2 Independence

Having defined the joint and the conditional probability, we can now define when two random variables are independent.

Definition 5.3 (Independent discrete random variables)**.** Let $\mathbf{x}$ and $\mathbf{y}$ be two discrete random variables. Two variables $\mathbf{x}$ and $\mathbf{y}$ are defined to be *statistically independent* (written as $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$) if the joint probability

$$\text{Prob}\{\mathbf{x} = x, \mathbf{y} = y\} = \text{Prob}\{\mathbf{x} = x\}\,\text{Prob}\{\mathbf{y} = y\}, \qquad \forall x, y \tag{3.5.53}$$

The definition can be easily extended to the continuous case.

Definition 5.4 (Independent continuous random variables)**.** Two continuous variables $\mathbf{x}$ and $\mathbf{y}$ are defined to be *statistically independent* (written as $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$) if the joint density

$$p(\mathbf{x} = x, \mathbf{y} = y) = p(\mathbf{x} = x)p(\mathbf{y} = y), \qquad \forall x, y \tag{3.5.54}$$

From the definition of independence and conditional density it follows that

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \Leftrightarrow p(\mathbf{x} = x|\mathbf{y} = y) = p(\mathbf{x} = x) \quad \forall x, y \tag{3.5.55}$$

In layman's terms, the independence of two variables means that we do not expect that the observed outcome of one variable will affect the probability of observing the other, or equivalently that knowing something about one variable adds no information about the other. For instance, hair colour and gender are independent. Knowing someone's hair colour adds nothing to the knowledge of his gender.

Height and weight are dependent, however. Knowing someone's height does not determine precisely their weight: nevertheless, you have less uncertainty about his probable weight after you have been told the height.

Though independence is symmetric

$$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \Leftrightarrow \mathbf{y} \perp\!\!\!\perp \mathbf{x}$$

it is neither reflexive (i.e. a variable is not independent of itself) nor transitive. In other terms, if $\mathbf{x}$ and $\mathbf{y}$ are independent and $\mathbf{y}$ and $\mathbf{z}$ are independent, then $\mathbf{x}$ and $\mathbf{z}$ need not be independent.

If we consider three instead of two variables, they are said to be mutually independent if and only if each pair of rv.s. is independent and

$$p(x, y, z) = p(x)p(y)p(z)$$

Also the relationship

$$\mathbf{x} \perp\!\!\!\perp (\mathbf{y}, \mathbf{z}) \Rightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{z}, \mathbf{x} \perp\!\!\!\perp \mathbf{y}$$

holds, but not the one in the opposite direction.

Note that in mathematical terms an independence assumption implies that a bivariate density function can be written in a simple form, i.e. as the product of two univariate densities. This results in an important benefit in terms of the size of the parametrisation. For instance, consider two discrete random variables $\mathbf{z}_1 \in \mathcal{Z}_1$, $\mathbf{z}_2 \in \mathcal{Z}_2$ such that the cardinality of the two ranges is k_1 and k_2, respectively. In the generic case, if $\mathbf{z}_1$ and $\mathbf{z}_2$ are not independent, the definition of the joint probability requires the definition of $k_1 k_2 - 1$ terms[5] (or parameters). In the independent case because of the property (3.5.54), the definition requires $k_1 - 1$ terms for $\mathbf{z}_1$ and $k_2 - 1$ terms for $\mathbf{z}_2$, so overall $k_1 + k_2 - 2$. This makes a big difference in case of large values of k_1 and k_2.

Independence allows an economic parametrisation in the multivariate case as well. Consider the case of a large number n of binary discrete r.v.s., i.e. each having a range made of two values. If we need to define the joint probability, we require $2^n - 1$ terms (or parameters) in the generic case. If the n variables are independent, this number is reduced to n.

Exercise

Check whether the variable $\mathbf{z}_1$ and $\mathbf{z}_2$ of the spam mail example are independent.

•

Note that hence-after, for the sake of brevity, we will limit to introduce definitions for continuous random variables only. All of them can however be extended to the discrete case too.

3.5.3 Chain rule

Given a set of n random variables, the *chain rule* (also called the general product rule) returns the joint density as a function of conditional densities:

$$p(z_n, \ldots, z_1) = p(z_n | z_{n-1}, \ldots, z_1) p(z_{n-1} | z_{n-2}, \ldots, z_1) \ldots p(z_2 | z_1) p(z_1) \quad (3.5.56)$$

This rule is convenient to simplify the representation of large variate distributions by describing them in terms of conditional probabilities.

[5] minus one because of the normalisation constraint

3.5.4 Conditional independence

Independence is not a stable relation. Though $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$, the r.v. $\mathbf{x}$ may become dependent with $\mathbf{y}$ once we observe the value z of a third variable $\mathbf{z}$. In the same way, two dependent variables $\mathbf{x}$ and $\mathbf{y}$ may become independent once the value of $\mathbf{z}$ is known. This leads us to introduce the notion of conditional independence.

Definition 5.5 (Conditional independence)**.** Two r.v.s $\mathbf{x}$ and $\mathbf{y}$ *are conditionally independent given the value* $\mathbf{z} = z$ ($\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} = z$) iff

$$p(\mathbf{x} = x, \mathbf{y} = y|\mathbf{z} = z) = p(\mathbf{x} = x|\mathbf{z} = z)p(\mathbf{y} = y|\mathbf{z} = z) \qquad \forall x, y \tag{3.5.57}$$

Two r.v.s $\mathbf{x}$ and $\mathbf{y}$ are *conditionally independent* given $\mathbf{z}$ ($\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$) iff they are conditionally independent for all values of $\mathbf{z}$.

Since from the chain rule (3.5.56) we may write

$$p(\mathbf{x} = x, \mathbf{y} = y|\mathbf{z} = z) = p(\mathbf{x} = x|\mathbf{z} = z)p(\mathbf{y} = y|\mathbf{x} = x, \mathbf{z} = z)$$

it follows that $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} = z$ implies the relation

$$p(\mathbf{y} = y|\mathbf{x} = x, \mathbf{z} = z) = p(\mathbf{y} = y|\mathbf{z} = z) \tag{3.5.58}$$

In plain words, the notion of conditional dependence makes formal the intuition that a variable may bring (or not) information about a second one, according to the context.

Note that the statement $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} = z$ means that $\mathbf{x}$ and $\mathbf{y}$ are independent if $\mathbf{z} = z$ occurs but does not say anything about the relation between $\mathbf{x}$ and $\mathbf{y}$ if $\mathbf{z} = z$ does not occur. It could follow that two variables are independent but not conditional independent (or the other way round). In general independence does not imply conditional independence and conditional independence does not imply independence [27] (as in the example below).

Example: pizzas, dependence and conditional independence

Let $\mathbf{y}$ a variable representing the quality of a pizza restaurant and $\mathbf{x}$ a variable quantifying the Italian assonance of the restaurant name. Intuitively, you would prefer (because of higher quality $\mathbf{y}$) a pizza served in the restaurant "Sole Mio" (large $\mathbf{x}$), rather than in the restaurant "Tot Straks" (low $\mathbf{x}$). In probabilistic terms, this means that $\mathbf{x}$ and $\mathbf{y}$ are dependent ($\mathbf{x} \not\perp\!\!\!\perp \mathbf{y}$), i.e. knowing $\mathbf{x}$ reduces the uncertainty we have about $\mathbf{y}$. However, it is not the restaurant owner who makes your pizza, but the cook (*pizzaiolo*). Let $\mathbf{z}$ represent the assonance of his name. Now you would prefer eating a pizza in a Belgian restaurant where the *pizzaiolo* has Italian origins rather than in an Italian restaurant with a Flemish cook. In probabilistic terms $\mathbf{x}$ and $\mathbf{y}$ become independent once $\mathbf{z}$ (the pizzaiolo's name) is known ($\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}$).

•

It can be shown that the following two assertions are equivalent

$$(\mathbf{x} \perp\!\!\!\perp (\mathbf{z}_1, \mathbf{z}_2)|\mathbf{y}) \Leftrightarrow (\mathbf{x} \perp\!\!\!\perp \mathbf{z}_1|(\mathbf{y}, \mathbf{z}_2)), \quad (\mathbf{x} \perp\!\!\!\perp \mathbf{z}_2|(\mathbf{y}, \mathbf{z}_1))$$

Also

$$(\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z}), (\mathbf{x} \perp\!\!\!\perp \mathbf{z}|\mathbf{y}) \Rightarrow (\mathbf{x} \perp\!\!\!\perp (\mathbf{y}, \mathbf{z}))$$

If $(\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z})$, $(\mathbf{z} \perp\!\!\!\perp \mathbf{y}|\mathbf{x})$, $(\mathbf{z} \perp\!\!\!\perp \mathbf{x}|\mathbf{y})$ then $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ are mutually independent. If $\mathbf{z}$ is a random vector, the *order* of the conditional independence is equal to the number of variables in $\mathbf{z}$.

3.5.5 Entropy in the continuous case

Consider a continuous r.v. $\mathbf{y}$. The *(differential) entropy* of $\mathbf{y}$ is defined by

$$H(\mathbf{y}) = -\int \log(p(y))p(y)dy = E_{\mathbf{y}}[-\log(p(y))] = E_{\mathbf{y}}\left[\log\frac{1}{p(y)}\right]$$

with the convention that $0\log 0 = 0$. Entropy is a functional of the distribution of $\mathbf{y}$ and is a measure of the predictability of a r.v. $\mathbf{y}$. The higher the entropy, the less reliable are our predictions about $\mathbf{y}$. For a scalar normal r.v. $\mathbf{y} \sim \mathcal{N}(\mu, \sigma^2)$

$$H(\mathbf{y}) = \frac{1}{2}\left(1 + \ln 2\pi\sigma^2\right) = \frac{1}{2}\left(\ln 2\pi e\sigma^2\right) \tag{3.5.59}$$

In the case of a normal random vector $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim \mathcal{N}(0, \Sigma)$

$$H(\mathbf{Y}) = \frac{1}{2}\left(\ln(2\pi e)^n \det(\Sigma)\right)$$

3.5.5.1 Joint and conditional entropy

Consider two continuous r.v.s $\mathbf{x}$ and $\mathbf{y}$ and their joint density $p(x, y)$. The *joint entropy* of $\mathbf{x}$ and $\mathbf{y}$ is defined by

$$H(\mathbf{x}, \mathbf{y}) = -\int\int \log(p(x, y))p(x, y)dxdy = \\ = E_{\mathbf{x},\mathbf{y}}[-\log(p(x, y))] = E_{\mathbf{x},\mathbf{y}}\left[\log\frac{1}{p(x, y)}\right]$$

The *conditional entropy* is defined as

$$H(\mathbf{y}|\mathbf{x}) = -\int\int \log(p(y|x))p(x, y)dxdy = E_{\mathbf{x},\mathbf{y}}[-\log(p(y|x))] = \\ = E_{\mathbf{x},\mathbf{y}}\left[\log\frac{1}{p(y|x)}\right] = E_{\mathbf{x}}[H(\mathbf{y}|x)]$$

This quantity quantifies the remaining uncertainty of $\mathbf{y}$ once $\mathbf{x}$ is known. Note that in general $H(\mathbf{y}|\mathbf{x}) \neq H(\mathbf{x}|\mathbf{y})$, $H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y})$ and that the *chain rule* holds

$$H(\mathbf{y}, \mathbf{x}) = H(\mathbf{y}|\mathbf{x}) + H(\mathbf{x}) \tag{3.5.60}$$

Also, conditioning reduces entropy

$$H(\mathbf{y}|\mathbf{x}) \leq H(\mathbf{y})$$

with equality if $\mathbf{x}$ and $\mathbf{y}$ are independent, i.e. $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$. This property formalises a fundamental principle underlying machine learning, data science and prediction in general, i.e. that by conditioning on some variables $\mathbf{x}$ (e.g. inputs) we may reduce the uncertainty about a variable $\mathbf{y}$ (target). Another interesting property is the *independence bound*

$$H(\mathbf{y}, \mathbf{x}) \leq H(\mathbf{y}) + H(\mathbf{x})$$

with equality if $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$.

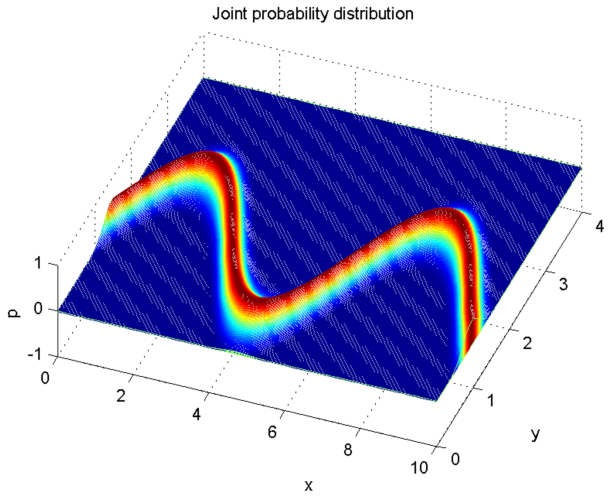


Figure 3.11: 3D visualisation of a bivariate joint density.

3.6 Bivariate continuous distribution

Let us consider two continuous r.v. $\mathbf{x}$ and $\mathbf{y}$ and their bivariate joint density function $p_{\mathbf{x},\mathbf{y}}(x,y)$. An example of bivariate joint density function is illustrated in Figure 3.11. From (3.5.50), we define *marginal density* the quantity

$$p_{\mathbf{x}}(x) = \int_{-\infty}^{\infty} p_{\mathbf{x},\mathbf{y}}(x,y)dy$$

and *conditional density* the quantity

$$p_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{p(x,y)}{p(x)} \tag{3.6.61}$$

which is, in loose terms, the probability that $\mathbf{y}$ belongs to an interval dy about y assuming that $\mathbf{x} = x$. Note that, if $\mathbf{x}$ and $\mathbf{y}$ are independent

$$p_{\mathbf{x},\mathbf{y}}(x,y) = p_{\mathbf{x}}(x)p_{\mathbf{y}}(y), \qquad p(y|x) = p_{\mathbf{y}}(y)$$

The definition of *conditional expectation* is obtained from (3.6.61) and (3.4.43).

Definition 6.1 (Conditional expectation)**.** The conditional expectation of $\mathbf{y}$ given $\mathbf{x} = x$ is

$$E_{\mathbf{y}}[\mathbf{y}|\mathbf{x} = x] = \int y p_{\mathbf{y}|\mathbf{x}}(y|x)dy = \mu_{\mathbf{y}|\mathbf{x}}(x) \tag{3.6.62}$$

From (3.3.29) we may derive that

$$E_{\mathbf{y}}[\mathbf{y}|\mathbf{x} = x] = \arg\min_{m} E_{\mathbf{y}}[(\mathbf{y} - m)^2|\mathbf{x} = x] \tag{3.6.63}$$

Note that $E_{\mathbf{y}}[\mathbf{y}|\mathbf{x} = x]$ is a function of x also known as the *regression function*.

The definition of *conditional variance* derives from (3.6.61) and (3.4.44).

Definition 6.2 (Conditional variance)**.**

$$\text{Var}\,[\mathbf{y}|\mathbf{x} = x] = \int (y - \mu_{\mathbf{y}|\mathbf{x}}(x))^2 p_{\mathbf{y}|\mathbf{x}}(y|x)dy \tag{3.6.64}$$

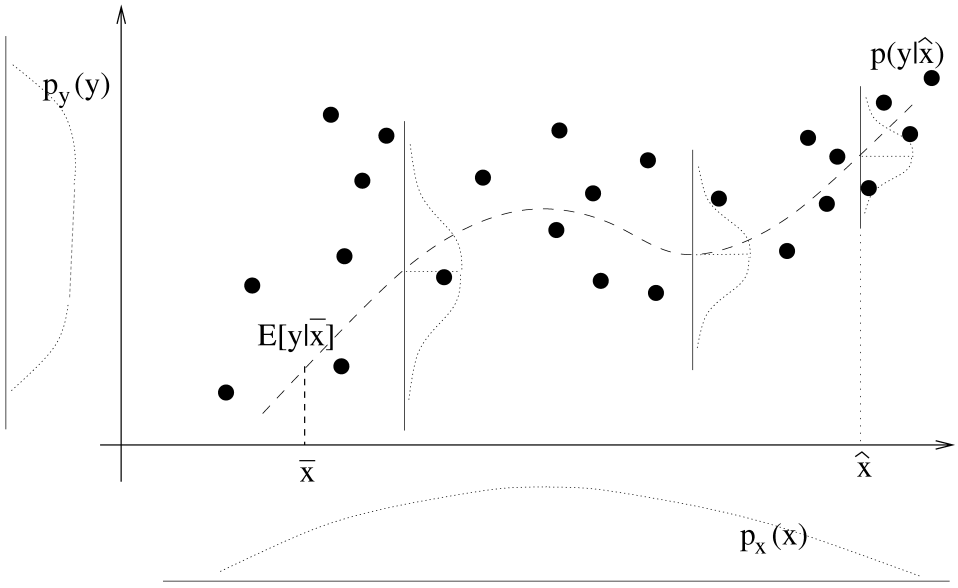


Figure 3.12: Bivariate distribution: the figure shows the two marginal distributions (beside the axis), the conditional expectation function (dashed line) and some conditional distributions (dotted).

Note that both these quantities are a function of x. If we replace the given value x by the r.v. $\mathbf{x}$ the terms $E_{\mathbf{y}}[\mathbf{y}|\mathbf{x}]$ and $\mathrm{Var}[\mathbf{y}|\mathbf{x}]$ are random, too.

Some important results on their expectation are contained in the following theorems [188].

Theorem 6.3. *For two r.v.s* $\mathbf{x}$ *and* $\mathbf{y}$*, assuming their expectations exist, we have that*

$$E_{\mathbf{x}}[E_{\mathbf{y}}[\mathbf{y}|\mathbf{x}=x]] = E_{\mathbf{y}}[\mathbf{y}] \tag{3.6.65}$$

and

$$Var[\mathbf{y}] = E_{\mathbf{x}}[\,Var[\mathbf{y}|\mathbf{x}=x]] + Var[E_{\mathbf{y}}[\mathbf{y}|\mathbf{x}=x]] \tag{3.6.66}$$

where $Var[\mathbf{y}|\mathbf{x}=x]$ *and* $E_{\mathbf{y}}[\mathbf{y}|\mathbf{x}=x]$ *are functions of* x*.*

We remind that for a bivariate function $f(x,y)$

$$E_{\mathbf{y}}[f(x,y)] = \int f(x,y)p_{\mathbf{y}}(y)dy, \quad E_{\mathbf{x}}[f(x,y)] = \int f(x,y)p_{\mathbf{x}}(x)dx.$$

A 2D representation of a bivariate continuous distribution is illustrated in Figure 3.12. It is worthy noting that, although the conditional distribution is bell-shaped, this is not necessarily the case for the marginal distributions.

3.6.1 Correlation

Consider two random variables $\mathbf{x}$ and $\mathbf{y}$ with means $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ and standard deviations $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$.

Definition 6.4 (Covariance)**.** The *covariance* between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathrm{Cov}[\mathbf{x},\mathbf{y}] = E[(\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}})] = E[\mathbf{xy}] - \mu_{\mathbf{x}}\mu_{\mathbf{y}} \tag{3.6.67}$$

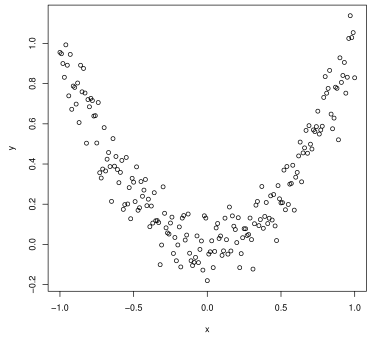


Figure 3.13: Dependent but uncorrelated random variables

A positive (negative) covariance means that the two variables are positively (inversely) related, i.e. that once one is above its mean, then the other tends to be above (below) its mean as well. The covariance can take any value in real numbers. A limitation of covariance is that it depends on variables' scales and units: for instance, if variables were measured in meters instead of centimetres, this would induce a change of their covariance. For this reason, it is common to replace covariance with correlation, a dimensionless measure of linear association.

Definition 6.5 (Correlation)**.** The *correlation coefficient* is defined as

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{Cov}[\mathbf{x}, \mathbf{y}]}{\sqrt{\mathrm{Var}\,[\mathbf{x}]\,\mathrm{Var}\,[\mathbf{y}]}} \tag{3.6.68}$$

It is easily shown that $-1 \leq \rho(\mathbf{x}, \mathbf{y}) \leq 1$. For this reason, the correlation is sometimes expressed as a percentage.

Definition 6.6 (Uncorrelated variables)**.** Two r.v.s $\mathbf{x}$ and $\mathbf{y}$ are said to be *uncorrelated* if $\rho(\mathbf{x}, \mathbf{y}) = 0$ or equivalently if

$$E[\mathbf{xy}] = E[\mathbf{x}]E[\mathbf{y}] \tag{3.6.69}$$

Note that if $\mathbf{x}$ and $\mathbf{y}$ are two independent random variables, then

$$E[\mathbf{xy}] = \int xyp(x,y)dxdy = \int xyp(x)p(y)dxdy = \int xp(x)dx \int yp(y)dy = E[\mathbf{x}]E[\mathbf{y}]$$

This means that independence implies uncorrelation. However, the contrary does not hold for a generic distribution. The equivalence between independence and uncorrelation

$$\rho(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{y} \tag{3.6.70}$$

holds only if $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian.

See Figure 3.13 for an example of uncorrelated but dependent variables.

Exercises

1. Let $\mathbf{x}$ and $\mathbf{y}$ two discrete independent r.v. such that

$$P_{\mathbf{x}}(-1) = 0.1, \quad P_{\mathbf{x}}(0) = 0.8, \quad P_{\mathbf{x}}(1) = 0.1$$

and

$$P_{\mathbf{y}}(1) = 0.1, \quad P_{\mathbf{y}}(2) = 0.8, \quad P_{\mathbf{y}}(3) = 0.1$$

If $\mathbf{z} = \mathbf{x} + \mathbf{y}$ show that $E[\mathbf{z}] = E[\mathbf{x}] + E[\mathbf{y}]$

2. Let $\mathbf{x}$ be a discrete r.v. which assumes $\{-1, 0, 1\}$ with probability $1/3$ and $\mathbf{y} = \mathbf{x}^2$. Let $\mathbf{z} = \mathbf{x} + \mathbf{y}$. Show that
   - $E[\mathbf{z}] = E[\mathbf{x}] + E[\mathbf{y}]$.
   - $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated but dependent random variables.

•

3.7 Normal distribution: the multivariate case

Let $\mathbf{z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^T$ be a $[n, 1]$ random vector . The vector is said to be *normally distributed* with parameters μ and Σ (also $\mathbf{z} \sim \mathcal{N}(\mu, \Sigma)$) if its probability density function is given by

$$p_{\mathbf{z}}(z) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)\right\} \tag{3.7.71}$$

where $\det(\Sigma)$ denotes the determinant of the matrix Σ. It follows that

- the mean $E[\mathbf{z}] = \mu$ is an $[n, 1]$ vector,
- the matrix
$$\Sigma = E[(\mathbf{z}-\mu)(\mathbf{z}-\mu)^T] \tag{3.7.72}$$
is the $[n, n]$ covariance matrix. This matrix is symmetric and positive semidefinite. It has $n(n+1)/2$ parameters: the diagonal terms Σ_{jj} are the variances $\mathrm{Var}[\mathbf{z}_j]$ of the vector components and the off-diagonal terms $\Sigma_{jk}, j \neq k$ are the covariance terms $\mathrm{Cov}[\mathbf{z}_j, \mathbf{z}_k]$. The inverse Σ^{-1} is also called the *concentration matrix*.

The quantity
$$\Delta = (z-\mu)^T \Sigma^{-1}(z-\mu) \tag{3.7.73}$$
which appears in the exponent of $p_{\mathbf{z}}$ is called the *Mahalanobis distance* from z to μ. It can be shown that the n-dimensional surfaces of constant probability density

- are hyper-ellipsoids on which Δ^2 is constant;
- their *principal axes* are given by the eigenvectors u_j, $j = 1, \ldots, n$ of Σ which satisfy
$$\Sigma u_j = \lambda_j u_j \qquad j = 1, \ldots, n$$
where λ_j are the corresponding eigenvalues.
- the eigenvalues λ_j give the variances along the principal directions (Figure 3.14).

If the covariance matrix Σ is *diagonal* then

- the contours of constant density are hyper-ellipsoids with the principal directions aligned with the coordinate axes.
- the components of $\mathbf{z}$ are then *statistically independent* since the distribution of $\mathbf{z}$ can be written as the product of the distributions for each of the components separately in the form
$$p_{\mathbf{z}}(z) = \prod_{j=1}^{n} p_{\mathbf{z}_j}(z_j)$$
- the total number of independent parameters in the distribution is $2n$ (n for the mean vector and n for the diagonal covariance matrix).
- if $\sigma_j = \sigma$ for all j, the contours of constant density are hyper-spheres.

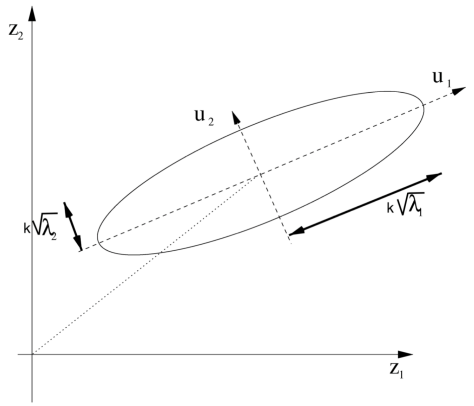


Figure 3.14: Contour curves of normal distribution for $n = 2$.

3.7.1 Bivariate normal distribution

Let us consider a bivariate ($n = 2$) normal density whose mean is $\mu = [\mu_1, \mu_2]^T$ and the covariance matrix is

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

The correlation coefficient is

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

It can be shown that the general bivariate normal density has the form

$$p(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{z_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{z_1-\mu_1}{\sigma_1}\right)\left(\frac{z_2-\mu_2}{\sigma_2}\right) + \left(\frac{z_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

A plot of a bivariate normal density with $\mu = [0, 0]$ and $\Sigma = [1.2919, 0.4546; 0.4546, 1.7081]$ and a corresponding contour curve are traced in Figure 3.15 by means of the script `gaussXYZ.R`.

We suggest the reader to play with the Shiny dashboard `gaussian.R` in order to visualize the impact of the parameters on the Gaussian distribution.

One of the important properties of the multivariate normal density is that all conditional and marginal probabilities are also normal. Using the relation

$$p(z_2|z_1) = \frac{p(z_1, z_2)}{p(z_1)}$$

we find that $p(z_2|z_1)$ is a normal distribution $\mathcal{N}(\mu_{2|1}, \sigma_{2|1}^2)$, where

$$\mu_{2|1} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(z_1 - \mu_1)$$
$$\sigma_{2|1}^2 = \sigma_2^2(1 - \rho^2)$$

Note that

- $\mu_{2|1}$ is a linear function of z_1: if the correlation coefficient ρ is positive, the larger z_1, the larger $\mu_{2|1}$.
- if there is no correlation between z_1 and z_2, the two variables are independent, i.e. we can ignore the value of z_1 to estimate μ_2.

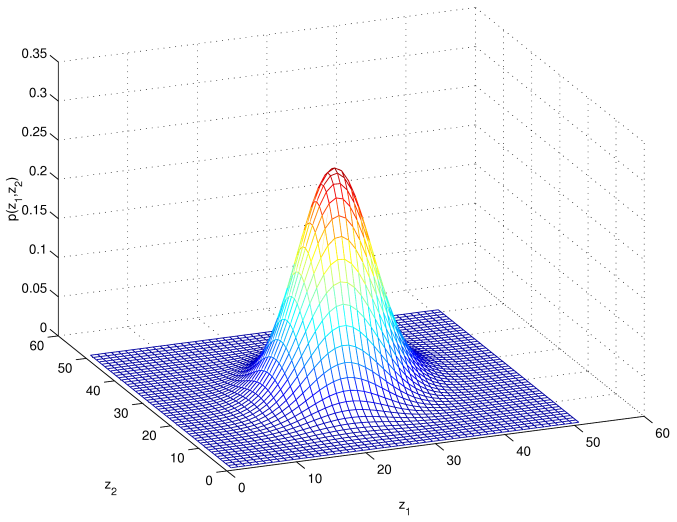


Figure 3.15: Bivariate normal density function

3.7.2 Gaussian mixture distribution

A continuous r.v. z has a Gaussian mixture distribution with m components if

$$p(\mathbf{z} = z) = \sum_{k=1}^{m} w_k \mathcal{N}(z; \mu_k, \Sigma_k) \tag{3.7.74}$$

where $\mathcal{N}(z; \mu_k, \Sigma_k)$ denotes the Normal density with mean μ_k and covariance Σ_k, and the mixture weights w_k satisfy

$$\sum_{k=1}^{m} w_k = 1, \quad 0 \leq w_k \leq 1$$

A Gaussian mixture is a linear superposition of m Gaussian components and, as such, has a higher expressive power than a unimodal Gaussian distribution: for instance, it can be used to model multimodal density distributions.

The script `gmm.R` samples a bidimensional mixture of Gaussians with 3 components with diagonal covariances. The density and the sampled points are in Figure 3.16. An interesting property of Gaussian mixtures is that they are *universal approximator* of densities which means that any smooth density can be approximated with any specific nonzero amount of error by a Gaussian mixture model (GMM) with enough components.

3.7.3 Linear transformations of Gaussian variables

If $\mathbf{z}_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbf{z}_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$ are independent Gaussian r.v.s., then the sum $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$ is a Gaussian r.v. $\mathbf{z} \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.

Given two real constants c_1 and c_2, the linear combination $\mathbf{z} = c_1\mathbf{z}_1 + c_2\mathbf{z}_2$ is a Gaussian r.v. $\mathbf{z} \sim \mathcal{N}(c_1\mu_1 + c_2\mu_2, c_1^2\Sigma_1 + c_2^2\Sigma_2)$.

If $\mathbf{z} \sim \mathcal{N}(\mu, \Sigma)$ is a $[n, 1]$ Gaussian random vector and $\mathbf{y} = A\mathbf{z}$, with A a $[n, n]$ real matrix, then $\mathbf{y} \sim \mathcal{N}(A\mu, A\Sigma A^T)$ is a Gaussian vector.

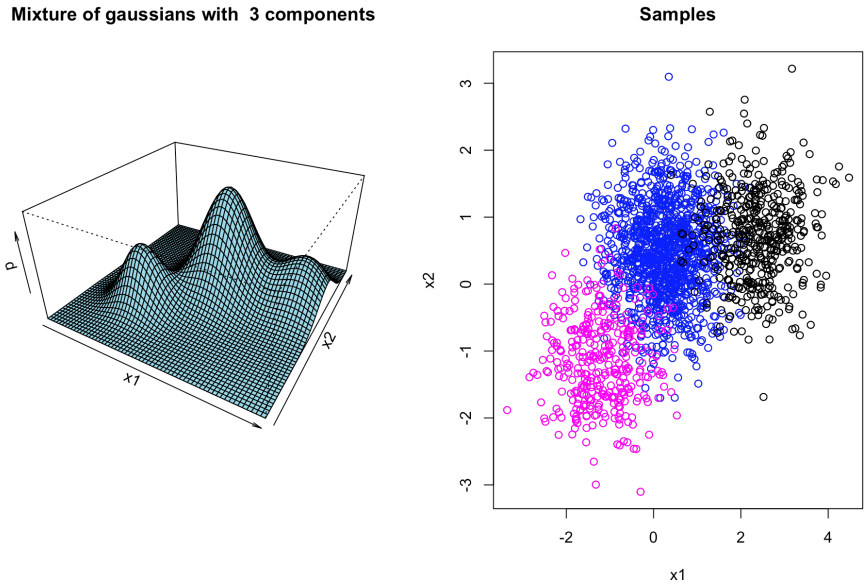


Figure 3.16: Density and observations of a bidimensional mixture of Gaussians with 3 components. Each colour corresponds to a different component.

3.8 Mutual information

Mutual information is one of the most widely used measures to convey the dependency of variables. It is a measure of the amount of information that one random variable contains about another random variable. It can also be considered as the *distance from independence* between the two variables. This quantity is always non-negative and zero if and only if the two variables are stochastically independent.

Given two random variables $\mathbf{x}$ and $\mathbf{y}$, their *mutual information* is defined in terms of their probabilistic marginal density functions $p_{\mathbf{x}}(x)$, $p_{\mathbf{y}}(y)$ and the joint $p_{(\mathbf{x},\mathbf{y})}(x,y)$:

$$I(\mathbf{x};\mathbf{y}) = \int\int \log \frac{p(x,y)}{p(x)p(y)} p(x,y)dxdy = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) \tag{3.8.75}$$

with the convention that $0 \log \frac{0}{0} = 0$. From (3.5.60), we derive

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) = H(\mathbf{y}) + H(\mathbf{x}) - H(\mathbf{x},\mathbf{y}) \tag{3.8.76}$$

Mutual information is null if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent, i.e.

$$I(\mathbf{x};\mathbf{y}) = 0 \Leftrightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{y}. \tag{3.8.77}$$

In other words, the larger the mutual information term, the stronger is the degree of dependency between two variables.

In the Gaussian case, an analytical link between correlation and mutual information exists. Let $(\mathbf{x},\mathbf{y})$ a normally distributed random vector with a correlation coefficient ρ. The mutual information between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$I(\mathbf{x};\mathbf{y}) = -\frac{1}{2}\log(1-\rho^2)$$

Equivalently the correlation coefficient (3.6.68) can be written as

$$\rho = \sqrt{1 - \exp(-2I(\mathbf{x};\mathbf{y}))}$$

In agreement with (3.8.77) and (3.6.70), it follows that in the Gaussian case

$$\rho(\mathbf{x},\mathbf{y}) = 0 \Leftrightarrow I(\mathbf{x};\mathbf{y}) = 0 \tag{3.8.78}$$

3.8.1 Conditional mutual information

Consider three r.v.s $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. The *conditional mutual information* is defined by

$$I(\mathbf{y};\mathbf{x}|\mathbf{z}) = H(\mathbf{y}|\mathbf{z}) - H(\mathbf{y}|\mathbf{x},\mathbf{z}) \tag{3.8.79}$$

It can also be written as

$$I(\mathbf{y};\mathbf{x}|\mathbf{z}) = \int\int \log \frac{p(x,y|z)}{p(x|z)p(y|z)} p(x,y,z)dxdydz$$

While mutual information quantifies the degree of (in)dependence between two variables, conditional mutual information quantifies the degree of conditional (in)dependence (Section 3.5.4) between three variables. The conditional mutual information is null iff $\mathbf{x}$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}$, i.e.

$$I(\mathbf{x};\mathbf{y}|\mathbf{z}) = 0 \Leftrightarrow \mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} \tag{3.8.80}$$

Note that $I(\mathbf{x};\mathbf{y}|\mathbf{z})$ can be null though $I(\mathbf{x};\mathbf{y}) > 0$, like in the pizzas example in Section 3.5.4. Also a symmetric configuration is possible, e.g. $I(\mathbf{x};\mathbf{y}) = 0$ but $I(\mathbf{x};\mathbf{y}|\mathbf{z}) > 0$ as in the case of complementary variables which will be discussed in Section 12.8.

3.8.2 Joint mutual information

This section derives the information of a pair of variables $(\mathbf{x}_1, \mathbf{x}_2)$ about a third one $\mathbf{y}$.

From (3.8.79) and (3.5.60) it follows:

$$\begin{aligned} I(\mathbf{x};\mathbf{y}|\mathbf{z}) = H(\mathbf{y}|\mathbf{z}) - H(\mathbf{y}|\mathbf{x},\mathbf{z}) = H(\mathbf{y}|\mathbf{z}) + H(\mathbf{x}|\mathbf{z}) - H\left((\mathbf{x},\mathbf{y})|\mathbf{z}\right) = \\ = H\left((\mathbf{x},\mathbf{z})\right) + H\left((\mathbf{y},\mathbf{z})\right) - H(\mathbf{z}) - H\left((\mathbf{x},\mathbf{y},\mathbf{z})\right) \end{aligned} \tag{3.8.81}$$

From (3.8.76) it follows

$$I((\mathbf{x}_1,\mathbf{x}_2);\mathbf{y}) = H(\mathbf{x}_1,\mathbf{x}_2) + H(\mathbf{y}) - H(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})$$

and

$$I(\mathbf{x}_1;\mathbf{y}) = H(\mathbf{x}_1) + H(\mathbf{y}) - H(\mathbf{x}_1,\mathbf{y})$$

From (3.8.81) it follows

$$\begin{aligned} I(\mathbf{x}_2;\mathbf{y}|\mathbf{x}_1) = H(\mathbf{y}|\mathbf{x}_1) - H(\mathbf{y}|\mathbf{x}_1,\mathbf{x}_2) = \\ = H(\mathbf{y},\mathbf{x}_1) - H(\mathbf{x}_1) - H(\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) + H(\mathbf{x}_1,\mathbf{x}_2) \end{aligned}$$

On the basis of the results above, we derive the *chain rule* of mutual information

$$\begin{aligned} &I(\mathbf{x}_1;\mathbf{y}) + I(\mathbf{x}_2;\mathbf{y}|\mathbf{x}_1) = \\ &= H(\mathbf{x}_1) + H(\mathbf{y}) - H(\mathbf{x}_1,\mathbf{y}) + H(\mathbf{y},\mathbf{x}_1) - H(\mathbf{x}_1) - H(\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) + H(\mathbf{x}_1,\mathbf{x}_2) = \\ &= H(\mathbf{y}) - H(\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) + H(\mathbf{x}_1,\mathbf{x}_2) = I((\mathbf{x}_1,\mathbf{x}_2);\mathbf{y}) \end{aligned} \tag{3.8.82}$$

This formula shows that the information that a pair of variables $(\mathbf{x}_1, \mathbf{x}_2)$ brings about a third variable $\mathbf{y}$ is not simply the sum of the two mutual information terms $I(\mathbf{x}_1; \mathbf{y})$ and $I(\mathbf{x}_2; \mathbf{y})$ but is the sum of $I(\mathbf{x}_1; \mathbf{y})$ and the conditional information of $\mathbf{x}_2$ and $\mathbf{y}$ given $\mathbf{x}_1$. This aspect is particularly important in the feature selection context (Section 12.8) where simplistic assumptions of monotonicity and additivity do not hold.

For $n > 2$ variables $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ the chain rule formulation is

$$I(\mathbf{X}; \mathbf{y}) = I(\mathbf{X}_{-i}; \mathbf{y}|\mathbf{x}_i) + I(\mathbf{x}_i; \mathbf{y}) = I(\mathbf{x}_i; \mathbf{y}|\mathbf{X}_{-i}) + I(\mathbf{X}_{-i}; \mathbf{y}) \quad i = 1, \ldots, n \tag{3.8.83}$$

where $\mathbf{X}_{-i}$ denote the $\mathbf{X}$ set with the ith term set aside.

3.8.3 Partial correlation coefficient

We have seen in Section 3.6.1 that correlation is a good measure of independence in the case of Gaussian distributions. The same role for conditional independence is played by partial correlation.

Definition 8.1 (First-order partial correlation)**.** Let us consider three r.v.s $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. The first-order partial correlation is

$$\rho_{\mathbf{xy}|\mathbf{z}} = \frac{\rho_{\mathbf{xy}} - \rho_{\mathbf{xz}}\rho_{\mathbf{zy}}}{\sqrt{(1 - \rho^2_{\mathbf{xz}})(1 - \rho^2_{\mathbf{yz}})}}$$

where $\rho_{\mathbf{xy}}$ is defined in (3.6.68).

This quantity returns a measure of the correlation between $\mathbf{x}$ and $\mathbf{y}$ once the value of $\mathbf{z}$ is known. It is possible to extend the partial correlation to the conditioning of two variables.

Definition 8.2 (Second-order correlation)**.**

$$\rho_{\mathbf{x}_1\mathbf{y}|\mathbf{z}\mathbf{x}_2} = \frac{\rho_{\mathbf{x}_1\mathbf{y}|\mathbf{z}} - \rho_{\mathbf{x}_1\mathbf{x}_2|\mathbf{z}}\rho_{\mathbf{y}\mathbf{x}_2|\mathbf{z}}}{\sqrt{(1 - \rho^2_{\mathbf{x}_1\mathbf{x}_2|\mathbf{z}})(1 - \rho^2_{\mathbf{y}\mathbf{x}_2|\mathbf{z}})}}$$

This can be used also to define a recurrence relationship where qth order partial correlations can be computed from $(q-1)$th order partial correlations.

Another interesting property is the link between partial correlation and concentration matrix (Section 3.7). Let Σ and $\Omega = \Sigma^{-1}$ denote the covariance and the concentration matrix of the normal set of variables $\mathbf{Z} \cup \{\mathbf{x}, \mathbf{y}\}$. The partial correlation coefficient $\rho_{\mathbf{xy}|\mathbf{Z}}$ can be obtained by matrix inversion:

$$\rho_{\mathbf{xy}|\mathbf{Z}} = \frac{-\omega_{\mathbf{xy}}}{\sqrt{\omega_{\mathbf{xx}}\omega_{\mathbf{xy}}}}$$

where $\omega_{\mathbf{xy}}$ is the element of the concentration matrix corresponding tp $\mathbf{x}$ and $\mathbf{y}$.

Consider a multivariate normal vector $\mathbf{X}$, such that $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$, $\mathbf{X}_S \subset \mathbf{X}$ and s is the dimension of $\mathbf{X}_S$. Then

$$\rho_{\mathbf{x}_i\mathbf{x}_j|\mathbf{X}_S} = 0 \Leftrightarrow I(\mathbf{x}_i, \mathbf{x}_j|\mathbf{X}_S) = 0$$

Note that this is the conditional version of the relation (3.8.78).

3.9 Functions of random variables and Monte Carlo simulation

For any function $g(\cdot)$ of the random variable $\mathbf{z}$

$$E[g(\mathbf{z})] = \int g(z)p_{\mathbf{z}}(z)dz \tag{3.9.84}$$

This is also know as the law of the unconscious statistician (LOTUS). Note that in general $E[g(\mathbf{z})] \neq g(E[\mathbf{z}])$, with the exception of the linear function $g(z) = az + b$ which will be discussed in the following section.

Exercise

Let $\mathbf{z}$ be a scalar r.v. and

$$g(z) = \begin{cases} 1 & z \in [a, b] \\ 0 & \text{else} \end{cases}$$

with $a < b$. Compute $E[g(\mathbf{z})]$.

•

For a generic g, the analytical computation or numerical integration of (3.9.84) may be extremely complex. A numerical alternative is represented by the Monte Carlo simulation which requires a pseudo-random generator of examples according to the distribution of $\mathbf{z}$. In a nutshell Monte Carlo computes $E[g(x)]$ by

1. generating a large number S of sample points $z_i \sim F_{\mathbf{z}}, i = 1, \ldots, S$,
2. computing $g(z_i)$,
3. returning the estimation

$$E[g(\mathbf{z})] \approx \frac{\sum_{i=1}^{S} g(z_i)}{S}$$

If S is sufficiently large, we may consider such approximation as reliable. The same procedure may be used to approximate other parameters of the distribution (e.g. the variance). In this book, we will have recourse to Monte Carlo simulation to provide a numerical illustration of probabilistic formulas or concepts (e.g. bias, variance and generalisation error), which otherwise might appear too abstract for the reader.

Monte Carlo computation

The script `mcarlo.R` contains the Monte Carlo computation of the mean and variance of $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ as well as the computation of $E[\mathbf{z}^2]$ and $E[|\mathbf{z}|]$.

The Shiny dashboard `mcarlo.R` visualises the result of some operations on a single and two random variables by using a Monte Carlo simulation.

•

3.10 Linear combinations of r.v.

The expected value of a linear combination of r.v.s is simply the linear combination of their respective expectation values

$$E[a\mathbf{x} + b\mathbf{y}] = aE[\mathbf{x}] + bE[\mathbf{y}], \qquad a \in \mathbb{R}, b \in \mathbb{R}$$

i.e., expectation is a linear statistic. On the contrary, the variance is not a linear statistic. We have

$$\begin{aligned} \text{Var}\left[a\mathbf{x} + b\mathbf{y}\right] &= a^2\text{Var}\left[\mathbf{x}\right] + b^2\text{Var}\left[\mathbf{y}\right] + 2ab\left(E[\mathbf{xy}] - E[\mathbf{x}]E[\mathbf{y}]\right) && (3.10.85)\\ &= a^2\text{Var}\left[\mathbf{x}\right] + b^2\text{Var}\left[\mathbf{y}\right] + 2ab\text{Cov}[\mathbf{x}, \mathbf{y}] && (3.10.86)\end{aligned}$$

where the quantity $\text{Cov}[\mathbf{x}, \mathbf{y}]$ is defined in (3.6.67).

Given n r.v. $\mathbf{z}_j, j = 1, \ldots, n$

$$\text{Var}\left[\sum_{j=1}^{n} c_j \mathbf{z}_j\right] = \sum_{j=1}^{n} c_j^2 \text{Var}\left[\mathbf{z}_j\right] + 2\sum_{i<j} c_i c_j \text{Cov}[\mathbf{z}_i, \mathbf{z}_j] \tag{3.10.87}$$

Let us consider now n random variables with the same variance σ^2 and mutual correlation ρ. Then the variance of their average is

$$\begin{aligned} \text{Var}\left[\frac{\sum_{j=1}^{n} \mathbf{z}_j}{n}\right] &= \frac{n\sigma^2}{n^2} + 2\frac{1}{n^2}\frac{n(n-1)}{2}\rho\sigma^2 = \\ &= \frac{\sigma^2}{n} + \rho\sigma^2 - \frac{\rho\sigma^2}{n} = (1-\rho)\frac{\sigma^2}{n} + \rho\sigma^2 \qquad (3.10.88)\end{aligned}$$

3.10.1 The sum of i.i.d. random variables

Suppose that $\mathbf{z}_1$,$\mathbf{z}_2$,...,$\mathbf{z}_N$ are i.i.d. (identically and independently distributed) random variables, discrete or continuous, each having a probability distribution with mean μ and variance σ^2. Let us consider the two derived r.v., that is the sum

$$\mathbf{S}_N = \mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_N$$

and the average

$$\bar{\mathbf{z}} = \frac{\mathbf{z}_1 + \mathbf{z}_2 + \cdots + \mathbf{z}_N}{N} \tag{3.10.89}$$

The following relations hold

$$E[\mathbf{S}_N] = N\mu, \qquad \text{Var}\left[\mathbf{S}_N\right] = N\sigma^2 \tag{3.10.90}$$

$$E[\bar{\mathbf{z}}] = \mu, \qquad \text{Var}\left[\bar{\mathbf{z}}\right] = \frac{\sigma^2}{N} \tag{3.10.91}$$

An illustration of these relations by simulation can be obtained by running the R script `sum_rv.R`.

3.11 Conclusion

The reader (in particular, if practitioner) might think that a chapter on probability theory is an unnecessary frill in a book on machine learning. The author has a different opinion. Probability extends the logical formalism and makes formal human patterns of reasoning under uncertainty (e.g. abduction). Also, probability provides an effective language to formalise the task of machine learning, i.e. using

some variables (e.g. inputs) to explain, provide information (or reduce uncertainty) about other ones (e.g. targets). According to Aristotles, *philosophy begins with wonder*. From a scientific perspective, wonder originates from uncertainty, and science has the role of reducing it by explanation. The author hopes that this chapter showed that uncertainty and information are not only philosophical concepts but quantities whose nature and relationship can be described in probabilistic terms.

So far, we only considered low variate settings, although the ambition of statistical machine learning is attacking complex high variate problems. For this reason, the next chapter will provide a probabilistic formalism to deal with high variate (and then complex) settings. What is still missing for the moment is the second major ingredient (besides uncertainty) of machine learning: data. Please be patient: the relation between uncertainty and observations will be discussed in Chapter 5, which introduces estimation as the statistical way of combining probabilistic models with real-world data.

3.12 Exercises

1. Suppose you collect a dataset about spam in emails. Let the binary variables $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ represent the occurrence of the words "Viagra", "Lottery" and "Won", respectively, in a email. Let the dataset of 20 emails being summarised as follows

Document	x_1 (Viagra)	x_2 (Lottery)	x_3 (Won)	y (Class)
E1	0	0	0	NOSPAM
E2	0	1	1	SPAM
E3	0	0	1	NOSPAM
E4	0	1	1	SPAM
E5	1	0	0	SPAM
E6	1	1	1	SPAM
E7	0	0	1	NOSPAM
E8	0	1	1	SPAM
E9	0	0	0	NOSPAM
E10	0	1	1	SPAM
E11	1	0	0	NOSPAM
E12	0	1	1	SPAM
E13	0	0	0	NOSPAM
E14	0	1	1	SPAM
E15	0	0	1	NOSPAM
E16	0	1	1	SPAM
E17	1	0	0	SPAM
E18	1	1	1	SPAM
E19	0	0	1	NOSPAM
E20	0	1	1	SPAM

   where

   - 0 stands for the case-insensitive absence of the word in the email.
   - 1 stands for the case-insensitive presence of the word in the email.

   Let $\mathbf{y} = 1$ denote a spam email and $\mathbf{y} = 0$ a no-spam email.

   The student should estimate on the basis of the frequency of the data above

   - $\text{Prob}\{\mathbf{x}_1 = 1, \mathbf{x}_2 = 1\}$
   - $\text{Prob}\{\mathbf{y} = 0 | \mathbf{x}_2 = 1, \mathbf{x}_3 = 1\}$
   - $\text{Prob}\{\mathbf{x}_1 = 0 | \mathbf{x}_2 = 1\}$
   - $\text{Prob}\{\mathbf{x}_3 = 1 | \mathbf{y} = 0, \mathbf{x}_2 = 0\}$
   - $\text{Prob}\{\mathbf{y} = 0 | \mathbf{x}_1 = 0, \mathbf{x}_2 = 0, \mathbf{x}_3 = 0\}$

- Prob $\{\mathbf{x}_1 = 0|\mathbf{y} = 0\}$
- Prob $\{\mathbf{y} = 0\}$

Solution:

- Prob $\{\mathbf{x}_1 = 1, \mathbf{x}_2 = 1\} = 0.1$
- Prob $\{\mathbf{y} = 0|\mathbf{x}_2 = 1, \mathbf{x}_3 = 1\} = 0$
- Prob $\{\mathbf{x}_1 = 0|\mathbf{x}_2 = 1\} = 0.8$
- Prob $\{\mathbf{x}_3 = 1|\mathbf{y} = 0, \mathbf{x}_2 = 0\} = 0.5$
- Prob $\{\mathbf{y} = 0|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0, \mathbf{x}_3 = 0\} = 1$
- Prob $\{\mathbf{x}_1 = 0|\mathbf{y} = 0\} = 0.875$
- Prob $\{\mathbf{y} = 0\} = 0.4$

2. Let us consider a fraud detection problem. Suppose we collect the following transactional dataset where $\mathbf{v} = 1$ means that the transaction came from a suspicious web site and $\mathbf{f} = 1$ means that the transaction is fraudulent.

	$\mathbf{f} = 1$	$\mathbf{f} = 0$
$\mathbf{v} = 1$	500	1000
$\mathbf{v} = 0$	1	10000

Estimate the following quantities by using the frequency as estimator of probability:

- Prob $\{\mathbf{f} = 1\}$
- Prob $\{\mathbf{v} = 0\}$
- Prob $\{\mathbf{f} = 1|\mathbf{v} = 1\}$
- Prob $\{\mathbf{v} = 1|\mathbf{f} = 1\}$

Use the Bayes theorem to compute Prob $\{\mathbf{v} = 1|\mathbf{f} = 1\}$ and show that the result is identical to the one computed before.

Solution:

- Prob $\{\mathbf{f} = 1\} = 501/11501 = 0.043$
- Prob $\{\mathbf{v} = 0\} = 10001/11501 = 0.869$
- Prob $\{\mathbf{f} = 1|\mathbf{v} = 1\} = 500/1500 = 1/3$
- Prob $\{\mathbf{v} = 1|\mathbf{f} = 1\} = 500/501$

By Bayes theorem: Prob $\{\mathbf{v} = 1|\mathbf{f} = 1\} = \frac{\text{Prob}_{\{\mathbf{f}=1|\mathbf{v}=1\}}\text{Prob}_{\{\mathbf{v}=1\}}}{\text{Prob}_{\{\mathbf{f}=1\}}} = \frac{1/3(1500/11501)}{501/11501} = 500/501$

3. Let us consider a dataset with 4 binary variables

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{y}$
1	1	0	1
0	0	1	0
0	1	0	0
1	1	1	1
0	0	0	0
0	1	0	0
0	1	1	0
0	0	1	0
0	0	0	0
0	1	0	0
1	1	1	1

Estimate the following quantities by using the frequency as estimator of probability

- Prob $\{\mathbf{y} = 1\}$
- Prob $\{\mathbf{y} = 1|\mathbf{x}_1 = 0\}$
- Prob $\{\mathbf{y} = 1|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0, \mathbf{x}_3 = 0\}$

Solution:

- Prob $\{\mathbf{y} = 1\} = 3/11$
- Prob $\{\mathbf{y} = 1|\mathbf{x}_1 = 0\} = 0$
- Prob $\{\mathbf{y} = 1|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0, \mathbf{x}_3 = 0\} = 0$

4. Let us consider a task with three binary inputs and one binary target where the input distribution is

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$P(x_1, x_2, x_3)$
0	0	0	0.2
0	0	1	0.1
0	1	0	0.1
0	1	1	0.1
1	0	0	0.1
1	0	1	0.1
1	1	0	0.1
1	1	1	0.2

and the conditional probability is

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$P(\mathbf{y} = 1|x_1, x_2, x_3)$
0	0	0	0.8
0	0	1	0.1
0	1	0	0.5
0	1	1	0.9
1	0	0	0.05
1	0	1	0.1
1	1	0	0.05
1	1	1	0.5

Compute

- Prob $\{\mathbf{x}_1 = 1, \mathbf{x}_2 = 1\}$
- Prob $\{\mathbf{y} = 0|\mathbf{x}_2 = 1, \mathbf{x}_3 = 0\}$
- Prob $\{\mathbf{x}_1 = 0|\mathbf{x}_2 = 1\}$
- Prob $\{\mathbf{x}_3 = 1|\mathbf{y} = 0, \mathbf{x}_2 = 1\}$
- Prob $\{\mathbf{y} = 0|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0, \mathbf{x}_3 = 0\}$
- Prob $\{\mathbf{x}_1 = 0|\mathbf{y} = 0\}$

Solution:

- Prob $\{\mathbf{x}_1 = 1, \mathbf{x}_2 = 1\}$=0.1+0.2=0.3
- Prob $\{\mathbf{y} = 0|\mathbf{x}_2 = 1, \mathbf{x}_3 = 0\}$ = Prob $\{\mathbf{y} = 0|\mathbf{x}_1 = 0, \mathbf{x}_2 = 1, \mathbf{x}_3 = 0\}$*Prob $\{\mathbf{x}_1 = 0|\mathbf{x}_2 = 1, \mathbf{x}_3 = 0\}$+ Prob $\{\mathbf{y} = 0|\mathbf{x}_1 = 1, \mathbf{x}_2 = 1, \mathbf{x}_3 = 0\}$*Prob $\{\mathbf{x}_1 = 1|\mathbf{x}_2 = 1, \mathbf{x}_3 = 0\} = 0.5*0.5+$ $0.95 * 0.5 = 0.725$
- Prob $\{\mathbf{x}_1 = 0|\mathbf{x}_2 = 1\} = (0.1 + 0.1)/(0.2 + 0.3) = 0.4$
- From the joint four variate distribution computed in the exercise below

$$\text{Prob}\{\mathbf{x}_3 = 1|\mathbf{y} = 0, \mathbf{x}_2 = 1\} = \frac{\text{Prob}\{\mathbf{x}_3 = 1, \mathbf{y} = 0, \mathbf{x}_2 = 1\}}{\text{Prob}\{\mathbf{y} = 0, \mathbf{x}_2 = 1\}} = \frac{0.11}{0.255} = 0.4313725$$

- Prob $\{\mathbf{y} = 0|\mathbf{x}_1 = 0, \mathbf{x}_2 = 0, \mathbf{x}_3 = 0\} = 1 - 0.8 = 0.2$

- From the joint four variate distribution computed in the exercise below

$$\text{Prob}\{\mathbf{x}_1 = 0|\mathbf{y} = 0\} = \frac{\text{Prob}\{\mathbf{x}_1 = 0, \mathbf{y} = 0\}}{\text{Prob}\{\mathbf{y} = 0\}} = \frac{0.19}{0.57} = 0.3333$$

5. Consider the probability distribution of the previous exercise. Is $\mathbf{y}$ conditionally independent of $\mathbf{x}_1$ given $\mathbf{x}_2$?

Solution:

According to Section 3.5.4, $\mathbf{y}$ is conditionally independent of $\mathbf{x}_1$ given $\mathbf{x}_2$ if for all values x_2:

$$\text{Prob}\{\mathbf{y} = y|\mathbf{x}_1 = x_1, \mathbf{x}_2 = x_2\} = \text{Prob}\{\mathbf{y} = y|\mathbf{x}_2 = x_2\}$$

Let us compute $\text{Prob}\{\mathbf{y} = 1|\mathbf{x}_1 = 1, \mathbf{x}_2 = x_2\}$ and $\text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = x_2\}$ for $\mathbf{x}_2 = 0$. From (3.1.21)

$$\begin{aligned}
&\text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1\} = \\
&\sum_{x_3} \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1, \mathbf{x}_3 = x_3\} \text{Prob}\{\mathbf{x}_3 = x_3|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1\} = \\
&= \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1\mathbf{x}_3 = 0\} \text{Prob}\{\mathbf{x}_3 = 0|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1\} + \\
&+ \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1, \mathbf{x}_3 = 1\} \text{Prob}\{\mathbf{x}_3 = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1\} = \\
&= 0.05 * 0.1/0.2 + 0.1 * 0.1/0.2 = 0.075
\end{aligned}$$

and

$$\begin{aligned}
&\text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0\} = \\
&= \sum_{x_1, x_3} \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = x_1, \mathbf{x}_3 = x_3\} \text{Prob}\{\mathbf{x}_1 = x_1, \mathbf{x}_3 = x_3|\mathbf{x}_2 = 0\} = \\
&= \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 0, \mathbf{x}_3 = 0\} \text{Prob}\{\mathbf{x}_1 = 0, \mathbf{x}_3 = 0|\mathbf{x}_2 = 0\} + \\
&+ \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 0, \mathbf{x}_3 = 1\} \text{Prob}\{\mathbf{x}_1 = 0, \mathbf{x}_3 = 1|\mathbf{x}_2 = 0\} + \\
&+ \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1, \mathbf{x}_3 = 0\} \text{Prob}\{\mathbf{x}_1 = 1, \mathbf{x}_3 = 0|\mathbf{x}_2 = 0\} + \\
&+ \text{Prob}\{\mathbf{y} = 1|\mathbf{x}_2 = 0, \mathbf{x}_1 = 1, \mathbf{x}_3 = 1\} \text{Prob}\{\mathbf{x}_1 = 1, \mathbf{x}_3 = 1|\mathbf{x}_2 = 0\} = \\
&= 0.8 * 0.2/0.5 + 0.1 * 0.1/0.5 + 0.05 * 0.1/0.5 + 0.1 * 0.1/0.5 = 0.37
\end{aligned}$$

Since those two values are different, the two variables are not conditionally independent.

An alternative would be first computing the joint distribution of the 4 variables and then deriving the conditional terms. Since

$$\text{Prob}\{y, x_1, x_2, x_3\} = \text{Prob}\{y|x_1, x_2, x_3\} \text{Prob}\{x_1, x_2, x_3\}$$

the joint distribution is :

$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$P(y, x_1, x_2, x_3)$
0	0	0	0	(1-0.8)*0.2=0.04
0	0	0	1	(1-0.1)*0.1=0.09
0	0	1	0	0.05
0	0	1	1	0.01
0	1	0	0	0.095
0	1	0	1	0.09
0	1	1	0	0.095
0	1	1	1	0.1
1	0	0	0	0.8*0.2=0.16
1	0	0	1	0.1*0.1=0.01
1	0	1	0	0.05
1	0	1	1	0.09
1	1	0	0	0.005
1	1	0	1	0.01
1	1	1	0	0.005
1	1	1	1	0.1

From the table above we compute the conditional terms as

$$\text{Prob}\{\mathbf{y}=1|\mathbf{x}_2=0\} = \frac{\text{Prob}\{\mathbf{y}=1,\mathbf{x}_2=0\}}{\text{Prob}\{\mathbf{x}_2=0\}} = \\ = \frac{0.16+0.01+0.01+0.005}{0.04+0.09+0.095+0.09+0.16+0.01+0.005+0.01} = 0.37$$

and

$$\text{Prob}\{\mathbf{y}=1|\mathbf{x}_2=0,\mathbf{x}_1=1\} = \frac{\text{Prob}\{\mathbf{y}=1,\mathbf{x}_1=1,\mathbf{x}_2=0\}}{\text{Prob}\{\mathbf{x}_1=1,\mathbf{x}_2=0\}} = \\ = \frac{0.005+0.01}{0.095+0.09+0.005+0.01} = 0.075$$

Since the results are (obviously) identical to the ones obtained with the first method, the conclusion is the same, i.e. the variables are conditionally dependent.

6. Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be three binary random variables denoting the pathological mutation of a given gene of the father, mother and child, respectively. The values 0 and 1 stand for the absence and presence of the mutation, respectively. Suppose that

   - the two parents have the same probability 0.5 of having a pathological mutation in a given gene
   - the variables $\mathbf{x}$ and $\mathbf{y}$ are independent
   - the child may inherit the mutation according to this conditional probability table

$\text{Prob}\{\mathbf{z}=1\|\mathbf{x}=x,\mathbf{y}=y\}$	x	y
0	0	0
0.6	0	1
0.4	1	0
0.7	1	1

   1. What is the probability that the child has no mutation if both parents are not affected?
   2. What is the probability that the father had a mutated gene if the child has the mutation and the mother is not affected?
   3. What is the probability that the father has a mutated gene if the child has the mutation and the mother is affected?
   4. What is the probability that the child has the mutation if the father has none?
   5. What is the probability that the father has a mutated gene if the child has the mutation?
   6. What is the probability that the father has a mutated gene if the child has no mutation?

Solution:

Let us derive first

$$P(\mathbf{z}=1|\mathbf{y}=0) = \\ P(\mathbf{z}=1|\mathbf{y}=0,\mathbf{x}=1)P(\mathbf{x}=1|\mathbf{y}=0) + P(\mathbf{z}=1|\mathbf{y}=0,\mathbf{x}=0)P(\mathbf{x}=0|\mathbf{y}=0) = \\ = P(\mathbf{z}=1|\mathbf{y}=0,\mathbf{x}=1)P(\mathbf{x}=1) + P(\mathbf{z}=1|\mathbf{y}=0,\mathbf{x}=0)P(\mathbf{x}=0) = \\ = 0.4*0.5 + 0*0.5 = 0.2$$

$$P(\mathbf{z}=1|\mathbf{y}=1) =$$
$$P(\mathbf{z}=1|\mathbf{y}=1,\mathbf{x}=1)P(\mathbf{x}=1|\mathbf{y}=1) + P(\mathbf{z}=1|\mathbf{y}=1,\mathbf{x}=0)P(\mathbf{x}=0|\mathbf{y}=1) =$$
$$= P(\mathbf{z}=1|\mathbf{y}=1,\mathbf{x}=1)P(\mathbf{x}=1) + P(\mathbf{z}=1|\mathbf{y}=1,\mathbf{x}=0)P(\mathbf{x}=0) =$$
$$= 0.7 * 0.5 + 0.6 * 0.5 = 0.65$$

$$P(\mathbf{z}=1|\mathbf{x}=1) =$$
$$P(\mathbf{z}=1|\mathbf{x}=1,\mathbf{y}=0)P(\mathbf{y}=0|\mathbf{x}=1) + P(\mathbf{z}=1|\mathbf{y}=1,\mathbf{x}=1)P(\mathbf{y}=1|\mathbf{x}=1) =$$
$$= 0.4 * 0.5 + 0.7 * 0.5 = 0.55$$

It follows

1. $$P(\mathbf{z}=0|\mathbf{x}=0,\mathbf{y}=0) = 1$$

2. $$P(\mathbf{x}=1|\mathbf{z}=1,\mathbf{y}=0) = \frac{P(\mathbf{z}=1|\mathbf{x}=1,\mathbf{y}=0)P(\mathbf{x}=1|\mathbf{y}=0)}{P(\mathbf{z}=1|\mathbf{y}=0} = \frac{0.4 * 0.5}{0.2} = 1$$

3. $$P(\mathbf{x}=1|\mathbf{z}=1,\mathbf{y}=1) = \frac{P(\mathbf{z}=1|\mathbf{x}=1,\mathbf{y}=1)P(\mathbf{x}=1|\mathbf{y}=1)}{P(\mathbf{z}=1|\mathbf{y}=1} = \frac{0.7 * 0.5}{0.65} = 0.538$$

4. $$P(\mathbf{z}=1|\mathbf{x}=0) = P(\mathbf{z}=1|\mathbf{x}=0,\mathbf{y}=1)P(\mathbf{y}=1|\mathbf{x}=0) + P(\mathbf{z}=1|\mathbf{x}=0,\mathbf{y}=0)P(\mathbf{y}=0|\mathbf{x}=0) =$$
$$= 0.6 * 0.5 + 0 = 0.3$$

5. $$P(\mathbf{x}=1|\mathbf{z}=1) = \frac{P(\mathbf{z}=1|\mathbf{x}=1)P(\mathbf{x}=1)}{P(\mathbf{z}=1} = \frac{0.55 * 0.5}{0.55 * 0.5 + 0.3 * 0.5} = 0.647$$

6. $$P(\mathbf{x}=1|\mathbf{z}=0) = \frac{P(\mathbf{z}=0|\mathbf{x}=1)P(\mathbf{x}=1)}{P(\mathbf{z}=0} = \frac{0.45 * 0.5}{0.45 * 0.5 + 0.7 * 0.5} = 0.3913$$

Chapter 4

Graphical models

Graphical Models combine probability theory and graph theory [111, 134, 117] to deal with two pervasive issues in applied mathematics and engineering: uncertainty and complexity. In particular, they rely on the notion of conditional independence (Section 3.1.6) to simplify the representation of complex high-variate probability distributions.

4.1 Conditional independence and multivariate distributions

One of the hardest challenges for machine learning is to model large variate tasks, i.e. tasks characterised by a large number of variables. Section 3.5.2 shows that an independence assumption reduces the size of the parameter set needed to describe a probability distribution with many variables. Unfortunately, the assumption of independence is very strong and rarely met in real tasks. Nevertheless, it is realistic to assume the existence of conditional independence (Section 3.5.4) relationships in large variate settings. This assumption implies *sparseness*, which is a dependence pattern where variables tend to interact with few others. If conditional independence between some variable holds, thanks to the (3.5.56), we reduce the size of the parameter set required to describe the joint probability distribution.

Consider for instance the case of $n = 4$ binary discrete r.v.s. In the generic case, we need $2^4 - 1 = 35$ parameters to encode such probability, i.e. a quantity exponential in the number of variables. This exponential nature makes the probabilistic modelling unfeasible (i.e. too many parameters to elicit) and unmanageable (i.e. too large required memory) in case of large n.

Let us now suppose that the 4 binary r.v.s are independent: in this case since

$$P(z_4, z_3, z_2, z_1) = P(z_4)P(z_3)P(z_2)P(z_1) \tag{4.1.1}$$

only 4 parameters are necessary to describe the joint distribution. No exponential explosion of the number of required parameters happens. However, this is a very simplistic and idealised setting, which rarely occurs in real interesting problems. Moreover, if all the variables were independent, there would be no need of supervised learning and predictive modelling since no variable brings information (or reduce uncertainty) about the other.

A more realistic assumption is to consider some variables as conditionally independent of others. For instance, suppose that $\mathbf{z}_4$ is conditionally independent of $\mathbf{z}_1$ and $\mathbf{z}_2$ given $\mathbf{z}_3$ ($\mathbf{z}_4 \perp\!\!\!\perp (\mathbf{z}_1, \mathbf{z}_2)|\mathbf{z}_3$)

$$P(z_4|z_3, z_2, z_1) = P(z_4|z_3) \tag{4.1.2}$$

and $\mathbf{z}_3$ is conditionally independent of $\mathbf{z}_1$ given $\mathbf{z}_2$ ($\mathbf{z}_3 \perp\!\!\!\perp \mathbf{z}_1|\mathbf{z}_2$)

$$P(z_3|z_2, z_1) = P(z_3|z_2) \tag{4.1.3}$$

From the discrete version of (3.5.56) we can write

$$P(z_4, z_3, z_2, z_1) = P(z_4|z_3, z_2, z_1)P(z_3|z_2, z_1)P(z_2|z_1)P(z_1)$$

From the conditional independence relations (4.1.2) and (4.1.3) we obtain the simplified expression

$$P(z_4, z_3, z_2, z_1) = P(z_4|z_3)P(z_3|z_2)P(z_2|z_1)P(z_1)$$

Note that the conditional probability $P(z_j|z_i)$ for two binary r.v.s can be now encoded by a conditional table with two single parameters, e.g. $P(\mathbf{z}_j = 1|\mathbf{z}_i = 1)$ and $P(\mathbf{z}_j = 1|\mathbf{z}_i = 0)$. It follows that thanks to such assumptions, we may describe the joint probability with 7 parameters only. The useful compactness of the representation is still more striking in the case of large n, continuous variables, or discrete r.v.s with a large range of values.

The representational advantage of conditional independence relationships is evident in Bayesian Networks, a formalism characterised by a correspondence between topological properties (e.g. connectivity in a directed graph) and probabilistic ones (notably independence). This formalism allows a compact, flexible, modular (since localized and then natural for humans) representation of joint distributions.

4.2 Directed acyclic graphs

A *directed graph* $\mathcal{G}$ is a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a finite non-empty set whose elements are called *nodes*, and $\mathcal{E}$ is a set of ordered pairs of distinct elements of $\mathcal{V}$. The elements of $\mathcal{E}$ are called *edges*. A directed cycle is a path from a node to itself. A directed graph is called a directed acyclic graph (DAG) if it contains no directed cycles.

Given a DAG and two nodes $\mathbf{z}_1 \in \mathcal{V}$, and $\mathbf{z}_2 \in \mathcal{V}$,

- $\mathbf{z}_2$ is called a *parent* of $\mathbf{z}_1$ if there is an edge from $\mathbf{z}_2$ to $\mathbf{z}_1$
- $\mathbf{z}_2$ is called a *descendent* of $\mathbf{z}_1$, and $\mathbf{z}_1$ is called an *ancestor* of $\mathbf{z}_2$ if there is a directed path from $\mathbf{z}_1$ to $\mathbf{z}_2$
- $\mathbf{z}_2$ is called a *non-descendent* of $\mathbf{z}_1$ if it is not a descendent of $\mathbf{z}_1$

Note that a node is not considered a descendant of itself.

4.3 Bayesian networks

DAGs are an effective way of representing multivariate distributions where the nodes denote random variables, and the topology (notably the absence of edges) encodes conditional independence assertions (e.g. elicited from an expert in this domain). A main advantage of the approach is the notion of modularity (a complex system is made by combining simpler parts) which makes possible visual interpretability.

A Bayesian Network (BN) is a pair $(\mathcal{G}, P)$ where $\mathcal{G}$ is a Directed Acyclic Graph (DAG) (i.e. graph with no loops from a variable back to itself) and P is a joint probability distribution over $\mathbf{Z}$, which is associated with $\mathcal{G}$ by the Markov condition.

Definition 3.1 (Markov condition)**.** Given a DAG graph $\mathcal{G}$ and the associated joint probability distribution P over $\mathbf{Z}$, the Markov condition (MC) holds if every variable is independent of its graphical non-descendants conditional on its parents.

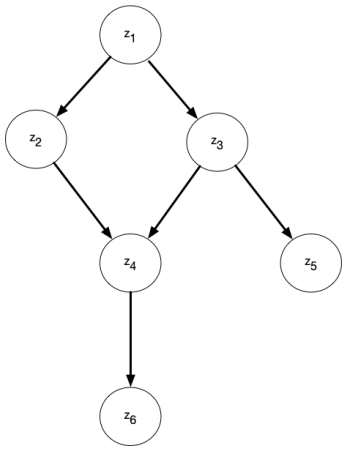


Figure 4.1: Bayesian Network.

If the Markov condition is satisfied (it is also said that $\mathcal{G}$ represents P) the following theorem holds.

Theorem 3.2. *If $(\mathcal{G}, P)$ satisfies the Markov condition, then P is equal to the product of its conditional distributions of all nodes given values of their parents, whenever these conditional distributions exist.*

This means that if we order the set of r.v. $\mathbf{z}_i$, such that if $\mathbf{z}_j$ is a descendant of $\mathbf{z}_k$, $\mathbf{z}_j$ follows $\mathbf{z}_k$ in the ordering ($k < j$), we have the *product form*

$$P(z_1, \ldots, z_n) = \prod_{i=1}^{n} P(z_i | \text{Parents}(\mathbf{z}_i))$$

where $\text{Parents}(\mathbf{z}_i)$ is the set of parents of the node $\mathbf{z}_i$ in $\mathcal{G}$.

Example

An example of BN is shown in Figure 4.1. Note that the enumeration of the variable indices satisfies the topological ordering mentioned before. Let us consider the node $\mathbf{z}_4$: the nodes $\mathbf{z}_2$ and $\mathbf{z}_3$ are its parents, $\mathbf{z}_1$ is its ancestor, and $\mathbf{z}_6$ is its descendant. The associate probability distribution may be factorised as follows:

$$P(Z) = P(z_6|z_4)P(z_5|z_3)P(z_4|z_3, z_2)P(z_3|z_1)P(z_2|z_1)P(z_1)$$

From the DAG, we can derive a number of conditional independent statements on the basis of the Markov Condition:

$$\mathbf{z}_6 \perp\!\!\!\perp (\mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_1, \mathbf{z}_5) | \mathbf{z}_4 \tag{4.3.4}$$
$$\mathbf{z}_4 \perp\!\!\!\perp (\mathbf{z}_1, \mathbf{z}_5) | (\mathbf{z}_2, \mathbf{z}_3) \tag{4.3.5}$$
$$\mathbf{z}_2 \perp\!\!\!\perp (\mathbf{z}_3, \mathbf{z}_5) | (\mathbf{z}_1) \tag{4.3.6}$$

Note, for instance, that $\mathbf{z}_4$ is not independent of $\mathbf{z}_6$ since it is a descendant. May you write more independence statements?

•

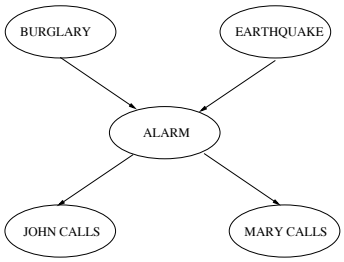


Figure 4.2: Alarm BN [164].

Example

This is a well-known example used in [164] to show a practical application of Bayesian Networks. Suppose you want to model a burglar alarm that is fairly reliable at detecting a burglary but also responds on occasion to minor earthquakes. You also have two neighbours, John and Mary, who promised to call you at work when they hear the alarm. John always calls when he hears the alarm but sometimes confuses the telephone ringing with the alarm and calls then, too. Mary likes loud music and sometimes misses the alarm. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary. We can describe the problem by using a BN (Figure 4.3) where all the variables are Boolean and denoted by capital letters to better remember their meaning.

The joint probability can be factorized as follows:

$$P(J, M, A, B, E) = P(J|A)P(M|A)P(A|B, E)P(B)P(E)$$

Suppose that the unconditional probability of a burglar (B) or an earthquake (E) are quite low, e.g. $P(B) = 0.001$ and $P(E) = 0.002$. Let the conditional probability tables be

B	E	P(A | B,E)	P(¬ A| B,E)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

A	P(J|A)	P(¬ J|A)
T	0.9	0.1
F	0.05	0.95

A	P(M|A)	P(¬ M|A)
T	0.7	0.3
F	0.01	0.99

What is the probability $\text{Prob}\{B = T|J = T\}$ (denoted $\text{Prob}\{B|J\}$ below), i.e. the probability that a burglar entered the house if John calls?

$$\text{Prob}\{B|J\} = \frac{\text{Prob}\{J|B\}\,\text{Prob}\{B\}}{\text{Prob}\{J\}} =$$

$$= \frac{\text{Prob}\{J|B\}\,\text{Prob}\{B\}}{\text{Prob}\{J|B\}\,\text{Prob}\{B\} + \text{Prob}\{J|\neg B\}\,\text{Prob}\{\neg B\}}$$

We have

$$\begin{aligned}\text{Prob}\{J|B\} &= \text{Prob}\{J|A,B\}\,\text{Prob}\{A|B\} + \text{Prob}\{J|\neg A,B\}\,\text{Prob}\{\neg A|B\} = \\ &= \text{Prob}\{J|A\}\,\text{Prob}\{A|B\} + \text{Prob}\{J|\neg A\}\,\text{Prob}\{\neg A|B\}\end{aligned}$$

Since

$$\begin{aligned}&\text{Prob}\{A|B\} = \text{Prob}\{A|B,E\} * \text{Prob}\{E\} + \text{Prob}\{A|B,\neg E\} * \text{Prob}\{\neg E\} = \\ &= 0.95 \cdot 0.002 + 0.94 \cdot (1 - 0.002) = 0.94 \\ &\text{Prob}\{A|\neg B\} = \text{Prob}\{A|\neg B,E\} * \text{Prob}\{E\} + \text{Prob}\{A|\neg B,\neg E\} * \text{Prob}\{\neg E\} = \\ &= 0.29 \cdot 0.002 + 0.001 \cdot (1 - 0.002) = 0.00158 \\ &\text{Prob}\{\neg A|\neg B\} = 1 - \text{Prob}\{A|\neg B\} = 0.9984\end{aligned}$$

it follows

$$\begin{aligned}&\text{Prob}\{J|B\} = 0.9 \cdot 0.94 + 0.05 \cdot (1 - 0.94) = 0.8490 \\ &\text{Prob}\{J|\neg B\} = \text{Prob}\{J|A\}\,\text{Prob}\{A|\neg B\} + \text{Prob}\{J|\neg A\}\,\text{Prob}\{\neg A|\neg B\} \\ &= 0.9 \cdot 0.00158 + 0.05 \cdot 0.9984 = 0.0513\end{aligned}$$

and

$$\text{Prob}\{B|J\} = \frac{0.8490 \cdot 0.001}{0.8490 \cdot 0.001 + 0.0513 \cdot (1 - 0.001)} = 0.016$$

You can retrieve the same results by running the script `alarm.R` which relies on first computing the entire joint probability distribution and then the ratio $\frac{\text{Prob}\{B,J\}}{\text{Prob}\{J\}}$.

•

Example

An interesting BN topology is known as *Naive Bayes* (Figure 4.3). In this case all variables $\mathbf{z}_i, 0 < i \leq n$ are conditionally independent given the variable $\mathbf{z}_0$

$$\mathbf{z}_i \perp\!\!\!\perp \mathbf{z}_j | \mathbf{z}_0 \qquad \forall i > 0, j > 0$$

It follows that the associated joint probability can be written as

$$P(Z) = P(z_0) \prod_{i=1}^{n} P(z_i|z_0)$$

Note that in this case, we need overall $2n+1$ parameters to encode the distribution, i.e. two parameters for each conditional distributions and one parameter to encode $P(z_0)$.

This probabilistic model is commonly used for probabilistic reasoning in medical diagnostic where $\mathbf{z}_0$ denotes the pathology class (the cause) and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ represent n symptoms (or effects) associated with the pathology. The assumption here is that symptoms depend only on the underlying pathology. The Naive Bayes principle also underlies a well-known classifier which will be presented in Section 10.2.3.1.

•

Definition 3.3 (Minimality condition)**.** Consider the BN $(\mathcal{G}, P)$ satisfying the MC. The BN satisfies the minimality condition iff for every proper subgraph $\mathcal{H}$ of $\mathcal{G}$ the pair $(\mathcal{H}, P)$ does not satisfy the MC.

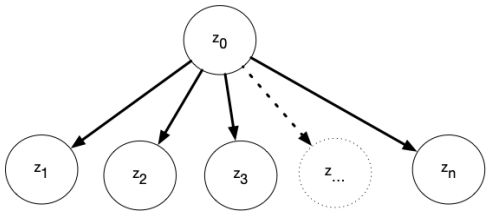


Figure 4.3: Naive Bayes topology.

4.3.1 Bayesian network and d-separation

The Markov Condition induces a set of conditional independence relations in a Bayesian Network. However, it is not easy to determine which other conditional independence relationships possibly hold.

In order to show the link between DAG topology and conditional independence, we introduce the criterion of *d-separation*. Let $\mathbf{Z}_{-(i,j)}$ the set obtained by removing $\mathbf{z}_i$ and $\mathbf{z}_j$ from $\mathbf{Z}$.

Definition 3.4 (d-separation)**.** In a DAG (Directed Acyclic Graph), two nodes $\mathbf{z}_i$ and $\mathbf{z}_j$ are *d-separated* by the *conditioning set* $\mathbf{S} \subseteq \mathbf{Z}_{-(i,j)}$ (denoted by $(\mathbf{z}_i \updownarrow \mathbf{z}_j|\mathbf{S})$) if every path from $\mathbf{z}_i$ to $\mathbf{z}_j$ is blocked by $\mathbf{S}$.

Two nodes are *d-connected* if they are not d-separated.

Definition 3.5 (Blocked path)**.** A path from $\mathbf{z}_i$ to $\mathbf{z}_j$ in a DAG is blocked by the conditioning set $\mathbf{S}$ if

- at least one diverging or serially connected (i.e. *non-collider*) node of the path is in $\mathbf{S}$, OR
- at least one converging node (*collider*) and all its descendants are not in $\mathbf{S}$

Example

Consider the graph $\mathcal{G}$ shown in Figure 4.1. If we consider the path $\mathbf{z}_2 \rightarrow \mathbf{z}_4 \leftarrow \mathbf{z}_3$ the node $\mathbf{z}_4$ is a collider. If we consider the path $\mathbf{z}_2 \rightarrow \mathbf{z}_4 \rightarrow \mathbf{z}_6$ the node $\mathbf{z}_4$ is a non-collider. It follows then that

- $\mathbf{z}_6$ is d-separated from $\mathbf{z}_2$ by the conditioning set $\mathbf{S} = \{\mathbf{z}_4\}$

$$(\mathbf{z}_6 \updownarrow \mathbf{z}_2|\mathbf{z}_4) \tag{4.3.7}$$

  since the only path is blocked (the serially connected node $\mathbf{z}_4 \in \mathbf{S}$ is in the path $\mathbf{z}_2 \rightarrow \mathbf{z}_4 \rightarrow \mathbf{z}_6$)

- $\mathbf{z}_2$ is not d-separated from $\mathbf{z}_3$ by the conditioning set $\mathbf{S} = \{\mathbf{z}_4\}$

$$(\mathbf{z}_3 \leftrightarrow \mathbf{z}_2|\mathbf{z}_4) \tag{4.3.8}$$

  since there is at least a path that is not blocked (the collider node $\mathbf{z}_4 \in \mathbf{S}$ is in the path $\mathbf{z}_2 \rightarrow \mathbf{z}_4 \rightarrow \mathbf{z}_3$)

•

A non-blocked path is also called *active*. The activity of a path depends on the activity of its (colliders and non-colliders) nodes:

- Non-colliders: when the conditioning set is empty, they are active. When they belong to the conditioning set they become inactive.
- Colliders: when the conditioning set is empty, they are inactive. They become active when they or some of their descendants are part of the conditioning set.

It follows that a path is not blocked or active when all the nodes are active.

R example

The R package `bnlearn` allows us to encode DAGs and performs checks of d-separation between sets of variables. This package is used in the script `dsep.R` to encode and then visualise the DAG in Figure 4.1. The script also uses the function `dsep` provided by the package `bnlearn` to check the existence of the d-separations corresponding to the conditional independence statements (4.3.7) and (4.3.8).

•

Note that, for the moment, d-separation is a pure graphical (and non probabilistic) notion related to the topology of the graph $\mathcal{G}$. In what follows, we will discuss how and when it may be informative about probabilistic properties.

4.3.2 D-separation and I-map

Definition 3.6 (I-map property)**.** A Bayesian Network $(\mathcal{G}, P)$ satisfies the I-map property if

$$\forall \mathbf{z}_i, \mathbf{z}_j \in \mathbf{Z}, \forall \mathbf{S} \subseteq \mathbf{Z}_{-(i,j)} : (\mathbf{z}_i \updownarrow \mathbf{z}_j | \mathbf{S}) \Rightarrow I(\mathbf{z}_i; \mathbf{z}_j | \mathbf{S}) = 0$$

i.e. if the d-separation $(\mathbf{z}_i \updownarrow \mathbf{z}_j | \mathbf{S})$ implies the conditional independence $\mathbf{z}_i \perp\!\!\!\perp \mathbf{z}_j | \mathbf{S}$ or, equivalently, the null conditional mutual information $I(\mathbf{z}_i; \mathbf{z}_j | \mathbf{S}) = 0$.

We remind the relation (3.8.80) between conditional mutual information and conditional independence.

The I-map property implies that the set of d-separations of $\mathcal{G}$ (denoted by $\mathcal{I}(\mathcal{G})$) is contained in the set of conditional independencies of P (denoted by $\mathcal{I}(P)$):

$$\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$$

Note that a completely connected DAG always satisfies the property above since $\mathcal{I}(\mathcal{G}) = \emptyset$.

It can be shown that if (and only if) a probability P satisfies the Markov condition for a given graph $\mathcal{G}$, each d-separation in the graph implies a conditional independence (MC $\Leftrightarrow$I-map).

In other terms, if one takes any instance of a distribution P which factorises according to the graph structure and $\mathcal{I}(P)$ is the list of all the conditional independence statements that can be obtained from P, if $\mathbf{z}_1$ and $\mathbf{z}_2$ are d-separated by $\mathbf{z}_3$, the independence $\mathbf{z}_1 \perp\!\!\!\perp \mathbf{z}_2 | \mathbf{z}_3$ belongs to $\mathcal{I}(P)$.

4.3.2.1 D-separation and faithfulness

If MC holds, d-separation implies independence. Can we also say the reverse, i.e. that d-connection (i.e. absence of d-separation) implies dependence? Do all distributions P factorised according to a graph G possess the dependencies related to the d-connection of the graph? Unless we make additional assumptions, the answer is no. The required additional assumption is called *faithfulness* between the graph and the distribution.

Definition 3.7 (Faithfulness)**.** A Bayesian Network $(\mathcal{G}, P)$ satisfies the Faithfulness property if

$$\forall \mathbf{z}_i, \mathbf{z}_j \in \mathbf{Z}, \forall \mathbf{S} \subseteq \mathbf{Z}_{-(i,j)} : (\mathbf{z}_i \leftrightarrow \mathbf{z}_j | \mathbf{S}) \Rightarrow I(\mathbf{z}_i; \mathbf{z}_j | \mathbf{S}) \neq 0$$

or equivalently if the conditional independence $\mathbf{z}_i \perp\!\!\!\perp \mathbf{z}_j | \mathbf{S}$ entails the d-separation

$$\forall \mathbf{z}_i, \mathbf{z}_j \in \mathbf{Z}, \forall \mathbf{S} \subseteq \mathbf{Z}_{-(i,j)} : I(\mathbf{z}_i; \mathbf{z}_j | \mathbf{S}) = 0 \Rightarrow (\mathbf{z}_i \updownarrow \mathbf{z}_j | \mathbf{S})$$

Faithfulness means that independence between two variables $\mathbf{z}_i$ and $\mathbf{z}_j$ (in information theory terms $I(\mathbf{z}_i; \mathbf{z}_j | \mathbf{S}) = 0$) implies d-separation, or equivalently that d-connection implies dependency.

When both the Markov condition and the faithfulness hold, there is a bijection between d-separation and conditional independence. The DAG is then said a *perfect map* of the joint probability distribution.

If faithfulness holds, we have a probabilistic independence interpretation of the graphical d-separation. This means, for instance, that when the conditioning set is empty non-colliders transmit information (dependence) while colliders do not (independence).

Example

Consider the BN in Figure 4.1 and suppose it is faithful. For an empty conditioning set, we can derive a number of relations of dependence, such as

$$\mathbf{z}_2 \not\perp\!\!\!\perp \mathbf{z}_6, \mathbf{z}_4 \not\perp\!\!\!\perp \mathbf{z}_5$$

Are there two independent variables?

•

Beware that many distributions have no perfect map in DAGs. The spectrum of probabilistic dependencies is, in fact, so rich that it cannot be cast into any representation scheme that uses a polynomial amount of storage ([Verma, 1987]).

So, how strong is the assumption of faithfulness? It is possible to show that a DAG is a perfect map for almost all distributions that factorise over $\mathcal{G}$ (i.e. all distributions except a set of measure zero) [117]. This means that assuming faithfulness is reasonable in most practical settings.

Example

Consider a symptom (e.g. headache) with two possible causes: a serious one (cancer) and a less serious one (virus). We can describe the problem by using a BN (Figure 4.4) where all the variables are Boolean. First of all, if we assume a perfect map, from d-separation we obtain that the variable C and V are independent but conditionally independent (i.e. conditioning on H the path from; C to V is unblocked).

Suppose that the a priori probability of the serious cause ($P(C) = 0.1$) is much lower than the one of the less serious one ($P(V) = 0.6$) and that the conditional probability table is

C	V	P(H | C,V)	P(¬ H| C,V)
T	T	0.95	0.05
T	F	0.8	0.2
F	T	0.8	0.2
F	F	0.1	0.9

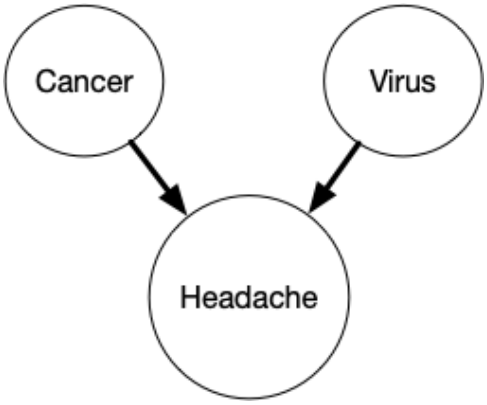


Figure 4.4: Common effect configuration.

From the script `expl_away.R` we compute the conditional probability of cancer in three different situations: headache only, headache and virus and headache but no virus:

$$P(C = T|H) = 0.1597846 \tag{4.3.9}$$
$$P(C = T|H, V) = 0.1165644 \tag{4.3.10}$$
$$P(C = T|H, \neg V) = 0.4705882 \tag{4.3.11}$$

Let us remark that the conditional probability (4.3.9) is higher than the a priori ($P(C) = 0.1$) if we observe that the patient has a headache. If we know that the patient is infected by a virus as well, the probability of cancer decreases to (4.3.10). We say that the virus explains away the cancer possibility. On the contrary, if we know that headache is present but the virus is absent, the cancer probability surges again to (4.3.11).

This non-monotone behaviour is caused by what is called the *explaining away effect*, i.e. if we have two common causes of an observed effect, knowing that one occurs (or not) reduces (or increases) the probability of the other. This is due to the fact that though virus and cancer are marginally independent ($P(C|V) = P(C)$), they are dependent once we condition on headache ($P(C|V, H) < P(C|H)$).

•

4.3.3 Skeleton and I-equivalence

In the case of perfect-map, a BN is fully specified by its conditional independent statements. A definition of equivalence follows, then:

Definition 3.8 (I-equivalence:)**.** Two graphs are I-equivalent if they have the same associated set of independencies.

All distributions that can be factorised on a graph $\mathcal{G}$, can also be factorised on an equivalent graph $\mathcal{G}'$.

In order to check the notion of equivalence visually, we introduce the notion of skeleton.

Definition 3.9 (Skeleton:)**.** The skeleton of a Bayesian Network graph $\mathcal{G}$ over $\mathbf{Z}$ is an undirected graph that contains an edge for every edge in $\mathcal{G}$

A sufficient (but not necessary) condition for equivalence of two graphs is that they have the same skeleton and the same set of v-structures. A *v-structure* occurs in

a DAG when there is a node having two entering edges (e.g. Figure 4.4). Complete graphs (e.g. completely connected triplets) are equivalent (no independence) but may have different v-structures.

A v-structure with no direct edge between its parents is also called *immorality*. A sufficient and necessary condition for the equivalence of two graphs is that they have the same skeleton and the same set of immoralities.

Observational equivalence places a limit on the ability to infer directionality from conditional probabilities alone. This means that there are classes of equivalence of graphs that cannot be distinguished using conditional independence tests. Two graphs that are I-equivalent cannot be distinguished without resorting to alternative strategies (e.g. manipulative experimentation or temporal information).

By considering conditional independence only, it is not possible to detect changes in graphs (e.g. reversing a single arc) that do not change the skeleton and which do not introduce or destroy a v-structure. The simplest example is provided by two graphs $\mathbf{z}_1 \rightarrow \mathbf{z}_2$ and $\mathbf{z}_2 \rightarrow \mathbf{z}_1$. They are equivalent (same skeleton and no v-structure) and they have associated an empty set of independencies.

4.3.4 Stable distributions

The use of the independence relationships made so far implies that the set of independences of the probability distribution associated with a graph depends only on the structure of the graph and not on the parametrisation.

This restriction is also known as *stability*: in other terms, we consider distributions whose independencies remain invariant to any change in the parameters. The stability assumption presumes that unstable independencies (i.e. dependencies that disappear with a parameter change) are unlikely to occur in the data, so all the independencies are structural.

In general, it is important to be aware of the limited expressibility of BNs. BNs cannot necessarily graphically represent all the independence properties of a given distribution. Consider, for instance, the distribution associated to $\mathbf{x}_1 \rightarrow \mathbf{y}_1 \leftarrow \mathbf{z} \rightarrow \mathbf{y}_2 \leftarrow \mathbf{x}_2$. If we marginalise the distribution wrt $\mathbf{z}$ (i.e. $\mathbf{z}$ is not observable), there is no DAG containing only the vertices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$, which represents the independence relations of the regional DAG without adding spurious independencies.

4.4 Markov networks

Markov networks (MN) are an undirected graphical representation of conditional independencies. Let us consider a set $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ of n random variables.

Definition 4.1. The conditional independent graph of $\mathbf{Z}$ is the undirected graph $\mathcal{G} = (V, E)$ where $V = \{1, \ldots, n\}$ and (i, j) is NOT in the edge set E iff

$$\mathbf{z}_i \perp\!\!\!\perp \mathbf{z}_j | \mathbf{Z}_{-\{i,j\}}$$

This graph is also called a *pairwise Markov graph*. Note that for n variables, there are $2^{\binom{n}{2}}$ potential undirected graphs.

4.4.1 Separating vertices, separated subsets and independence

As in the directed case, it is possible to use topological notions (e.g. separation of vertices in the network) to deduce probabilistic properties (e.g. conditional (in)dependencies). Given an undirected graph $\mathcal{G} = (V, E)$,

- a subset of vertices separates two vertices i and j if every path joining the two vertices contains at least one vertex from the separating subset
- a subset of vertices separates two subsets V_a and V_b in $\mathcal{G}$ if it separates every pair of vertices $i \in V_a$, $j \in V_b$.

The last property is also called the *global Markov property*. In general, it can be show that the set of distributions that satisfies the pairwise Markov property satisfies as well the global Markov property.

Given an undirected independence graph $\mathcal{G} = (V, E)$ it can be shown that:

- if V can be partitioned into two subsets V_b and V_c, such that there is no path between any vertex in V_b and any vertex in V_c then

$$\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j, \quad \text{forall } \mathbf{x}_i \in V_b \text{ and } \mathbf{x}_j \in V_c$$

- if V_a is any subset of vertices of G that separates i and j then

$$\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j | \mathbf{X}_a$$

4.4.2 Directed and undirected representations

BNs and MNs are two closely related yet different, representations and it is important not to confuse them. So why consider undirected representations besides directed Bayesian networks? What is their main difference? First of all, let us remind the Box golden rule of modelling: no model representation is perfect or exhaustive, all of them are wrong, but sometimes some of them are useful. Markov networks visualise conditional independence properties in distributions without having recourse to any notion of ordering. In that sense, they are more adequate when the considered problem is not explicitly associated to a specific ordering of variables or is characterised by symmetric relations. At the same time, asymmetric relationships (e.g. cause and effect, past and future) fit well the BN formalism.

As an example, let us consider two probabilistic distributions of n random variables. In the first case, the n variables represent a quantity measured during n consecutive time instants. In the second case, the n variables measure the same quantity over n different spatial locations. In both cases properties of conditional independence might help in representing and reasoning on the distribution. However, only the first case takes advantage of a Bayesian Network representation which encodes the explicit and asymmetric time ordering. An undirected representation, where a notion of symmetric neighbourhood is present, is more suitable to the spatial distribution task.

A second interesting issue is whether a MN is equivalent to a BN which have been deprived of the edge directionality. The answer is not so simple. Consider a DAG $\mathcal{G}$ and a faithful probability P. Let U be the undirected skeleton associated with $\mathcal{G}$ and U' be the undirected conditional independence graph associated with P. Which relationship exists between U and U'? U and U' generally are not the same, but the relation $U \subseteq U'$ holds. As shown by Wermuth and Lauritzen, U and U' are the same iff $\mathcal{G}$ does not contain any unshielded collider.

4.5 Conclusions

The graphical modelling formalism enables a modular representation of large variate problems thanks to the correspondence between topological properties and probabilistic notions. Graphical models are then effective tools that can be used to

represent and communicate the relations between a large number of variables and to perform probabilistic reasoning.

In general, an effective modelling approach has to manage the trade-off between complexity and fidelity to reality. Graphical modelling uses the notion of conditional independence to address such issue.

Note that the adoption of conditional independence assumptions to simplify representations is pervasive in mathematical modelling and human reasoning: think, for instance, to the notion of state in dynamical systems, which makes the future behaviour independent from the past given the present. Simplifying by conditioning is also a peculiar characteristic of human causal reasoning: once we find the cause of a certain phenomenon, we can disregard all other variables as irrelevant.

We will see that in machine learning, graphical modelling is a powerful way to explain why some variables are more important or relevant than others (Chapter 12). At the same time, machine learning strategies may be used to infer compact, graphical (and sometimes causal) representations from data (Chapter 13).

Chapter 5

Parametric estimation and testing

Given the correct probabilistic model of a phenomenon, we may derive the properties of observable data by logical deduction. The theory of statistics is designed to reverse the deductive process (Chapter 2). It takes measured data and uses them to propose a probabilistic model, to estimate its parameters and eventually to validate it. This chapter will focus on the estimation methodology, intended as the inductive process which leads from observed data to a probabilistic description of reality. We will focus here on the *parametric* approach, which assumes that we know all about the probabilistic model except the value of a finite number of parameters. Parametric estimation algorithms build estimates from data and, more important, statistical measures to assess their quality. There are two main approaches to parametric estimation:

Classical or frequentist: it is based on the idea that sample data are the sole quantifiable form of relevant information and that the parameters are *fixed but unknown*. It is related to the frequency view of probability (Section 3.1.4).

Bayesian approach: the parameters are supposed to be *random variables*, having a distribution *prior* to data observation and a distribution *posterior* to data observation. This approach assumes that there exists something beyond data, (i.e. a human sense of uncertainty or a subjective degree of belief), and that this belief can be described in the probabilistic form.

It is well known, however, that in large-sample problems, frequentist and Bayesian approaches tend to produce similar numerical results and that in small-medium settings, though the two outcomes may not coincide, their difference is usually small. For those reasons and, mainly for reasons of space, we will limit here to consider the classical approach. It is important, however, not to underestimate the important role of the Bayesian estimation philosophy, which led recently to a large amount of research in Bayesian data analysis and important applications in machine learning [77].

5.1 Classical approach

The classical approach to parameter estimation dates back to the period 1920-35 when J. Neyman and E.S. Pearson, stimulated by problems in biology and industry, concentrated on the principles for testing hypothesis and R.A. Fisher, interested in agricultural issues, focused on the estimation from data.

We will introduce estimation by considering a simple univariate setting. Let $\mathbf{z}$ be a continuous r.v. and suppose that

1. we know the analytical form of the distribution family

$$F_{\mathbf{z}}(z) = F_{\mathbf{z}}(z, \theta)$$

   but the parameter vector $\theta \in \Theta$ is unknown,

2. we have access to a set D_N of N i.i.d. measurements of $\mathbf{z}$, called *sample data*.

In the general case, few parameters are not enough to describe a function, like the density function: in that sense, parametric densities are an obvious simplification. An example of a parametric distribution function is the Normal distribution (Section (3.4.2)), where the parameter vector is $\theta = [\mu, \sigma]$. The goal of the estimation procedure is to find a value $\hat{\theta}$ of the parameter θ so that the parameterised distribution $F_{\mathbf{z}}(z, \hat{\theta})$ closely matches the distribution of data.

The notation *i.i.d.* stands for **i**dentically and **i**ndependently **d**istributed. Identically distributed means that all the observations have been sampled from the same distribution, that is

$$\text{Prob}\{\mathbf{z}_i = z\} = \text{Prob}\{\mathbf{z}_j = z\} \qquad \text{for all } i, j = 1, \ldots, N \text{ and } z \in \mathcal{Z}$$

Independently distributed means that the fact that we have observed a certain value z_i does not influence the probability of observing the value z_j, that is

$$\text{Prob}\{\mathbf{z}_j = z | \mathbf{z}_i = z_i\} = \text{Prob}\{\mathbf{z}_j = z\}$$

Example

Here you find some examples of estimation problems:

1. Let $D_N = \{20, 31, 14, 11, 19, \ldots\}$ be the times in minutes spent the last 2 weeks to go home. What is the mean time to reach my house from ULB?

2. Consider the car traffic in the boulevard Jacques. Suppose that the measures of the inter-arrival times are $D_N = \{10, 11, 1, 21, 2, \ldots\}$ seconds. What does this imply about the mean inter-arrival time?

3. Consider the students of the last year of Computer Science. What is the variance of their grades?

4. Let $\mathbf{z}$ be the r.v. denoting tomorrow's temperature. How can I estimate its mean value on the basis of past observations?

•

Parametric estimation is a *mapping* from the space of the sample data to the space of parameters Θ. The two possible outcomes are:

1. some specific value of Θ. In this case, we have the so-called *point estimation*.

2. some particular region of Θ. In this case, we obtain an *interval of confidence* on the value of the parameter.

5.1.1 Point estimation

Consider a random variable $\mathbf{z}$ with a parametric distribution $F_{\mathbf{z}}(z, \theta)$, $\theta \in \Theta$. The unknown parameter can be written as a function(al) of F

$$\theta = t(F)$$

This corresponds to the fact that θ is a characteristic of the population described by $F_{\mathbf{z}}(\cdot)$. For instance the expected value parameter $\mu = t(F) = \int z dF(z)$ is a functional of F.

Suppose now that we have observed a set of N i.i.d. values $D_N = \{z_1, z_2, \ldots, z_N\}$. A *point estimate* is an example of *statistic*, where by statistic it is generally meant any function of the sample data D_N. In other terms a *point estimate* is a function

$$\hat{\theta} = g(D_N) \tag{5.1.1}$$

of the sample dataset D_N, where $g(\cdot)$ stands for the estimation algorithm, that is the procedure which returns the estimation starting from a dataset D_N. Note that, from a machine learning perspective, it is more appropriate to consider g, rather than a conventional mathematical function, as a generic algorithm taking the sample dataset as an input and returning an estimation as output[1].

There are two main issues in estimation and, more generally, in data analysis, statistics and machine learning: how to construct an estimator (i.e. which form should g take) and how to assess the quality of the returned estimation $\hat{\theta}$. In Sections 5.3 and 5.8 we will discuss two strategies for defining an estimator; the *plug-in principle* and the *maximum likelihood*. In Section 5.5 we will present the statistical measures most commonly adopted to assess an estimator accuracy.

Before introducing the plug-in principle, we need, however, to present the notion of empirical distribution.

5.2 Empirical distributions

Suppose we have observed a i.i.d. random sample of size N from a probability distribution $F_{\mathbf{z}}(\cdot)$

$$F_{\mathbf{z}} \rightarrow \{z_1, z_2, \ldots, z_N\}$$

The *empirical distribution probability* $\hat{F}$ is defined as the *discrete distribution* that assigns probability $1/N$ to each value z_i, $i = 1, \ldots, N$. In other words, $\hat{F}$ assigns to a set A in the sample space of $\mathbf{z}$ its empirical probability

$$\text{Prob}\{\mathbf{z} \in A\} \approx \frac{\#z_i \in A}{N}$$

that is the proportion of the observations in D_N which occur in A.

It can be proved that the vector of observed frequencies in $\hat{F}$ is a *sufficient* statistic for the true distribution $F(\cdot)$, i.e. all the information about $F(\cdot)$ contained in D_N is also contained in $\hat{F}(\cdot)$.

Consider now the distribution function $F_{\mathbf{z}}(z)$ of a continuous rv $\mathbf{z}$ and a set of N observations $D_N = \{z_1, \ldots, z_N\}$. Since

$$F_{\mathbf{z}}(z) = \text{Prob}\{\mathbf{z} \leq z\}$$

we define $N(z)$ as the number of observations in D_N that do not exceed z. We obtain then the empirical estimate of $F(\cdot)$

$$\hat{F}_{\mathbf{z}}(z) = \frac{N(z)}{N} = \frac{\#z_i \leq z}{N} \tag{5.2.2}$$

[1] For instance, an awkward, yet acceptable, estimation algorithm could take the dataset, discard all the examples except the third one and return it as the estimation.

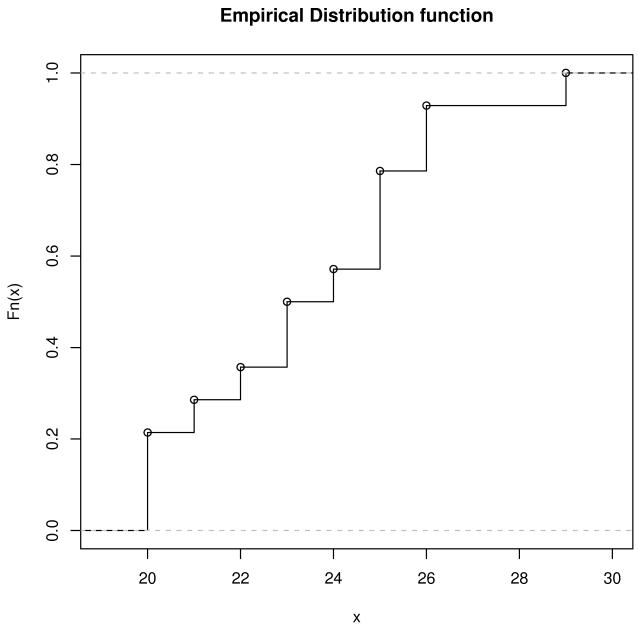


Figure 5.1: Empirical distribution.

This function is a staircase function with discontinuities at the points z_i (Figure 5.1).

Example

Suppose that our dataset is made of the following $N = 14$ observations

$$D_N = \{20, 21, 22, 20, 23, 25, 26, 25, 20, 23, 24, 25, 26, 29\}$$

The empirical distribution function $\hat{F}_{\mathbf{z}}$ (which can be traced by running the script `cumdis.R`) is plotted in Figure 5.1.

5.3 Plug-in principle to define an estimator

Consider an r.v. $\mathbf{z}$ and sample dataset D_N drawn from the parametric distribution $F_{\mathbf{z}}(z, \theta)$. The main issue of estimation is how to define an estimate of θ. A possible solution is given by the *plug-in principle*, that is a simple method of estimating parameters from observations. The *plug-in estimate* of a parameter (or target) θ is defined to be:

$$\hat{\theta} = t(\hat{F}(z)) \tag{5.3.3}$$

obtained by replacing the distribution function with the empirical distribution in the analytical expression of the parameter.

The following section will discuss the plug-in estimators of the first two moments of a probability distribution.

5.3.1 Sample average

Consider an r.v. $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$ such that

$$\theta = E[\mathbf{z}] = \int z dF(z)$$

with θ unknown. Suppose we have available the sample $F_{\mathbf{z}} \rightarrow D_N$, made of N observations. The *plug-in* point estimate of θ is given by the *sample average*

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} z_i = \hat{\mu} \tag{5.3.4}$$

Note that the sample average is not a parameter (i.e. it is not a function of the probability distribution $F_{\mathbf{z}}$) but a statistic (i.e. a function of the dataset D_N).

5.3.2 Sample variance

Consider a r.v. $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$ where the mean μ and the variance σ^2 are unknown. Suppose we have available the sample $F_{\mathbf{z}} \rightarrow D_N$. Once we have the sample average $\hat{\mu}$, the *plug-in* estimate of σ^2 is given by the *sample variance*

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \hat{\mu})^2 \tag{5.3.5}$$

The presence of $N-1$ instead of N at the denominator will be explained later. Note also that the following relation holds for all z_i

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{\mu})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} z_i^2\right) - \hat{\mu}^2$$

The expression of the plug-in estimators of other interesting probabilistic parameters are in the Appendix (D).

5.4 Sampling distribution

Given a dataset D_N of N observations sampled from $\mathbf{z}$, let us consider a point estimate

$$\hat{\theta} = g(D_N) \tag{5.4.6}$$

Note that since D_N is the outcome of N realisations of a r.v. $\mathbf{z}$, the vector D_N can be considered as the realisation of a random vector $\mathbf{D}_N$[2].

By applying the transformation g to the random variable $\mathbf{D}_N$ we obtain the random variable

$$\hat{\boldsymbol{\theta}} = g(\mathbf{D}_N) \tag{5.4.7}$$

which is called the *point estimator* of θ. A key point is the following: *while θ is an (unknown) fixed value, the estimator $\hat{\boldsymbol{\theta}}$ is a random variable.* For instance, if we aim to estimate $\theta = \mu$ (expected value of $\mathbf{z}$) the parameter μ is an unknown and fixed value while the average $\hat{\boldsymbol{\mu}}$ is a random variable (since it is a function of a random dataset).

[2]This is not a mathematical detail but an essential aspect of the data-driven discovery process under uncertainty. Every model learned from data, or more in general all knowledge acquired from data, is built on random foundations and, as such, it is a random quantity and has to be assessed as such.

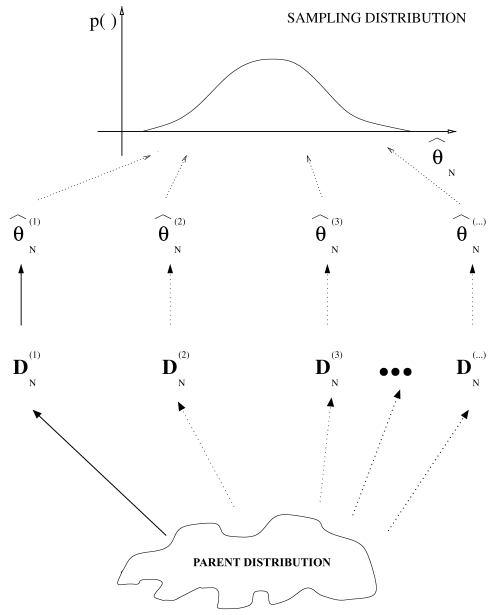


Figure 5.2: From the parametric parent distribution of $F_{\mathbf{z}}(\cdot, \theta)$ (underlying the data generation) to the sampling distribution of the estimator $\hat{\boldsymbol{\theta}}_N$. Each dataset has the same size N.

The probability distribution of the r.v. $\hat{\boldsymbol{\theta}}$ is called the *sampling distribution*, while the distribution of the r.v. $\mathbf{z}$ (with parameter θ) is called the *parent distribution*. An example of the process bringing from the parent to the sampling distribution is plotted in Figure 5.2. Note that the sampling distribution, though a theoretical quantity, is of great significance in estimation since it quantifies the estimator's accuracy in probabilistic terms, or, in simpler words, the gap between the estimation and the parameter θ.

5.4.1 Shiny dashboard

The dashboard `estimation.R` (Appendix G) provides an interactive visualisation of the sampling distribution of the plug-in estimators of the parameters (mean and variance) of a Normal parent distribution $\mathbf{z}$. We invite the reader to modify the values N , μ and σ and to observe the impact on the sampling distribution. Note that the sampling distribution is obtained by a Monte Carlo simulation of the process illustrated in Figure 5.2. The simulation (Algorithm 1 and related R code in Table 5.1) consists in repeating a number (adjustable) of trials where for each trial a sample dataset of size N is generated and the plug-in estimations are computed. The dashboard shows the histograms of the estimations.

Algorithm 1 Monte Carlo simulation to generate a sampling distribution

1: $S = \{\}$
2: **for** $r = 1$ to R **do**
3: $F_{\mathbf{z}} \rightarrow D_N = \{z_1, z_2, \ldots, z_N\}$ ▷ // pseudo-random sample generation
4: $\hat{\theta} = g(D_N)$ ▷ // estimation computation
5: $S = S \cup \{\hat{\theta}\}$
6: **end for**
7: Plot histogram of S
8: Compute statistics of S (mean, variance)
9: Study distribution of S with respect to θ (e.g. estimate bias)

```
mu<-0  # parameter
R<-10000  # number trials
N<-20   # size dataset
S<-numeric(R)
for (r in 1:R){
     D<-rnorm(N,mean=mu,sd=10)
#  pseudo-random sample generation

S[r]<-mean(D)
#  compute estimate
}
hist(S)
#  Plot histogram of S

bias=mean(S)-mu
#  Estimate bias
```

Table 5.1: R version of Algorithm 1 pseudo-code to generate the sampling distribution of $\hat{\boldsymbol{\mu}}$.

5.5 The assessment of an estimator

Once defined an estimator $\hat{\boldsymbol{\theta}}$ (e.g. in algorithmic or mathematical form), it is possible to assess its accuracy from its sampling distribution.

5.5.1 Bias and variance

The following measures rely on the sampling distribution[3] to assess the estimator accuracy.

Definition 5.1 (Bias of an estimator)**.** An estimator $\hat{\boldsymbol{\theta}}$ of θ is said to be *unbiased* if and only if

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] = \theta$$

Otherwise, it is said to be *biased* with bias

$$\text{Bias}[\hat{\boldsymbol{\theta}}] = E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}] - \theta \tag{5.5.8}$$

Definition 5.2 (Variance of an estimator)**.** The variance of an estimator $\hat{\boldsymbol{\theta}}$ of θ is the variance of its sampling distribution

$$\text{Var}\left[\hat{\boldsymbol{\theta}}\right] = E_{\mathbf{D}_N}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2]$$

[3]please note that we refer to the $\hat{\boldsymbol{\theta}}$ distribution and not to the $\mathbf{z}$ distribution

Definition 5.3 (Standard error)**.** The square root of the variance

$$\hat{\boldsymbol{\sigma}} = \sqrt{\text{Var}\left[\hat{\boldsymbol{\theta}}\right]}$$

is called the *standard error* of the estimator $\hat{\boldsymbol{\theta}}$.

An unbiased estimator is an estimator that, on average, has the right value but averaged over what? It is important to retain that this average is over different realisations of the dataset D_N as made explicit by the notation $E_{\mathbf{D}_N}[\hat{\boldsymbol{\theta}}]$, represented visually by Figure 5.2 and simulated by the Monte Carlo repetitions in Section (5.4.1).

Note that different unbiased estimators may exist for a parameter θ. Also, a biased estimator with a known bias (i.e. not depending on θ) is equivalent to an unbiased estimator since we can easily compensate for the bias. We will see in Section 5.5.3 that for some specific estimators it is possible to derive analytically the bias. Unfortunately, in general, the bias is not measurable since this would require the knowledge of θ which is in fact the target of our estimation procedure: nevertheless, the notion of bias is an important theoretical quantity to reason about the accuracy of an estimation process.

Sometimes we are accurate (e.g. unbiased) in estimating θ though we are interested in $f(\theta)$. Given a generic transformation $f(\cdot)$, if $\hat{\boldsymbol{\theta}}$ is unbiased for θ this does not imply that that $f(\hat{\boldsymbol{\theta}})$ is unbiased for $f(\theta)$ as well. This implies, for instance, that the standard error $\hat{\boldsymbol{\sigma}}$ is not an unbiased estimator of standard deviation σ despite $\hat{\boldsymbol{\sigma}}^2$ being an unbiased estimator of σ^2.

5.5.2 Estimation and the game of darts

An intuitive manner of visualising the notion of sampling distribution of an estimator and the related concepts of bias and variance is to use the analogy of the darts game.

The unknown parameter θ can be seen as the darts game target and the estimator $\hat{\boldsymbol{\theta}}$ as a player. Figure 5.3 shows the target (black dot) together with the distribution of the draws of two different players: the C (cross) player and the R (round) player. In terms of our analogy the cross player/estimator has small variance but large bias, while the round one has small bias and large variance. Which one is the best?

Now it's your turn to draw the shot distribution of a player with low bias and low variance and of a player which large bias and large variance.

5.5.3 Bias and variance of $\hat{\boldsymbol{\mu}}$

This section shows that for a generic r.v. $\mathbf{z}$ and an i.i.d. dataset D_N, the sample average $\hat{\boldsymbol{\mu}}$ is an unbiased estimator of the mean $E[\mathbf{z}]$.

Consider a random variable $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$. Let μ and σ^2 the mean and the variance of $F_{\mathbf{z}}(\cdot)$, respectively. Suppose we have observed the i.i.d. sample $D_N \leftarrow F_{\mathbf{z}}$. From (5.3.4) we obtain

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\mu}}] = E_{\mathbf{D}_N}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\right] = \frac{\sum_{i=1}^{N}E[\mathbf{z}_i]}{N} = \frac{N\mu}{N} = \mu \tag{5.5.9}$$

This means that the *sample average estimator* is not biased, whatever the distribution $F_{\mathbf{z}}(\cdot)$ is. And what about its variance? Since according to the i.i.d. assumption $\text{Cov}[\mathbf{z}_i, \mathbf{z}_j] = 0$, for $i \neq j$, from (3.10.85) we obtain that the variance of the *sample*

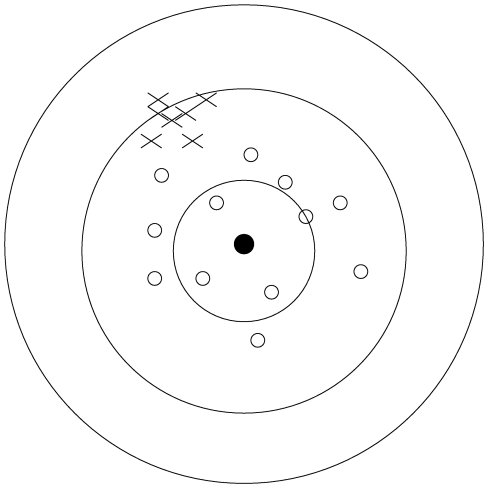


Figure 5.3: The dart analogy: the target is the unknown parameter, the round dots represent some realisations of the estimator R, while the crosses represent some realisations of the estimator C.

average estimator is

$$\text{Var}\left[\hat{\boldsymbol{\mu}}\right] = \text{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\right] = \frac{1}{N^2}\text{Var}\left[\sum_{i=1}^{N}\mathbf{z}_i\right] = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}. \tag{5.5.10}$$

In fact, $\hat{\boldsymbol{\mu}}$ acts like the "round player" in darts game (Figure 5.3) with some variance but no bias.

You can visualise the bias and variance of the sample average estimator by running the Shiny dashboard `estimation.R` introduced in Section 5.4.

5.5.4 Bias of the estimator $\hat{\boldsymbol{\sigma}}^2$

Let us study now the bias of the estimator of the variance of $\mathbf{z}$.

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\sigma}}^2] = E_{\mathbf{D}_N}\left[\frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{z}_i - \hat{\boldsymbol{\mu}})^2\right] \tag{5.5.11}$$

$$= \frac{N}{N-1}E_{\mathbf{D}_N}\left[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \hat{\boldsymbol{\mu}})^2\right] \tag{5.5.12}$$

$$= \frac{N}{N-1}E_{\mathbf{D}_N}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^2\right) - \hat{\boldsymbol{\mu}}^2\right] \tag{5.5.13}$$

Since $E[\mathbf{z}^2] = \mu^2 + \sigma^2$ and $\text{Cov}[\mathbf{z}_i, \mathbf{z}_j] = 0$, the first term inside the $E[\cdot]$ is

$$E_{\mathbf{D}_N}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^2\right)\right] = \frac{1}{N}\sum_{i=1}^{N}E_{\mathbf{D}_N}\left[\mathbf{z}_i^2\right] = \frac{1}{N}N(\mu^2 + \sigma^2)$$

Since $E\left[\left(\sum_{i=1}^{N}\mathbf{z}_i\right)^2\right] = N^2\mu^2 + N\sigma^2$ the 2nd term is

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\mu}}^2] = \frac{1}{N^2} E_{\mathbf{D}_N}\left[\left(\sum_{i=1}^{N} \mathbf{z}_i\right)^2\right] = \frac{1}{N^2}(N^2\mu^2 + N\sigma^2) = \mu^2 + \sigma^2/N$$

It follows that

$$E_{\mathbf{D}_N}[\hat{\boldsymbol{\sigma}}^2] \;=\; \frac{N}{N-1}\left((\mu^2+\sigma^2) - (\mu^2+\sigma^2/N)\right) \;=\; \frac{N}{N-1}\left(\frac{N-1}{N}\sigma^2\right) \;=\; \sigma^2$$

This result justifies our definition (5.3.5). Once the term $N-1$ is inserted at the denominator, the sample variance estimator is not biased.

Some points are worth considering:

- The results (5.5.9),(5.5.10) and (5.5.11) are *independent* of the family of the distribution $F(\cdot)$.
- According to (5.5.10), the variance of $\hat{\boldsymbol{\mu}}$ is $1/N$ times the variance of $\mathbf{z}$. This is a formal justification of the reason why taking averages on a large number of observations is recommended: the larger N, the smaller is $\text{Var}\,[\hat{\boldsymbol{\mu}}]$, so a bigger N for a given σ^2 implies a better estimate of μ.
- According to the central limit theorem (Section C.7), under quite general conditions on the distribution $F_{\mathbf{z}}$, the distribution of $\hat{\boldsymbol{\mu}}$ will be approximately normal as N gets large, which we can write as

$$\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu, \sigma^2/N) \quad \text{for } N \to \infty$$

- The *standard error* $\sqrt{\text{Var}\,[\hat{\boldsymbol{\mu}}]} = \frac{\sigma}{\sqrt{N}}$ is a common measure of statistical accuracy. Roughly speaking, if the estimator is not biased and the conditions of the central limit theorem apply, we expect $\hat{\boldsymbol{\mu}}$ to be less than one standard error away from μ about 68% of the time, and less than two standard errors away from μ about 95% of the time (see Table 3.2) .

Script

You can visualize the bias and variance of the sample variance estimator by running the following R script `sam_dis2.R` or by running the Shiny dashboard `estimation.R` introduced in Section 5.4..

5.5.5 A tongue-twister exercise

It sounds like a tongue-twister but it is important that the reader takes some time to reason on the substantial difference between two quantities like

1. the variance of an estimator and
2. the estimator of the variance.

The first quantity is denoted by $\text{Var}\left[\hat{\boldsymbol{\theta}}\right]$, is a real number and measures the accuracy of an estimator. It has been introduced in Section 5.5.

The second is denoted $\hat{\boldsymbol{\sigma}}^2$, is a random quantity since it is an estimator and its properties (e.g. bias) has been discussed in Section 5.5.4.

Now, if you understand the difference between the two quantities above, you could reason on $\text{Var}\left[\hat{\boldsymbol{\sigma}}^2\right]$, which is nothing more than the *variance of the estimator of the variance*. Clear, isn't it? And what about the *estimator of the variance of the estimator of the variance*?

5.5.6 Bias/variance decomposition of MSE

Bias and variance are two independent criteria to assess the quality of an estimator. As shown in Figure 5.3 we could have two estimators behaving in opposite ways: the first has large bias and low variance, while the second has large variance and small bias. How can we choose among them? We need a measure able to combine or merge the two to a single criteria. This is the role of the *mean-square error* (MSE) measure.

When $\hat{\boldsymbol{\theta}}$ is a biased estimator of θ, its accuracy is usually assessed by its MSE rather than simply by its variance. The MSE is defined by

$$\text{MSE} = E_{\mathbf{D}_N}[(\theta - \hat{\boldsymbol{\theta}})^2]$$

For a generic estimator it can be shown that

$$\text{MSE} = (E[\hat{\boldsymbol{\theta}}] - \theta)^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right] = \left[\text{Bias}[\hat{\boldsymbol{\theta}}]\right]^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right] \tag{5.5.14}$$

i.e., the mean-square error is equal to the sum of the variance and the squared bias of the estimator . Here it is the analytical derivation

$$\text{MSE} = E_{\mathbf{D}_N}[(\theta - \hat{\boldsymbol{\theta}})^2] = E_{\mathbf{D}_N}[(\theta - E[\hat{\boldsymbol{\theta}}] + E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] = \tag{5.5.15}$$

$$= E_{\mathbf{D}_N}[(\theta - E[\hat{\boldsymbol{\theta}}])^2] + E_{\mathbf{D}_N}[(E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] + E_{\mathbf{D}_N}[2(\theta - E[\hat{\boldsymbol{\theta}}])(E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})] = \tag{5.5.16}$$

$$= E_{\mathbf{D}_N}[(\theta - E[\hat{\boldsymbol{\theta}}])^2] + E_{\mathbf{D}_N}[(E[\hat{\boldsymbol{\theta}}] - \hat{\boldsymbol{\theta}})^2] + 2(\theta - E[\hat{\boldsymbol{\theta}}])(E[\hat{\boldsymbol{\theta}}] - E[\hat{\boldsymbol{\theta}}]) = \tag{5.5.17}$$

$$= (E[\hat{\boldsymbol{\theta}}] - \theta)^2 + \text{Var}\left[\hat{\boldsymbol{\theta}}\right] \tag{5.5.18}$$

This decomposition is typically called the *bias-variance* decomposition. Note that, if an estimator is unbiased then its MSE is equal to its variance.

5.5.7 Consistency

Suppose that the sample data contains N independent observations $z_1, \ldots, z_N$ of a univariate random variable. Let the estimator of θ based on N observations be denoted $\hat{\boldsymbol{\theta}}_N$. As N becomes larger, we might reasonably expect that $\hat{\boldsymbol{\theta}}_N$ improves as estimator of θ (in other terms it gets closer to θ). The notion of consistency formalizes this concept.

Definition 5.4. The estimator $\hat{\boldsymbol{\theta}}_N$ is said to be *weakly consistent* if $\hat{\boldsymbol{\theta}}_N$ converges to θ in probability, that is

$$\forall \epsilon > 0 \qquad \lim_{N\to\infty} \text{Prob}\left\{|\hat{\boldsymbol{\theta}}_N - \theta| \leq \epsilon\right\} = 1$$

Definition 5.5. The estimator $\hat{\boldsymbol{\theta}}_N$ is said *strongly consistent* if $\hat{\boldsymbol{\theta}}_N$ converges to θ with probability 1 (or almost surely).

$$\text{Prob}\left\{\lim_{N\to\infty} \hat{\boldsymbol{\theta}}_N = \theta\right\} = 1$$

For a scalar θ the property of convergence guarantees that the sampling distribution of $\hat{\boldsymbol{\theta}}_N$ becomes less disperse as $N \to \infty$. In other terms a consistent estimator is asymptotically unbiased. It can be shown that a sufficient condition for weak consistency of unbiased estimators $\hat{\boldsymbol{\theta}}_N$ is that $\text{Var}\left[\hat{\boldsymbol{\theta}}_N\right] \to 0$ as $N \to \infty$.

It is important to remark that the property of unbiasedness (for finite-size samples) and consistency are largely unrelated.

Exercise

Consider an estimator of the mean that takes into consideration only the first 10 sample points, whatever the total number $N > 10$ of observations is. Is such estimator consistent?

•

5.5.8 Efficiency

Suppose we have two *unbiased and consistent* estimators. How to choose between them?

Definition 5.6 (Relative efficiency)**.** Let us consider two unbiased estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. If

$$\text{Var}\left[\hat{\boldsymbol{\theta}}_1\right] < \text{Var}\left[\hat{\boldsymbol{\theta}}_2\right]$$

we say that $\hat{\boldsymbol{\theta}}_1$ is *more efficient* than $\hat{\boldsymbol{\theta}}_2$.

If the estimators are biased, typically the comparison is done on the basis of the mean square error.

Exercise

Suppose $z_1, \ldots, z_N$ is a random sample of observations from a distribution with mean θ and variance σ^2. Study the unbiasedness and the consistency of the three estimators of the mean μ:

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum_{i=1}^{N} z_i}{N}$$
$$\hat{\theta}_2 = \frac{N\hat{\theta}_1}{N+1}$$
$$\hat{\theta}_3 = z_1$$

•

5.6 The Hoeffding's inequality

A probabilistic measure of the discrepancy between the estimator $\hat{\boldsymbol{\mu}}$ and the quantity $\mu = E[\mathbf{z}]$ to be estimated is returned by the Hoeffding's inequality.

Theorem 6.1. *[101] Let $\mathbf{z}_1, \ldots, \mathbf{z}_N$ be independent bounded random variables such that $\mathbf{z}_i$ falls in the interval $[a_i, b_i]$ with probability one. Let their sum be $\mathbf{S}_N = \sum_{i=1}^{N} \mathbf{z}_i$. Then for any $\varepsilon > 0$ we have*

$$Prob\left\{|\mathbf{S}_N - E[\mathbf{S}_N]| > \varepsilon\right\} \leq \exp\left\{-2\varepsilon^2 / \sum_{i=1}^{N}(b_i - a_i)^2\right\}$$

Corollary 6.2. *If the variables $\mathbf{z}_1, \ldots, \mathbf{z}_N$ are independent and identically distributed, the following bound on the discrepancy between the sample mean $\hat{\mu} = \frac{\sum_{i=1}^{N} z_i}{N}$ and the expected value $E[\mathbf{z}]$ holds*

$$Prob\left\{|\hat{\boldsymbol{\mu}} - E[\mathbf{z}]| > \varepsilon\right\} \leq \exp\left\{-2N\varepsilon^2/(b-a)^2\right\}$$

Assume that δ is a confidence parameter, that is we are $100(1-\delta)\%$ confident that the estimate $\hat{\mu}$ is within the accuracy ε of the true expectation. It is possible to derive the expression

$$\varepsilon(N) = \sqrt{\frac{(b-a)^2 \log(2/\delta)}{2N}}$$

which measures with confidence $1-\delta$ how the sample mean $\hat{\mu}$, estimated on the basis of N points, is close to the expectation $E[\mathbf{z}]$. We can also determine the number of observations N necessary to obtain an accuracy ε and a confidence δ by using the relation

$$N > \frac{(b-a)^2 \log(2/\delta)}{2\varepsilon^2}$$

Hoeffding's bound is a general bound that only relies on the assumption that sample points are drawn independently. Bayesian bounds are another example of statistical bounds which give tighter results under the assumption that the examples are drawn from a normal distribution.

5.7 Sampling distributions for Gaussian r.v.s

The results in Section 5.5 are independent of the type of distribution function $F_{\mathbf{z}}$. Additional results are available in the specific case of a normal random variable.

Let $\mathbf{z}_1, \ldots, \mathbf{z}_N$ be i.i.d. realisation of $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ and let us consider the following sample statistics

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i, \quad \widehat{\mathbf{SS}} = \sum_{i=1}^{N} (\mathbf{z}_i - \hat{\boldsymbol{\mu}})^2, \quad \hat{\boldsymbol{\sigma}}^2 = \frac{\widehat{\mathbf{SS}}}{N-1}$$

It can be shown that the following relations hold

- $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu, \sigma^2/N)$ and $N(\hat{\boldsymbol{\mu}}-\mu)^2 \sim \sigma^2\chi_1^2$ where the χ^2 distribution is presented in Appendix C.2.2.
- $\mathbf{z}_i - \mu \sim \mathcal{N}(0, \sigma^2)$, so $\sum_{i=1}^{N}(\mathbf{z}_i - \mu)^2 \sim \sigma^2\chi_N^2$.
- $\sum_{i=1}^{N}(\mathbf{z}_i - \mu)^2 = \widehat{\mathbf{SS}} + N(\hat{\mu} - \mu)^2$.
- $\widehat{\mathbf{SS}} \sim \sigma^2\chi_{N-1}^2$ or equivalently $\frac{(N-1)\hat{\boldsymbol{\sigma}}^2}{\sigma^2} \sim \chi_{N-1}^2$. See R script `sam_dis2.R`.
- $\sqrt{N}(\hat{\boldsymbol{\mu}} - \mu)/\hat{\boldsymbol{\sigma}} \sim \mathcal{T}_{N-1}$ where $\mathcal{T}$ stands for the Student distribution (Section C.2.3).
- if $E[|\mathbf{z} - \mu|^4] = \mu_4$ then $\mathrm{Var}\left[\hat{\boldsymbol{\sigma}}^2\right] = \frac{1}{N}\left(\mu_4 - \frac{N-3}{N-1}\sigma^4\right)$.

5.8 The principle of maximum likelihood

Maximum-likelihood is a major strategy used in statistics to design an estimator, i.e. the algorithm g in (5.4.7). Its rationale is to transform a problem of estimation into a problem of optimisation. Let us consider

1. a density distribution $p_{\mathbf{z}}(z, \theta)$ which depends on a parameter $\theta \in \Theta$,
2. a dataset $D_N = \{z_1, z_2, \ldots, z_N\}$ i.i.d. drawn from this distribution.

According to (3.5.54), the joint probability density of the i.i.d. dataset is the product

$$p_{\mathbf{D}_N}(D_N, \theta) = \prod_{i=1}^{N} p_{\mathbf{z}}(z_i, \theta) = L_N(\theta) \tag{5.8.19}$$

where for a fixed D_N, $L_N(\cdot)$ is a function of θ and is called the *empirical likelihood* of θ given D_N.

The principle of maximum likelihood was first used by Lambert around 1760 and by D. Bernoulli about 13 years later. It was detailed by Fisher in 1920. The idea is simple: given an unknown parameter θ and a sample data D_N, the maximum likelihood estimate $\hat{\theta}$ is the value for which the empirical likelihood $L_N(\theta)$ has a maximum

$$\hat{\theta}_{\text{ml}} = \arg\max_{\theta \in \Theta} L_N(\theta)$$

The estimator $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is called the maximum likelihood estimator (m.l.e.). In practice, it is usual to consider the log-likelihood $l_N(\theta)$ instead of $L_N(\theta)$. Since $\log(\cdot)$ is a monotone function, we have

$$\hat{\theta}_{\text{ml}} = \arg\max_{\theta \in \Theta} L_N(\theta) = \arg\max_{\theta \in \Theta} \log(L_N(\theta)) = \arg\max_{\theta \in \Theta} l_N(\theta) \tag{5.8.20}$$

The likelihood function quantifies the relative abilities of the various parameter values to *explain* the observed data. The principle of m.l. is that the value of the parameter under which the obtained data would have had highest probability of arising must be intuitively our best estimator of θ. In other terms the likelihood can be considered a measure of how plausible the parameter values are in light of the data. Note however that the likelihood function is NOT a probability function: for instance, in general, it does not integrate to 1 (with respect to θ). In terms of conditional probability, $L_N(\theta)$ represents the probability of the observed dataset given θ and not the probability of θ (which is not a r.v. in the frequentist approach) given D_N.

Example

Consider a binary variable (e.g. a coin tossing) which takes $z = 15$ times the value 1 (e.g. "Tail") in $N = 40$ trials. Suppose that the probabilistic model underlying the data is Binomial (Section C.1.2) with an unknown probability $\theta = p$. We want to estimate the unknown parameter $\theta = p \in [0, 1]$ on the basis of the empirical evidence from the N trials. The likelihood $L(p)$ is a function of (only) the unknown parameter p. By applying the maximum likelihood technique we have

$$\hat{\theta}_{\text{ml}} = \hat{p} = \arg\max_{p} L_N(p) = \arg\max_{p} \binom{N}{z} p^z (1-p)^{(N-z)} = \arg\max_{p} \binom{40}{15} p^{15} (1-p)^{(25)}$$

Figure 5.4 plots $L(p)$ versus $p \in [0, 1]$ (R script `ml_bin.R`). The most likely value of p is the value where $L(\cdot)$ attains its maximum. According to Figure 5.4 this value is $\hat{p} = z/N$. The log-likelihood for this model is

$$l_N(p) = \log L_N(p) = \log \binom{N}{z} + z \log(p) + (N - z) \log(1 - p) = \\ = \log \binom{40}{15} + 15 \log p + 25 \log(1 - p)$$

The reader can analytically find the maximum of this function by differentiating $l(p)$ with respect to p.

•

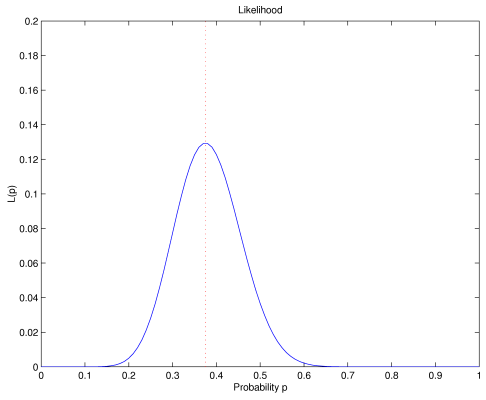


Figure 5.4: Likelihood function

5.8.1 Maximum likelihood computation

In many situations the log-likelihood $l_N(\theta)$ is particularly well behaved in being continuous with a single maximum away from the extremes of the range of variation of θ. Then $\hat{\theta}_{\text{ml}}$ is obtained simply as the solution of

$$\frac{\partial l_N(\theta)}{\partial \theta} = 0$$

subject to

$$\frac{\partial^2 l_N(\theta)}{\partial \theta^2}\Big|_{\hat{\theta}_{\text{ml}}} < 0$$

to ensure that the identified stationary point is a maximum.

5.8.2 Maximum likelihood in the Gaussian case

Let D_N be a random sample from the r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$. It is possible to derive analytically the expression of the maximum likelihood estimators of the mean and variance of $\mathbf{z}$. According to (5.8.19), the likelihood of the N observations is

$$L_N(\mu, \sigma^2) = \prod_{i=1}^{N} p_{\mathbf{z}}(z_i, \mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left[\frac{-(z_i - \mu)^2}{2\sigma^2} \right]$$

and the log-likelihood is

$$l_N(\mu, \sigma^2) = \log L_N(\mu, \sigma^2) = \log \left[\prod_{i=1}^{N} p_{\mathbf{z}}(z_i, \mu, \sigma^2) \right] =$$

$$= \sum_{i=1}^{N} \log p_{\mathbf{z}}(z_i, \mu, \sigma^2) = -\frac{\sum_{i=1}^{N} (z_i - \mu)^2}{2\sigma^2} + N \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)$$

Note that, for a given σ, maximising the log-likelihood is equivalent to minimising the sum of squares of the difference between z_i and the mean. Taking the derivatives with respect to μ and σ^2 and setting them equal to zero, we obtain

$$\hat{\mu}_{\text{ml}} = \frac{\sum_{i=1}^{N} z_i}{N} = \hat{\mu} \tag{5.8.21}$$

$$\hat{\sigma}^2_{\text{ml}} = \frac{\sum_{i=1}^{N} (z_i - \hat{\mu}_{\text{ml}})^2}{N} \neq \hat{\sigma}^2 \tag{5.8.22}$$

Note that the m.l. estimator (5.8.21) of the mean coincides with the sample average (5.3.4) but that the m.l. estimator (5.8.22) of the variance differs from the sample variance (5.3.5) in terms of the denominator.

In the multivariate Normal case, where $\mathbf{z}$ is a vector with $[n, 1]$ mean μ and $[n, n]$ covariance matrix Σ, the maximum likelihood estimators are

$$\hat{\mu}_{\mathrm{ml}} = \frac{\sum_{i=1}^{N} z_i}{N} \tag{5.8.23}$$

$$\hat{\Sigma}_{\mathrm{ml}} = \frac{\sum_{i=1}^{N} (z_i - \hat{\mu}_{\mathrm{ml}})(z_i - \hat{\mu}_{\mathrm{ml}})^T}{N} \tag{5.8.24}$$

where z_i and $\hat{\mu}$ are $[n, 1]$ vectors.

Exercise

- Let $\mathbf{z} \sim \mathcal{U}(0, M)$ follow a uniform distribution and $F_{\mathbf{z}} \to D_N = \{z_1, \ldots, z_N\}$. Find the maximum likelihood estimator of M.
- Let $\mathbf{z}$ have a Poisson distribution, i.e.

$$p_{\mathbf{z}}(z, \lambda) = \frac{e^{-\lambda} \lambda^z}{z!}$$

  If $F_{\mathbf{z}}(z, \lambda) \to D_N = \{z_1, \ldots, z_N\}$, find the m.l.e. of λ

•

In case of generic distributions $F_{\mathbf{z}}$ computational difficulties may arise: for example in some cases no explicit solution might exist for $\partial l_N(\theta)/\partial\theta = 0$. Iterative numerical methods must be used in this case. The computational cost becomes heavier if we consider a vector of parameters instead of a scalar θ or when there are several relative maxima of the function l_N.

Another complex situation occurs when $l_N(\theta)$ is discontinuous, or have a discontinuous first derivative, or a maximum at an extremal point.

R script

Suppose we know the analytical form of a one dimensional function $f(x) : I \to \mathbb{R}$ but not the analytical expression of its extreme points. In this case numerical optimisation methods can be applied. The implementation of some continuous optimisation routines is available in the R statistical tool.

Consider for example the function $f(x) = (x - 1/3)^2$ and $I = [0, 1]$. The value of the point x where f takes a minimum value can be approximated numerically by this set of R commands

```
f <- function (x,a) (x-a)^2
xmin <- optimize(f, c(0, 1), tol = 0.0001, a = 1/3)
xmin
```

These routines may be applied to solve the problem of maximum likelihood estimation which is nothing more than a particular case of optimisation problem. Let D_N be a random sample drawn from $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$. The negative log-likelihood function of the N observations can be written in R by

```
eml <- function(m,D,var) {
         N<- length(D)
         Lik<-1
```

```
    for (i in 1:N)
     Lik<-Lik*dnorm(D[i],m,sqrt(var))
    -log(Lik)
}
```

and the numerical minimisation of $-l_N(\mu, s^2)$ for a given $\sigma = s$ in the interval $I = [-10, 10]$ can be written in R as

```
xmin<-optimize( eml,c(-10,10),D=DN,var=s)
```

In order to run the above code and compute numerically the m.l. solution we invite the reader to run the R script `emp_ml.R`.

•

5.8.3 Cramer-Rao lower bound

Assume that θ is a scalar parameter, that the first two derivatives of $L_N(\theta)$ with respect to θ exist for all θ and that certain operations of integration and differentiation may be interchanged. Let $\hat{\boldsymbol{\theta}}$ be an unbiased estimator of θ and $l_N(\theta) = \log_e[L_N(\theta)]$. Suppose that the regularity condition

$$E\left[\frac{\partial \boldsymbol{l}_N(\theta)}{\partial \theta}\right] = 0 \tag{5.8.25}$$

holds where the quantity $\partial l(\theta)/\partial\theta$ is called *score*. The Cramer-Rao bound is a lower bound to the variance of the estimator $\hat{\boldsymbol{\theta}}$ which states that

$$\text{Var}\left[\hat{\boldsymbol{\theta}}\right] \geq \frac{1}{E\left[\left(\frac{\partial \boldsymbol{l}_N(\theta)}{\partial \theta}\right)^2\right]} = -\frac{1}{NE\left[\left(\frac{\partial^2 \boldsymbol{l}_N(\theta)}{\partial \theta^2}\right)\right]} = \frac{1}{I_N}$$

where the denominator term I_N is known as the Fisher information. Note that $\frac{\partial^2 l_N(\theta)}{\partial \theta^2}$ is the second derivative of $l_N(\cdot)$ and, as such, it defines the curvature of the log-likelihood function. At the maximum $\hat{\theta}$, the second derivative takes a negative value. Also, the larger its absolute value the larger is the curvature around the function peak and then the lower is the uncertainty about the m.l. estimation [143].

An estimator having a variance as low as $1/I_N$ is called a *Minimum Variance Bound (MVB) estimator.*

Example

Consider a r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known and the unknown parameter is $\theta = \mu$. Let us consider the bound on the variance of the estimator (5.8.21). Since

$$\frac{\partial \log p(z, \theta)}{\partial \theta} = \frac{z - \theta}{\sigma^2}$$

$$\frac{\partial^2 \log p(z, \theta)}{\partial \theta^2} = -\frac{1}{\sigma^2}$$

It follows that

$$\text{Var}\left[\hat{\boldsymbol{\theta}}\right] \geq \frac{1}{\frac{N}{\sigma^2}} = \frac{\sigma^2}{N}$$

From (5.5.10) it derives then that the m.l. estimator (5.8.21) of the mean μ is minimum variance.

•

5.8.4 Properties of m.l. estimators

Under the (strong) assumption that the probabilistic model structure is known, the maximum likelihood technique features the following properties:

- $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is *asymptotically unbiased* but usually biased in small-size samples (e.g. $\hat{\sigma}^2_{\text{ml}}$ in (5.8.22)).
- $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is consistent.
- If $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is the m.l.e. of θ and $\gamma(\cdot)$ is a *monotone* function then $\gamma(\hat{\boldsymbol{\theta}}_{\text{ml}})$ is the m.l.e. of $\gamma(\theta)$.
- If $\gamma(\cdot)$ is a non monotonic function, then even if $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is an unbiased estimator of θ, the m.l.e. $\gamma(\hat{\boldsymbol{\theta}}_{\text{ml}})$ of $\gamma(\theta)$ is usually biased.
- the variance of $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is often difficult to determine. For large-size samples we can use as approximation

$$\left(-E\left[\frac{\partial^2 \boldsymbol{l}_N}{\partial\theta^2}\right]\right)^{-1} \text{ or } \left(-\frac{\partial^2 l_N}{\partial\theta^2}\Big|_{\hat{\theta}_{\text{ml}}}\right)^{-1}$$

- $\hat{\boldsymbol{\theta}}_{\text{ml}}$ is asymptotically normally distributed, that is

$$\hat{\boldsymbol{\theta}}_{\text{ml}} \sim \mathcal{N}(\theta, [I_N(\theta)]^{-1}), \qquad N \to \infty$$

5.9 Interval estimation

Unlike point estimation which is based on a one-to-one mapping from the space of data to the space of parameters, interval estimation maps D_N to an interval of Θ. A point estimator is a function which, given a dataset D_N generated from $F_{\mathbf{z}}(z,\theta)$, returns an estimate of θ. An *interval estimator* is a transformation which, given a dataset D_N, returns an interval estimate $[\underline{\theta}, \bar{\theta}]$ of θ. While an estimator is a random variable, an interval estimator is a random interval. Let $\underline{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ be the random lower and the upper bounds respectively. While an interval either contains or not a certain value, a random interval has a certain probability of containing a value. Suppose that

$$\text{Prob}\left\{\underline{\boldsymbol{\theta}} \leq \theta \leq \bar{\boldsymbol{\theta}}\right\} = 1-\alpha \qquad \alpha \in [0,1] \tag{5.9.26}$$

then the random interval $[\underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}]$ is called a $100(1-\alpha)\%$ confidence interval of θ. If (5.9.26) holds, we expect that by repeating the sampling of D_N and the construction of the confidence interval many times, our confidence interval will contain the true θ at least $100(1-\alpha)\%$ of the time. Notice, however, that being θ a fixed unknown value, at each realisation D_N the interval $[\underline{\theta}, \bar{\theta}]$ either contains or not the true θ. Therefore, from a frequentist perspective, it is erroneous to think that $1-\alpha$ is the probability of θ belonging to the interval $[\underline{\theta}, \bar{\theta}]$ computed for a given D_N. In fact, $1-\alpha$ is not the probability of the event $\theta \in [\underline{\theta}, \bar{\theta}]$ (since θ is fixed) but the probability that the interval estimation procedure returns a (random) interval $[\underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}]$ containing θ.

While a point estimator is characterised by bias and variance (Section 5.5), an interval estimator is characterised by its *endpoints* $\underline{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ (or its width) and by its *confidence* α. In Figure 5.3 we used an analogy between point estimation and dart game to illustrate the bias/variance notions. In the case of interval estimation the best analogy is provided by the *horseshoes* game[4] (Figure 5.5). A horseshoe

[4] `https://en.wikipedia.org/wiki/Horseshoes`

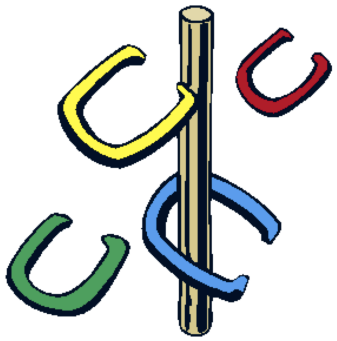


Figure 5.5: Horseshoes game as an analogy of interval estimation

player is like an interval estimator and her interval estimation corresponds to the tossing of a horseshoe. The horseshoe width corresponds to the interval size and the probability of encircling the stake corresponds to the confidence α.

5.9.1 Confidence interval of μ

Consider a random sample D_N of a r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. Suppose we want to estimate μ with the estimator $\hat{\boldsymbol{\mu}}$. From Section 5.7 we have that $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu, \sigma^2/N)$ is Gaussian distributed. From (3.4.47) it follows that

$$\frac{\hat{\boldsymbol{\mu}} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$$

and consequently, according to the Definition 4.7

$$\text{Prob}\left\{-z_{\alpha/2} \leq \frac{\hat{\boldsymbol{\mu}} - \mu}{\sigma/\sqrt{N}} \leq z_{\alpha/2}\right\} = 1 - \alpha \tag{5.9.27}$$

$$\text{Prob}\left\{\hat{\boldsymbol{\mu}} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \hat{\boldsymbol{\mu}} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha \tag{5.9.28}$$

where z_α is the upper critical point of the standard Gaussian distribution. It follows that $\underline{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}} - z_\alpha\sigma/\sqrt{N}$ is a lower $1-\alpha$ confidence bound for μ while $\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}} + z_\alpha\sigma/\sqrt{N}$ is an upper $1-\alpha$ confidence bound for μ. By varying α we can vary the width and the confidence of the interval.

Example

Let $\mathbf{z} \sim \mathcal{N}(\mu, 0.01)$ and $D_N = \{10, 11, 12, 13, 14, 15\}$. We want to estimate the confidence interval of μ with level $\alpha = 0.1$. Since $N = 6$, $\hat{\mu} = 12.5$, and

$$\epsilon = z_{\alpha/2}\sigma/\sqrt{N} = 1.645 \cdot 0.01/\sqrt{6} = 0.0672$$

the 90% confidence interval for the given D_N is

$$\{\mu : |\hat{\boldsymbol{\mu}} - \mu| \leq \epsilon\} = \{12.5 - 0.0672 \leq \mu \leq 12.5 + 0.0672\}$$

•

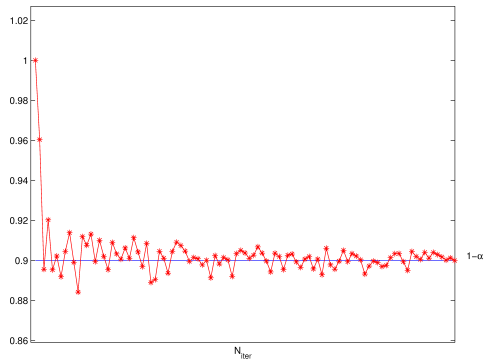


Figure 5.6: Fraction of times that the interval of confidence contains the parameter μ vs. the number of repetitions for $\alpha = 0.1$

R script

The R script `confidence.R` allows the test of the formula (5.9.27) by simulation. The user sets μ, σ, N, α and a number of iterations N_{iter}.

The script generates N_{iter} times $D_N \sim \mathcal{N}(\mu, \sigma^2)$ and computes $\hat{\mu}$. The script returns the percentage of times that

$$\hat{\mu} - \frac{z_{\alpha/2}\sigma}{\sqrt{N}} < \mu < \hat{\mu} + \frac{z_{\alpha/2}\sigma}{\sqrt{N}}$$

This percentage versus the number of iterations is plotted in Figure 5.6 (R script `confidence.R`). We can easily check that this percentage converges to $100(1-\alpha)\%$ for $N_{\text{iter}} \to \infty$.

•

Consider now the interval of confidence of μ when the variance σ^2 is not known. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}^2$ be the estimators of μ and σ^2 computed on the basis of the i.i.d. dataset D_N. From Section 5.7, it follows that

$$\frac{\hat{\mu} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{N}}} \sim \mathcal{T}_{N-1}$$

Analogously to (5.9.28) we have

$$\text{Prob}\left\{\hat{\boldsymbol{\mu}} - t_{\alpha/2}\frac{\sigma}{\sqrt{N}} \leq \mu \leq \hat{\boldsymbol{\mu}} + t_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha \tag{5.9.29}$$

where t_α is the upper critical point of the Student distribution.

Example

Let $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 unknown and $D_N = \{10, 11, 12, 13, 14, 15\}$. We want to estimate the confidence region of μ with level $\alpha = 0.1$. We have $\hat{\mu} = 12.5$, $\hat{\sigma}^2 = 3.5$. According to (5.9.29) we have

$$\epsilon = t_{\{\alpha/2, N-1\}}\hat{\sigma}/\sqrt{N} = 2.015 * 1.87/\sqrt{6} = 1.53$$

The $(1-\alpha)$ confidence interval of μ is

$$\hat{\mu} - \epsilon < \mu < \hat{\mu} + \epsilon$$

•

Example

We want to estimate θ, the proportion of people who support the politics of Mr. Berlusconi amongst a very large population. We want to define how many interviews are necessary to have a confidence interval of 6% width with a significance of 5%. We interview N persons and estimate θ as

$$\hat{\theta} = \frac{x_1 + \cdots + x_N}{N} = \frac{S}{N}$$

where $x_i = 1$ if the ith person supports Berlusconi and $x_i = 0$ otherwise. Note that $\mathbf{S}$ is a binomial variable. We have

$$E[\hat{\boldsymbol{\theta}}] = \theta, \qquad \text{Var}\left[\hat{\boldsymbol{\theta}}\right] = \text{Var}\left[S/N\right] = \frac{N(\theta)(1-\theta)}{N^2} = \frac{\theta(1-\theta)}{N} \leq \frac{1}{4N}$$

If we approximate the distribution of $\hat{\boldsymbol{\theta}}$ by $\mathcal{N}(\theta, \frac{\theta(1-\theta)}{N})$ it follows that $\frac{\hat{\boldsymbol{\theta}}-\theta}{\sqrt{\theta(\theta-1)N}} \sim \mathcal{N}(0,1)$. The following relation holds

$$\begin{aligned}
&\text{Prob}\left\{\hat{\boldsymbol{\theta}} - 0.03 \leq \theta \leq \hat{\boldsymbol{\theta}} + 0.03\right\} = \\
&\qquad \text{Prob}\left\{-\frac{0.03}{\sqrt{\theta(1-\theta)/N}} \leq \frac{\hat{\boldsymbol{\theta}} - \theta}{\sqrt{\theta(1-\theta)/N}} \leq \frac{0.03}{\sqrt{\theta(1-\theta)/N}}\right\} = \\
&\qquad\qquad \Phi\left(\frac{0.03}{\sqrt{\theta(1-\theta)/N}}\right) - \Phi\left(-\frac{0.03}{\sqrt{\theta(1-\theta)/N}}\right) \geq \\
&\qquad\qquad\qquad \geq \Phi(0.03\sqrt{4N}) - \Phi(-0.03\sqrt{4N})
\end{aligned}$$

In order to have this probability to be at least 0.95 we need $0.03\sqrt{4N} \geq 1.96$ or equivalently $N \geq 1068$.

•

5.10 Combination of two estimators

Consider two unbiased estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ of the same parameter θ

$$E[\hat{\boldsymbol{\theta}}_1] = \theta \qquad E[\hat{\boldsymbol{\theta}}_2] = \theta$$

having equal and non zero variance

$$\text{Var}\left[\hat{\boldsymbol{\theta}}_1\right] = \text{Var}\left[\hat{\boldsymbol{\theta}}_2\right] = v$$

and being uncorrelated, i.e. $\text{Cov}[\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2] = 0$. Let $\hat{\boldsymbol{\theta}}_{\text{cm}}$ be the combined estimator

$$\hat{\boldsymbol{\theta}}_{\text{cm}} = \frac{\hat{\boldsymbol{\theta}}_1 + \hat{\boldsymbol{\theta}}_2}{2}$$

This estimator has the nice properties of being unbiased

$$E[\hat{\boldsymbol{\theta}}_{\text{cm}}] = \frac{E[\hat{\boldsymbol{\theta}}_1] + E[\hat{\boldsymbol{\theta}}_2]}{2} = \theta \tag{5.10.30}$$

and with a smaller variance than the original estimators

$$\text{Var}\left[\hat{\boldsymbol{\theta}}_{\text{cm}}\right] = \frac{1}{4}\text{Var}\left[\hat{\boldsymbol{\theta}}_1 + \hat{\boldsymbol{\theta}}_2\right] = \frac{\text{Var}\left[\hat{\boldsymbol{\theta}}_1\right] + \text{Var}\left[\hat{\boldsymbol{\theta}}_2\right]}{4} = \frac{v}{2} \tag{5.10.31}$$

This trivial computation shows that the simple average of two unbiased estimators with a non zero variance returns a combined estimator with reduced variance.

5.10.1 Combination of m estimators

Here, we report the general formula of the linear combination of a number m of estimators [175, 177]. Assume we want to estimate the unknown parameter θ by combining a set of m estimators $\{\hat{\boldsymbol{\theta}}_j\}$, $j = 1, \ldots, m$. Let

$$E[\hat{\boldsymbol{\theta}}_j] = \mu_j \quad \text{Var}\left[\hat{\boldsymbol{\theta}}_j\right] = v_j \quad \text{Bias}[\hat{\boldsymbol{\theta}}_j] = b_j$$

be the expected values, the variances and the bias of the m estimators, respectively.

We are interested in estimating θ by forming a linear combination

$$\hat{\boldsymbol{\theta}}_{\text{cm}} = \sum_{j=1}^{m} w_j \hat{\boldsymbol{\theta}}_j = w^T \hat{\boldsymbol{\theta}} \tag{5.10.32}$$

where $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_m]^T$ is the vector of estimators and $w = [w_1, \ldots, w_m]^T$ is the weighting vector.

The mean-squared error of the combined system is

$$\begin{aligned}\text{MSE} = E[(\hat{\boldsymbol{\theta}}_{\text{cm}} - \theta)^2] &= E[(w^T\hat{\boldsymbol{\theta}} - E[w^T\hat{\boldsymbol{\theta}}])^2] + (E[w^T\hat{\boldsymbol{\theta}}] - \theta)^2 \\ &= E[(w^T(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}]))^2] + (w^T\mu - \theta)^2 = \\ &= w^T\Omega w + (w^T\mu - \theta)^2\end{aligned}$$

where Ω is a $[m \times m]$ covariance matrix whose ij^{th} term is

$$\Omega_{ij} = E[(\hat{\boldsymbol{\theta}}_i - \mu_i)(\hat{\boldsymbol{\theta}}_j - \mu_j)]$$

and $\mu = (\mu_1, \ldots, \mu_m)^T$ is the vector of expected values. Note that the MSE error has a variance term (dependent on the covariance of the single estimators) and a bias term (dependent on the bias of the single estimators).

5.10.1.1 Linear constrained combination

A commonly used constraint is

$$\sum_{j=1}^{m} w_j = 1, \quad w_j \geq 0, \quad j = 1, \ldots, m \tag{5.10.33}$$

This means that the combined estimator is unbiased if the individual estimators are unbiased. Let us write w as

$$w = (u^T g)^{-1} g$$

where $u = (1, \ldots, 1)^T$ is an m-dimensional vector of ones, $g = (g_1, \ldots, g_m)^T$ and $g_j > 0, \forall j = 1, \ldots, m$.

The constraint can be enforced in minimising the MSE by using the Lagrangian function

$$L = w^T\Omega w + (w^T\mu - \theta)^2 + \lambda(w^T u - 1)$$

with λ Lagrange multiplier.

The optimum is achieved if we set

$$g^* = [\Omega + (\mu - \theta u)(\mu - \theta u)^T]^{-1} u$$

With unbiased estimators ($\mu = \theta$) we obtain

$$g^* = \Omega^{-1} u$$

and with uncorrelated estimators

$$g_j^* = \frac{1}{\text{v}_j} \qquad j = 1, \ldots, m \tag{5.10.34}$$

This means that the optimal term g_j^* of each estimator is inversely proportional to its own variance.

5.11 Testing hypothesis

Hypothesis testing is together with estimation a major area of statistical inference. A *statistical hypothesis* is an assertion or conjecture about the distribution of one or more random variables. A *test* of a statistical hypothesis is a rule or procedure for deciding whether to reject the assertion on the basis of the observed data. The basic idea is formulate some statistical hypothesis and look to see whether the data provides any evidence to reject the hypothesis. Examples of hypothesis tests follow:

- Consider the model of the traffic in the boulevard. Suppose that the measures of the inter-arrival times are $D_N = \{10, 11, 1, 21, 2, \dots\}$ seconds. Can we say that the mean inter-arrival time θ is different from 10?
- We want to know the effect of a drug on rats' survival to cancer. We randomly divide some rats in two groups and we administrate a drug only to one of them. Is the survival rate of the groups the same?
- Consider the grades of two different school sections. Section A had $\{15, 10, 12, 19, 5, 7\}$. Section B had $\{14, 11, 11, 12, 6, 7\}$. Can we say that Section A had better grades than Section B?
- Consider two protein coding genes and their expression levels in a cell. Are the two genes *differentially expressed* ?

A statistical test is a procedure that aims to answer such questions.

5.11.1 Types of hypothesis

We start by declaring the *working (basic, null) hypothesis* H to be tested, in the form $\theta = \theta_0$ or $\theta \in \omega \subset \Theta$, where θ_0 or ω are *given*.

The hypothesis can be

simple: this means that it fully specifies the distribution of the r.v. $\mathbf{z}$.

composite: this means that it partially specifies the distribution of $\mathbf{z}$.

For example if D_N is a random sample of size N drawn from $\mathcal{N}(\mu, \sigma^2)$ the hypothesis $H : \mu = \mu_0, \sigma = \sigma_0$, (with μ_0 and σ_0 known values) is simple while the hypothesis $H : \mu = \mu_0$ is composite since it leaves open the value of σ in $(0, \infty)$.

5.11.2 Types of statistical test

Suppose we have sampled a dataset $D_N = \{z_1, \dots, z_N\}$ from a distribution $F_{\mathbf{z}}$ and we have declared a null hypothesis H about F. The three most common types of statistical test are:

Pure significance test: data D_N are used to assess the inferential evidence against H.

Significance test: the inferential evidence against H is used to judge whether H is inappropriate. This test returns a decision rule for rejecting or not rejecting H.

Hypothesis test: data D_N are used to assess the hypothesis H against a specific alternative hypothesis $\bar{H}$.This test returns a rule for rejecting H in favour of $\bar{H}$.

The three tests will be discussed in the following sections.

5.11.3 Pure significance test

Consider a simple null hypothesis H. Let $t(\mathbf{D}_N)$ be a statistic (i.e. a function of the dataset) such that the larger its value the more it casts doubt on H. The quantity $t(\mathbf{D}_N)$ is called *test statistic* or *discrepancy measure*. Suppose that the distribution of $t(\mathbf{D}_N)$ under H is known. This is possible since the function $t(\cdot)$ is fixed by the user and the simple hypothesis H entirely specifies the distribution of $\mathbf{z}$ and consequently the distribution of $t(\mathbf{D}_N)$. Let $t_N = t(D_N)$ the observed value of t calculated on the basis of the sample data D_N. Let us define the *p-value* quantity as

$$p = \text{Prob}\{t(\mathbf{D}_N) > t_N | H\} \tag{5.11.35}$$

i.e. the probability of observing a statistic greater than t_N if the hypothesis H were true. Note that in the expression (5.11.35), the term $t(\mathbf{D}_N)$ is a random variable having a known distribution, while t_N is a value computed on the basis of the observed dataset.

If the p quantity is small then the sample data D_N are highly inconsistent with H, and p (*significance probability or significance level*) is the measure of such inconsistency. If p is small, then either a rare event has occurred or perhaps H is not true. In other terms, if H were true, the quantity p would be the proportion of situations where we would observe a degree of inconsistency, at least to the extent represented by t_N. The smaller the p-value, the stronger the evidence against H[5].

Note that p depends on D_N since different D_N would yield different values of t_N and consequently different values of $p \in [0, 1]$. Moreover, it can be shown that, if the null hypothesis is true, the p-value has a Uniform $U[0, 1]$ distribution. Also, in a frequentist perspective, we cannot say that p is the probability that H is true but rather that p *is the probability that the dataset* D_N *is observed given that* H *is true.*

5.11.4 Tests of significance

The test of significance proposes the following decision rule: if p is less than some stated value α, we reject H. Once a *critical level* α is chosen, and the dataset D_N is observed, the rule rejects H at level α if

$$P\{t(\mathbf{D}_N) > t_\alpha | H) = \alpha \tag{5.11.36}$$

This is equivalent to choosing some *critical value* t_α and to reject H if $t_N > t_\alpha$. This implies the existence of two regions in the space of sample data:

critical region: this is the set of values of D_N

$$S_0 = \{D_N : t(D_N) > t_\alpha\}$$

such that if $D_N \in S_0$, we reject the null hypothesis H.

non-critical region: this is the set of values of D_N such that there is no reason to reject H on the basis of the level-α test.

The principle is that we will accept H unless what we observed has a too small probability of happening when H is true. The upper bound of this probability is α, i.e. the significance level α is the highest p-value for which we reject H. Note that the p-value changes with the observed data (i.e. it is a random variable) while α is a level fixed by the user.

[5] It is common habit in life-sceince research to consider a p-value smaller than 0.05 (0.01) a (very) strong evidence against H

Example

Let D_N consist of N independent observations of $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, with known variance σ^2. We want to test the hypothesis $H : \mu = \mu_0$ with μ_0 known. Consider as test statistic the quantity $t(\mathbf{D}_N) = |\hat{\boldsymbol{\mu}} - \mu_0|$ where $\hat{\boldsymbol{\mu}}$ is the sample average estimator. If H is true we know from Section 5.4 that $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu_0, \sigma^2/N)$. Let us calculate the value $t(D_N) = |\hat{\mu} - \mu_0|$ and fix a significance level $\alpha = 10\%$. This means that the decision rule needs the definition of the value t_α such that

$$\text{Prob}\{t(\mathbf{D}_N) > t_\alpha | H\} = \text{Prob}\{|\hat{\boldsymbol{\mu}} - \mu_0| > t_\alpha | H\} = \\ \text{Prob}\{(\hat{\boldsymbol{\mu}} - \mu_0 > t_\alpha) \cup (\hat{\boldsymbol{\mu}} - \mu_0 < -t_\alpha) | H\} = 0.1$$

For a Normal variable $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, we have that

$$\text{Prob}\{|\mathbf{z} - \mu| > 1.645\sigma\} = \text{Prob}\left\{\frac{|\mathbf{z} - \mu|}{\sigma} > 1.645\right\} = 2 * 0.05$$

It follows that being $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\mu_0, \sigma^2/N)$

$$\text{Prob}\left\{|\hat{\boldsymbol{\mu}} - \mu_0| > 1.645\sigma/\sqrt{N}\right\} = 0.05 + 0.05 = 0.1$$

and consequently

$$t_\alpha = 1.645\sigma/\sqrt{N} \tag{5.11.37}$$

The critical region is

$$S_0 = \left\{D_N : |\hat{\mu} - \mu_0| > 1.645\sigma/\sqrt{N}\right\}$$

•

Example

Suppose that $\sigma = 0.1$ and that we want to test if $\mu = \mu_0 = 10$ with a significance level 10%. Let $N = 6$ and $D_N = \{10, 11, 12, 13, 14, 15\}$. From the dataset we compute

$$\hat{\mu} = \frac{10 + 11 + 12 + 13 + 14 + 15}{6} = 12.5$$

and

$$t(D_N) = |\hat{\mu} - \mu_0| = 2.5$$

Since according to (5.11.37) $t_\alpha = 1.645 * 0.1/\sqrt{6} = 0.0672$, and $t(D_N) > t_\alpha$, the observations D_N are in the critical region. The conclusion is: *the hypothesis* $H : \mu = 10$ *is rejected* and the probability that we are making an error by rejecting H is smaller than 0.1.

•

5.11.5 Hypothesis testing

So far we have dealt with single hypothesis tests. Let us now consider two mutually exclusive hypothesis: H and $\bar{H}$. Suppose we have a dataset $\{z_1, \ldots, z_N\} \sim F$ drawn from a distribution F. On the basis of this dataset, one hypothesis will be *accepted* and the other one *rejected*. In this case, given the stochastic setting, two type of errors are possible.

Type I error. This is the kind of error we make when *we reject H but H is true.*
For a given critical level t_α the probability of making this error is

$$\text{Prob}\{t(\mathbf{D}_N) > t_\alpha | H\} = \alpha \tag{5.11.38}$$

Type II error. This is the kind of error we make when *we accept H and H is false*. In order to define this error, we are forced to declare an alternative hypothesis $\bar{H}$ as a formal definition of what is meant by H being *false*. The probability of type II error is

$$\text{Prob}\left\{t(\mathbf{D}_N) \leq t_\alpha | \bar{H}\right\} \tag{5.11.39}$$

that is the probability that the test leads to acceptance of H when in fact $\bar{H}$ holds.

Note that

- when the alternative hypothesis is composite, there could be no unique Type II error.
- although H and $\bar{H}$ are complementary events, the quantity (5.11.39) cannot be derived from (5.11.38) (see Equation (3.1.23)).

Example

In order to better illustrate these notions, let us consider the analogy with a murder trial, where the suspect is Mr. Bean. The null hypothesis H is "Mr. Bean is innocent". The dataset is the amount of evidence collected by the police against Mr. Bean. The Type I error is the error that we make if, Mr. Bean being innocent, we send him to death-penalty. The Type II error is the error that we make if, being Mr. Bean guilty, we acquit him. Note that the two hypotheses have different philosophical status (asymmetry). H is a conservative hypothesis, not to be rejected unless evidence against Mr Bean's innocence is clear. This means that a type I error is *more serious* than a type II error (*benefit of the doubt*).

•

Example

Let us consider a professor who has to decide on the basis of empirical evidence whether a student copied or not during a class test. The null hypothesis H is that the student is honest. The alternative hypothesis $\bar{H}$ is that the student has cheated. Let the empirical evidence t_N be represented by the number of lines of the classwork that a student shares with one of his classmates.

Suppose that a student passes if the professor thinks she has not copied ($t_N \leq t_\alpha = 2$) while he fails otherwise.

•

5.11.6 The hypothesis testing procedure

In general terms a hypothesis testing procedure can be decomposed in the following steps:

1. Declare the null and the alternative hypothesis
2. Choose the numeric value α of the type I error (e.g. the risk I want to run when I reject the null hypothesis).
3. Define a test statistic.
4. Determine the critical value t_α of the test statistic that leads to a rejection of H according to the Type I error defined in Step 2.

5. Among the set of tests of level α, choose the test that minimises the probability of type II error.

6. Obtain the data and determine whether the observed value of the test statistic leads to an acceptation or rejection of H.

Note that a number of tests, having a different type II error, can guarantee the same type I error. An appropriate choice of test as a function of the type II error is therefore required and will be discussed in the following section.

5.11.7 Choice of test

The choice of test and consequently the choice of the partition $\{S_0, S_1\}$ is based on two steps

1. Define a significance level α, that is the probability of type I error (or the probability of incorrectly rejecting H)

$$\text{Prob}\{\text{reject } H|H\} = \text{Prob}\{\mathbf{D}_N \in S_0|H\} = \alpha$$

2. Among the set of tests $\{S_0, S_1\}$ of level α, choose the test that minimises the probability of type II error

$$\text{Prob}\left\{\text{accept } H|\bar{H}\right\} = \text{Prob}\left\{\mathbf{D}_N \in S_1|\bar{H}\right\}$$

that is the probability of incorrectly accepting H. This is equivalent to maximising the *power of the test*

$$\text{Prob}\left\{\text{reject } H|\bar{H}\right\} = \text{Prob}\left\{\mathbf{D}_N \in S_0|\bar{H}\right\} = 1 - \text{Prob}\left\{\mathbf{D}_N \in S_1|\bar{H}\right\}$$

which is the probability of *correctly* rejecting H. Note that for a given significance level, the higher the power, the better !

Example

In order to reason about the Type II error, let us consider an r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, where σ is known and a set of N iid observations are given. We want to test the null hypothesis $\mu = \mu_0 = 0$, with $\alpha = 0.1$ Consider three different tests and the associated critical regions S_0

1. $|\hat{\mu} - \mu_0| > 1.645\sigma/\sqrt{N}$

2. $\hat{\mu} - \mu_0 > 1.282\sigma/\sqrt{N}$ (Figure 5.7)

3. $|\hat{\mu} - \mu_0| < 0.126\sigma/\sqrt{N}$ (Figure 5.8)

Assume that the area blackened in Figure (5.7) equals the area blackened in Figure (5.8). For all these tests $\text{Prob}\{\mathbf{D}_N \in S_0|H\} \leq \alpha$, hence the significance level (i.e. Type I error) is the same. However if $\bar{H} : \mu_1 = 10$ the type II error of the three tests is significantly different. Which test is the best one, that is the one which guarantees the lowest Type II error?

•

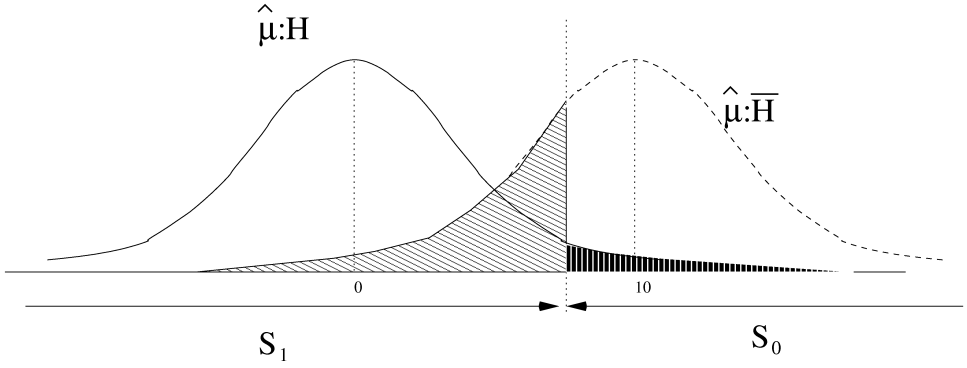


Figure 5.7: On the left: distribution of the test statistic $\hat{\boldsymbol{\mu}}$ if $H : \mu_0 = 0$ is true. On the right: distribution of the test statistic $\hat{\boldsymbol{\mu}}$ if $\bar{H} : \mu_1 = 10$ is true. The interval marked by S_1 denotes the set of observed $\hat{\mu}$ values for which H is accepted (non-critical region). The interval marked by S_0 denotes the set of observed $\hat{\mu}$ values for which H is rejected (critical region). The area of the black pattern region on the right equals Prob $\{\mathbf{D}_N \in S_0|H\}$, i.e. the probability of rejecting H when H is true (Type I error). The area of the grey shaded region on the left equals the probability of accepting H when H is false (Type II error).

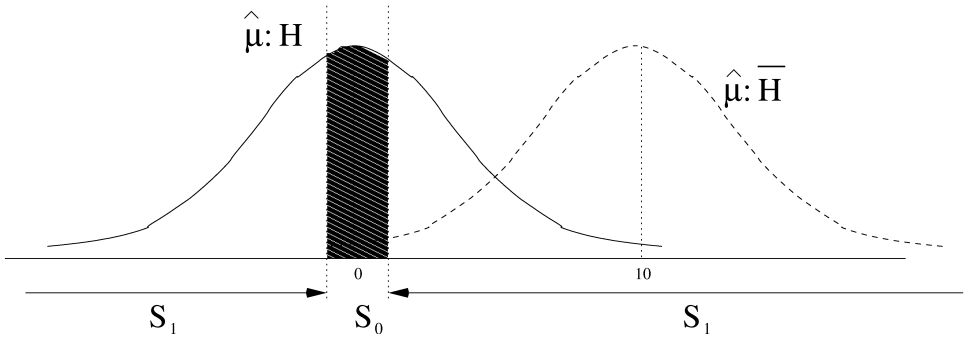


Figure 5.8: On the left: distribution of the test statistic $\hat{\boldsymbol{\mu}}$ if $H : \mu_0 = 0$ is true. On the right: distribution of the test statistic $\hat{\boldsymbol{\mu}}$ if $\bar{H} : \mu_1 = 10$ is true. The two intervals marked by S_1 denote the set of observed $\hat{\mu}$ values for which H is accepted (non-critical region). The interval marked by S_0 denotes the set of observed $\hat{\mu}$ values for which H is rejected (critical region). The area of the pattern region equals Prob $\{\mathbf{D}_N \in S_0|H\}$, i.e. the probability of rejecting H when H is true (Type I error). Which area corresponds to the probability of the Type II error?

5.11.8 UMP level-α test

Given a significance level α we denote by *uniformly most powerful (UMP)* test, the test

1. which satisfies
$$\text{Prob}\{\text{reject H}|H\} = \text{Prob}\{\mathbf{D}_N \in S_0|H\} = \alpha$$
2. for which
$$\text{Prob}\left\{\text{reject H}|\bar{H}\right\} = \text{Prob}\left\{\mathbf{D}_N \in S_0|\bar{H}\right\}$$
is maximed simultaneously for all $\theta \in \Theta_{\bar{H}}$.

How is it possible to find UMP tests? In a simple case, an answer is given by the Neyman-Pearson lemma.

5.11.9 Likelihood ratio test

Consider the simplest case $\Theta = \{\theta_0, \theta_1\}$, where $H : \theta = \theta_0$ and $\bar{H} : \theta = \theta_1$ and θ_0, θ_1 are two different values of the parameter of a r.v. $\mathbf{z}$. Let us denote the two likelihoods by $L_0(\theta)$ and $L_1(\theta)$, respectively.

The idea of Neyman and Pearson was to base the acceptance/rejection of H on the relative values $L(\theta_0)$ and $L(\theta_1)$. In other terms we reject H if the *likelihood ratio*
$$\frac{L(\theta_1)}{L(\theta_0)}$$
is sufficiently big.

We reject H only if the sample data D_N are sufficiently more probable when $\theta = \theta_1$ than when $\theta = \theta_0$.

Lemma 2 (Neyman-Pearson lemma)**.** Let $H : \theta = \theta_0$ and $\bar{H} : \theta = \theta_1$. If a partition $\{S_0, S_1\}$ of the sample space $\mathcal{D}$ is defined by
$$S_0 = \{D_N : L(\theta_1) > kL(\theta_0)\} \qquad S_1 = \{D_N : L(\theta_1) < kL(\theta_0)\}$$
with $\int_{S_0} p(D_N, \theta_0) dD_N = \alpha$, then $\{S_0, S_1\}$ is the most powerful level-α test of H against $\bar{H}$.

This lemma demonstrates that among all tests of level $\leq \alpha$, the likelihood ratio test is the optimal procedure, i.e. it has the smallest probability of type II error.

Although, for a generic distribution, the definition of an optimal test is very difficult, all the tests that will be described in the following are optimal in the UMP sense.

5.12 Parametric tests

Suppose we want to test an assertion about a random variable with a known parametric distribution $F(\cdot, \theta)$. Besides the distinction between simple and composite tests presented in Section 5.11.1, there are two more ways of classifying hypothesis tests:

One-sample vs. two-sample: one-sample tests concern an hypothesis about the properties of a single r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$ while two-sample test concern the relationship between two r.v. $\mathbf{z}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathbf{z}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Single-sided (one-tailed) vs. Two-sided (two-tailed): in single-sided tests the region of rejection concerns only one tail of the distribution of the null hypothesis. This means that $\bar{H}$ indicates the predicted direction of the difference. In two-sided tests the region of rejection concerns both tails of the null distribution. This means that $\bar{H}$ does not indicate the predicted direction of the difference.

The most common parametric tests rely on hypothesis of normality. A non-exhaustive list of conventional parametric test is available in the following table:

Name	single/two sample	known	H	$\bar{H}$
z-test	single	σ^2	$\mu = \mu_0$	$\mu \neq \mu_0$
z-test	two	$\sigma_1^2 = \sigma_2^2$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
t-test	single		$\mu = \mu_0$	$\mu \neq \mu_0$
t-test	two		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
χ^2-test	single	μ	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$
χ^2-test	single		$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$
F-test	two		$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$

The columns H and $\bar{H}$ contain the parameter taken into consideration by the test.

All the parametric test procedures can be decomposed into five main steps:

1. Define the null hypothesis and the alternative one.
2. Fix the probability α of having a Type I error.
3. Choose a test statistic $t(D_N)$.
4. Define the critical value t_α that satisfies the Type I error constraint.
5. Collect the dataset D_N, compute $t(D_N)$ and decide if the hypothesis is either accepted or rejected.

Note that the first 4 steps are independent of the data and should be carried out *before* the collection of the dataset. A more detailed description of some of these tests is contained in the following sections and Appendix C.3.

5.12.1 z-test (single and one-sided)

Consider a random sample $D_N \leftarrow \mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$ with μ unknown and σ^2 known. Let us see in detail how the five steps of the testing procedure are instantiated in this case.

STEP 1:

Consider the null hypothesis and the alternative (composite and one-sided)

$$H : \mu = \mu_0; \qquad \bar{H} : \mu > \mu_0$$

STEP 2: fix the value α of the type I error.

STEP 3: If H is true then the distribution of $\hat{\boldsymbol{\mu}}$ is $\mathcal{N}(\mu_0, \sigma^2/N)$. This means that the test statistic $t(D_N)$ is

$$t_N = t(D_N) = \frac{(\hat{\boldsymbol{\mu}} - \mu_0)\sqrt{N}}{\sigma} \sim \mathcal{N}(0, 1)$$

STEP 4: determine the critical value t_α.

We reject the hypothesis H if $t_N > t_\alpha = z_\alpha$ where z_α is such that Prob $\{\mathcal{N}(0,1) > z_\alpha\} = \alpha$.

Example: for $\alpha = 0.05$ we would take $z_\alpha = 1.645$ since 5% of the standard normal distribution lies to the right of 1.645. Note that the value z_α for a given α can be obtained by the R command `qnorm(alpha,lower.tail=FALSE)`.

STEP 5: Once the dataset D_N is measured, the value of the test statistic is

$$t_N = \frac{(\hat{\mu} - \mu_0)\sqrt{N}}{\sigma}$$

and the hypothesis is either accepted ($t_N \leq z_\alpha$) or rejected.

Example z-test

Consider a r.v. $\mathbf{z} \sim \mathcal{N}(\mu, 1)$. We want to test $H : \mu = 5$ against $\bar{H} : \mu > 5$ with significance level 0.05. Suppose that the dataset is $D_N = \{5.1, 5.5, 4.9, 5.3\}$. Then $\hat{\mu} = 5.2$ and $z_N = (5.2 - 5) * 2/1 = 0.4$. Since this is less than 1.645, we do not reject the null hypothesis.

•

5.12.2 t-test: single sample and two-sided

Consider a random sample from $\mathcal{N}(\mu, \sigma^2)$ with σ^2 *unknown*. Let

$$H : \mu = \mu_0; \qquad \bar{H} : \mu \neq \mu_0$$

Let

$$t(D_N) = t_N = \frac{\sqrt{N}(\hat{\mu} - \mu_0)}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(z_i - \hat{\mu})^2}} = \frac{(\hat{\mu} - \mu_0)}{\sqrt{\frac{\hat{\sigma}^2}{N}}}$$

a statistic computed using the data set D_N.

If the hypothesis H holds, from Sections C.2.3 and 5.7 it follows that $t(\mathbf{D}_N) \sim \mathcal{T}_{N-1}$ is a r.v. with a Student distribution with $N-1$ degrees of freedom. The size α t-test consists in rejecting H if

$$|t_N| > k = t_{\alpha/2,N-1}$$

where $t_{\alpha/2,N-1}$ is the upper α point of a $\mathcal{T}$-distribution on $N-1$ degrees of freedom, i.e.

$$\text{Prob}\left\{\boldsymbol{t}_{N-1} > t_{\alpha/2,N-1}\right\} = \alpha/2, \quad \text{Prob}\left\{|\boldsymbol{t}_{N-1}| > t_{\alpha/2,N-1}\right\} = \alpha.$$

where $\boldsymbol{t}_{N-1} \sim \mathcal{T}_{N-1}$. In other terms H is rejected when t_N is too large.

Note that the value $t_{\alpha/2,N-1}$ for a given N and α can be obtained by the R command `qt(alpha/2,N-1,lower.tail=TRUE)`.

Example [64]

Suppose we want an answer to the following question: *Does jogging lead to a reduction in pulse rate?*. Let us engage eight non jogging volunteers in a one-month jogging programme and let us take their pulses before and after the programme

pulse rate before	74	86	98	102	78	84	79	70
pulse rate after	70	85	90	110	71	80	69	74
decrease	4	1	8	-8	7	4	10	-4

Let us assume that the decreases are randomly sampled from $\mathcal{N}(\mu, \sigma^2)$ where σ^2 is unknown. We want to test $H : \mu = \mu_0 = 0$ against $\bar{H} : \mu \neq 0$ with a significance $\alpha = 0.05$. We have $N = 8$, $\hat{\mu} = 2.75$, $T = 1.263$, $t_{\alpha/2,N-1} = 2.365$ Since $|T| \leq t_{\alpha/2,N-1}$, the data is not sufficient to reject the hypothesis H. In other terms the experiment does not provide enough evidence that jogging leads to reduction in pulse rate.

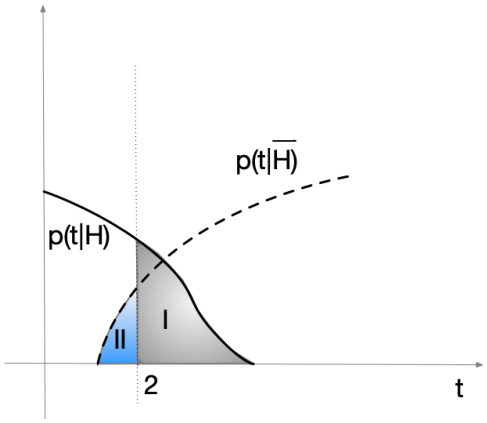


Figure 5.9: On the left: distribution of the test statistic (number of identical lines) if H is true, i.e. the student is honest. Typically honest students have very few lines in common with others though it could happen by chance that such number is more than 2. On the right: distribution of the test statistic (number of identical lines) if $\bar{H}$ is true, i.e. the student is dishonest. Typically dishonest students have several lines in common with others though some of them are cunning enough to conceal it.

•

So far we assumed that the distribution of the test statistic is known under the null hypothesis. In this case it is possible to fix a priori the Type I error. But what about if we do not know anything about the distribution? Is it possible to assess a posteriori the quality (in terms of errors of Type I or II) of a certain test (e.g. using a certain threshold) ?

5.13 A posteriori assessment of a test

Let us consider the professor example at page 122 and the hypothesis test strategy which leads to the refusal of a student when $t_N > t_\alpha = 2$. The distributions of the t_N statistic for an honest student and a dishonest one have no known parametric form and are plotted in Figure 5.9.

However, the professor has no access to such information. He has then no a priori way to measure or control the Type I error rate (equivalent to the grey area in Figure 5.9).

Nevertheless, it is possible to compute the Type I and Type II error rate posteriori if we have access to the decisions of the professor and the real nature of student (honest or dishonest). Suppose that N students took part in the exam and that N_0 did not copy while N_1 copied. According to the professor $\hat{N}_N$ were considered honest and passed the exam, while $\hat{N}_P$ were considered dishonest and rejected. Because of the overlapping of the distributions in Figure 5.9, it happens that $F_P > 0$ honest students (the ones in the grey area) failed and $F_N > 0$ dishonest students (the ones in the blue area) passed. Note that the honest students who failed indeed did not copy but they had by chance more than one line in common with a classmate. At the same time there are dishonest students who succeeded by copying but who were clever enough to avoid more than 2 identical lines.

The resulting situation can be summarised in this table where we associated the null hypothesis H to the minus sign (non guilty or honest) and the hypothesis $\bar{H}$ to the plus sign.

	Passed	Failed	
H: Honest student (-)	T_N	F_P	$N_N = T_N + F_P$
H: Guilty student (+)	F_N	T_P	$N_P = F_N + T_P$
	$\hat{N}_N = T_N + F_N$	$\hat{N}_P = F_P + T_P$	N

In this table F_P is the number of False Positives, i.e. the number of times that the professor predicted the student as guilty (+) but in reality she was innocent (-). The ratio F_P/N represents an estimate of the type I error (probability of rejecting the null hypothesis when it is true). The term F_N represents the number of False Negatives, i.e. the number of times that the professor predicted the student as honest (-) but in reality he had copied (+). The ratio F_N/N is an estimation of the type II error (probability of accepting the null hypothesis when it is false).

5.14 Conclusion

The reader wishing to know more about machine learning could be disappointed. She has been reading more than one hundred pages and he has still the sensation that she did not learn much about machine learning. All he read seems very far from intelligent agents, neural networks and fancy applications... Nevertheless, she already came across the most important notions of machine learning: conditional probability, estimation and bias/variance trade-off. Is it all about that? From an abstract perspective, yes. All the fancy algorithms that will be presented afterwards (or that the reader is used to hear about) are nothing more (often without designer knowledge) *estimators of conditional probability, and as such, submitted to a bias/variance tradeoff*. Such algorithms are accurate and useful only if they manage well such trade-off.

But we can go a step further and see the bias/variance tradeoff not only as a statistical concept but as a metaphor of human attitude towards models and data, beliefs and experience, ideology and observations, preconceptions and events [6]. Humans define models (not only in science but also in politics, economics, religion) to represent the regularity of nature. Now, reality often escapes or diverges from such regularity. In front of the gap between the Eden of regularity and the natural Hell of observations, humans waver between two extremes: i) negate or discredit reality and reduce all divergences to some sort of noise (measurement error) or ii) adapt, change their belief, to incorporate discording data and measures in their model (or preconceptions).

The first attitude is exposed to bias (or dogmatism or worse conspiracy thinking): the second to variance (or instability). A biased human learner behaves as an estimator which is insensitive to data: her strength derives from the intrinsic robustness and coherence, and his weakness is due to the (in)sane attitude of disregarding data and flagrant evidence. On the other side, a highly variant human learner adapts rapidly and swiftly to data and observations, but he can be easily criticised for his excessive instability, for going where the wind blows.

When real events do not confirm your expectations (or what your parents, teachers or media told you), what is the best attitude to take? Is there an optimal attitude? Which side are you on?

5.15 Exercises

1. Derive analytically the bias of the sample average estimator in a non i.i.d. setting.

2. Derive analytically the variance of the sample average estimator in an i.i.d. setting.

[6] `https://tinyurl.com/y2514xyp`

3. Consider a regression problem where

$$\mathbf{y} = \sin(\mathbf{x}) + \mathbf{w}$$

and $\mathbf{x}$ is uniformly distributed on the interval $[0, 2\pi]$ and $\mathbf{w} = \mathcal{N}(1, 1)$ is a Normal variable with both mean and variance equal to 1. Let us consider a predictor $h(x)$ that is distributed like $\mathbf{w}$. Compute the bias and variance of the predictor in the following coordinates: $x = 0$, $x = \pi$, $x = \pi/2$.

Solution:

- $x = 0$ Bias=0, Var=1
- $x = \pi$ Bias=0, Var= 1
- $x = \pi/2$ Bias=1, Var=1

4. Let us consider a dataset $D_N = \{z_1, \ldots, z_{20}\}$ of 20 observations generated according to an uniform distribution over the interval $[-1, 1]$. Suppose I want to estimate the expected value of the distribution. Compute the bias and variance of the following estimators:

- $\hat{\boldsymbol{\theta}}_1 = \frac{\sum_{i=1}^{10} \mathbf{z}_i}{10}$
- $\hat{\boldsymbol{\theta}}_2 = \hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{20} \mathbf{z}_i}{20}$
- $\hat{\boldsymbol{\theta}}_3 = -1$
- $\hat{\boldsymbol{\theta}}_4 = 1$
- $\hat{\boldsymbol{\theta}}_5 = \mathbf{z}_2$

Suppose I want to estimate the variance of the distribution. Compute the bias of the following estimators:

- $\hat{\boldsymbol{\sigma}}_1^2 = \frac{\sum(\mathbf{z}_i - \hat{\mu})^2}{19}$
- $\hat{\boldsymbol{\sigma}}_2^2 = \frac{\sum(\mathbf{z}_i - \hat{\mu})^2}{20}$
- $\hat{\boldsymbol{\sigma}}_3^2 = 1/3$

Solution: Note that $\theta = 0$ and $\sigma_{\mathbf{z}}^2 = 1/3$

$\hat{\boldsymbol{\theta}}_1 : B_1 = 0, V_1 = 0.03$. Justification: $E[\hat{\boldsymbol{\theta}}_1] = \theta$ and Var $\left[\hat{\boldsymbol{\theta}}_1\right] = \sigma^2/10$

$\hat{\boldsymbol{\theta}}_2 : B_2 = 0, V_2 = 0.015$. Justification: $E[\hat{\boldsymbol{\theta}}_2] = \theta$ and Var $\left[\hat{\boldsymbol{\theta}}_2\right] = \sigma^2/20$

$\hat{\boldsymbol{\theta}}_3 : B_3 = -1, V_3 = 0$. Justification: $E[\hat{\boldsymbol{\theta}}_3] = -1$ and Var $\left[\hat{\boldsymbol{\theta}}_3\right] = 0$ since constant

$\hat{\boldsymbol{\theta}}_4 : B_4 = 1, V_4 = 0$. Justification: $E[\hat{\boldsymbol{\theta}}_4] = 1$ and Var $\left[\hat{\boldsymbol{\theta}}_4\right] = 0$ since constant

$\hat{\boldsymbol{\theta}}_5 : B_5 = 0, V_5 = 0.33$. Justification: $E[\hat{\boldsymbol{\theta}}_5] = \theta$ and Var $\left[\hat{\boldsymbol{\theta}}_5\right] = \sigma^2$

$\hat{\boldsymbol{\sigma}}_1^2 : B = 0$. Justification: sample variance is unbiased then $E[\hat{\boldsymbol{\sigma}}_1^2] = \sigma_{\mathbf{z}}^2$

$\hat{\boldsymbol{\sigma}}_2^2 : B - 1/60 = -0.0166$. Justification: Note first that $\hat{\boldsymbol{\sigma}}^2 = \frac{19}{20}\frac{\sum(\mathbf{z}_i - \hat{\mu})^2}{19}$. Then

$$E[\hat{\boldsymbol{\sigma}}_2^2] = \frac{19}{20} E\left[\frac{\sum(\mathbf{z}_i - \hat{\mu})^2}{19}\right] = \frac{19}{20}\sigma_{\mathbf{z}}^2$$

then

$$E[\hat{\boldsymbol{\sigma}}_2^2] - \sigma_{\mathbf{z}}^2 = \frac{19}{20}\sigma_{\mathbf{z}}^2 - \sigma_{\mathbf{z}}^2 = -\sigma_{\mathbf{z}}^2/20$$

$\hat{\boldsymbol{\sigma}}_3^2 : B = 0$. Justification $E[1/3] = 1/3 = \sigma_{\mathbf{z}}^2$

5. Let us consider the following observations of the random variable $\mathbf{z}$

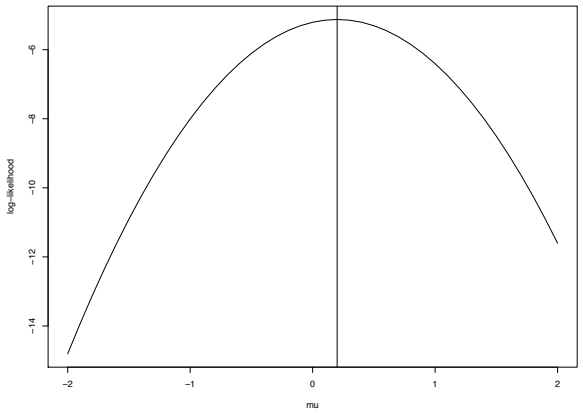
$$D_N = \{0.1, -1, 0.3, 1.4\}$$

Write the analytical form of the likelihood function of the mean μ for a Gaussian distribution with a variance $\sigma^2 = 1$. The student should:

1. Trace the log-likelihood function on the graph paper
2. Determine graphically the maximum likelihood estimator.
3. Discuss the result.

Solution: Since $N = 4$ and $\sigma = 1$

$$L(\mu) = \prod_{i=1}^{N} p(\mathbf{z}_i, \mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp \frac{-(z_i - \mu)^2}{2}$$



Note that $\hat{\mu}_{ml}$ coincides with the sample average $\hat{\mu} = 0.2$ of D_N.

6. Suppose you want to estimate the expectation μ of the uniform r.v. $\mathbf{z} \sim \mathcal{U}[-2, 3]$ by using a dataset of size $N = 10$. By using R and its random generator, first plot the sampling distribution then estimate the bias and the variance of the following estimators:

1. $\hat{\theta} = \sum_{i=1}^{N} \frac{z_i}{N}$
2. $\hat{\theta} = \min_{i=1}^{N} z_i$
3. $\hat{\theta} = \max_{i=1}^{N} z_i$
4. $\hat{\theta} = z_1$
5. $\hat{\theta} = z_N$
6. $\hat{\theta} = \sum_{i=1} N \frac{|z_i|}{N}$
7. $\hat{\theta} = \text{median}_i z_i$
8. $\hat{\theta} = \max_{i=1}^{N} w_i$ where $\mathbf{w} \sim \mathcal{N}(0, 1)$.
9. $\hat{\theta} = 1$

Before each random generation set the seed to zero.

7. The student should first create a dataset of $N = 1000$ observations according to the dependency

$$\mathbf{y} = g(\beta_0 + \beta_1 \mathbf{x}) + \mathbf{w}$$

where $\mathbf{x} \sim \mathcal{U}[-1, 1]$, $\beta_0 = 1$, $\beta_1 = -1$, $\mathbf{w} \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.1)$, $g(x) = \frac{e^x}{1+e^x}$.

Then by using the same dataset he should:

- estimate by maximum likelihood the parameters β_0 and β_1,
- plot the contour of the likelihood function, showing in the same graph the values of the parameters and their estimations.

Hint: use a grid search to perform the maximisation.

8. The student should first create a dataset of $N = 1000$ observations according to the dependence

$$\text{Prob}\{\mathbf{y} = 1|x\} = g(\beta_0 + \beta_1 x)$$

where $\mathbf{x} \sim \mathcal{U}[0,1]$, $\beta_0 = 1$, $\beta_1 = -1$, $g(x) = \frac{e^x}{1+e^x}$ and $y \in \{0,1\}$.

Then by using the same dataset she should:

- estimate by maximum likelihood the parameters β_0 and β_1,
- plot the contour of the likelihood function, showing in the same graph the values of the parameters and their estimations.

Hint: use a grid search to perform the maximisation.

9. Let $\mathbf{z} \sim \mathcal{N}(1,1)$, D_N a training set of N i.i.d. observations z_i and $\hat{\mu}_N$ the related sample average estimator.

   1. Compute analytically
   $$E_{\mathbf{z},\mathbf{D}_N}[(\mathbf{z} - \hat{\boldsymbol{\mu}}_N)^2]$$
   Hint: consider that $\mathbf{z} = \theta + \mathbf{w}$ where $\theta = E[\mathbf{z}]$ and $\mathbf{w} \sim \mathcal{N}(0,1)$.
   2. Compute analytically
   $$E_{\mathbf{z},\mathbf{D}_N}[(\mathbf{z} - \hat{\boldsymbol{\mu}}_N)]$$
   3. Validate by Monte Carlo simulation the two theoretical results above.

   Solution: Since $E[\hat{\boldsymbol{\mu}}] = \mu$, $\text{Var}[\hat{\boldsymbol{\mu}}] = \sigma^2_{\mathbf{w}}/N$ and $\mathbf{w}$ is independent of $\mathbf{D}_N$:

$$\begin{aligned} E_{\mathbf{z},\mathbf{D}_N}[(\mathbf{z} - \hat{\boldsymbol{\mu}}_N)^2] &= E_{\mathbf{z},\mathbf{D}_N}[(\theta + \mathbf{w} - \hat{\boldsymbol{\mu}}_N)^2] = \\ &= E_{\mathbf{z},\mathbf{D}_N}[\mathbf{w}^2 + 2\mathbf{w}(\theta - \hat{\boldsymbol{\mu}}_N) + (\theta - \hat{\boldsymbol{\mu}}_N)^2] = \\ &= E_{\mathbf{z}}[\mathbf{w}^2] + E_{\mathbf{D}_N}[(\theta - \hat{\boldsymbol{\mu}}_N)^2] = \sigma^2_{\mathbf{w}} + \sigma^2_{\mathbf{w}}/N = 1 + 1/N \end{aligned}$$

   R code to perform Monte Carlo validation :

```
rm(list=ls())
N=5
S=10000
sdw=1  ## noise variance
E=NULL
for (s in 1:S){
  DN=rnorm(N,1,sdw)
  muhat=mean(DN)
  z=rnorm(1,1,sdw)
  e=z-muhat
  E=c(E,e^2)
}

cat("th=",sdw^2+sdw^2/N, "MC estimation=", mean(E),"\n")
```

10. Let us supposed that the only measurement of a Gaussian random variable $\mathbf{z} \sim \mathcal{N}(\mu, 1)$ is the interval $[-3.5, 1.5]$. Estimate μ by maximum-likelihood and show the likelihood-function $L(\mu)$. Hint: use the R function `pnorm`.

11. Let us suppose that 12 of the 31 days of August in Brussels are rainy. Estimate the probability of a rainy day by maximum likelihood by using the Binomial distribution (Section C.1.2).

Chapter 6

Nonparametric approaches to estimation and testing

6.1 Nonparametric methods

In the previous chapter, we considered estimation problems where the probability distribution is known, parameters' value (e.g. mean and/or variance) aside. Such estimation methods are called *parametric*. The meaningfulness of a parametric test depends entirely on the validity of the assumptions made about the analytical form of the distribution. However, in real configurations, it is not uncommon for the experimenter to question parametric assumptions.

Consider a random sample $D_N \leftarrow \mathbf{z}$ collected through some experimental observation and for which no hint about the underlying probability distribution $F_{\mathbf{z}}(\cdot)$ is available. Suppose we want to estimate a parameter of interest θ of the distribution of $\mathbf{z}$ by using the plug-in estimate $\hat{\theta} = t(\hat{F})$ (Section 5.3). What can we say about the accuracy of the estimator $\hat{\boldsymbol{\theta}}$? As shown in Section 5.5.3, for some specific parameters (e.g. mean and variance) the accuracy can be estimated independently of the parametric distribution. In most cases, however, the assessment of the estimator is not possible unless we know the underlying distribution. What to do, hence, if the distribution is not available? A solution is provided by the so-called *nonparametric* or *distribution-free* methods that work independently on any specific assumption about the probability distribution.

The adoption of these methods enjoyed considerable success in the last decades thanks to the evolution and parallelisation of computational processing power. In fact, most techniques for nonparametric estimation and testing are based on *resampling procedures*, which require a large number of repeated (and almost similar) computations on the data.

This chapter will deal with two resampling strategies for estimation and two resampling strategies for hypothesis testing, respectively.

Jacknife: this approach to nonparametric estimation relies on repeated computations of the statistic of interest for all the combinations of the data where one or more of the original examples are removed. It will be presented in Section 6.3.

Bootstrap: this approach to nonparametric estimation aims to estimate the sampling distribution of an estimator by sampling (with replacement) from the original data. It will be introduced in Section 6.4.

Randomisation: This is a resampling without replacement testing procedure. It

consists in taking the original data and either scrambling the order or the association of the original data. It will be discussed in Section 6.5.

Permutation: This is a resampling two-sample hypothesis-testing procedure based on repeated permutations of the dataset. It will be presented in Section 6.6.

6.2 Estimation of arbitrary statistics

Consider a set D_N of N data points sampled from a scalar r.v. $\mathbf{z}$. Let $E[\mathbf{z}] = \mu$ the parameter to be estimated. In Section 5.3.1 we derived the bias and the variance of the estimator $\hat{\boldsymbol{\mu}}$:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} z_i, \qquad \text{Bias}[\hat{\boldsymbol{\mu}}] = 0, \qquad \text{Var}\left[\hat{\boldsymbol{\mu}}\right] = \frac{\sigma^2}{N}$$

Consider now another quantity of interest, for example, the median or a mode of the distribution. While it is easy to design a plug-in estimate of these quantities, their accuracy is difficult to be computed. In other terms, *given an arbitrary estimator* $\hat{\boldsymbol{\theta}}$*, the analytical form of the variance* $Var\left[\hat{\boldsymbol{\theta}}\right]$ *and the bias* $Bias[\hat{\boldsymbol{\theta}}]$ *is typically not available.*

Example

According to the plug-in principle (Section 5.3) we can design other estimators besides sampled mean and variance, like:

- Estimation of skewness (3.3.36) of $\mathbf{z}$: see Equation (D.0.2).
- Estimation of correlation (3.6.67) between $\mathbf{x}$ and $\mathbf{y}$: : see Equation (D.0.3).

What about the accuracy (e.g. bias, variance) of such estimators?

•

Example

Let us consider an example of estimation taken from an experimental medical study [64]. The goal of the study is to show bioequivalence between an old and a new version of a patch designed to infuse a certain hormone in the blood. Eight subjects take part in the study. Each subject has his hormone levels measured after wearing three different patches: a placebo, an "old" patch and a "new" patch. It is established by the Food and Drug Administration (FDA) that the new patch will be approved for sale only if the new patch is bioequivalent to the old one according to the following criterion:

$$\theta = \frac{|E(\text{new}) - E(\text{old})|}{E(\text{old}) - E(\text{placebo})} \leq 0.2 \tag{6.2.1}$$

Let us consider the following plug-in estimator (Section 5.3) of (6.2.1)

$$\hat{\boldsymbol{\theta}} = \frac{|\hat{\boldsymbol{\mu}}_{\text{new}} - \hat{\boldsymbol{\mu}}_{\text{old}}|}{\hat{\boldsymbol{\mu}}_{\text{old}} - \hat{\boldsymbol{\mu}}_{\text{placebo}}}$$

Suppose we have collected the following data (details in [64])

subj	plac	old	new	z=old-plac	y=new-old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
...	...	...	...	...	...
8	18806	29044	26325	10238	-2719
mean:				6342	-452.3

The estimate is

$$\hat{\theta} = t(\hat{F}) = \frac{|\hat{\mu}_{\text{new}} - \hat{\mu}_{\text{old}}|}{\hat{\mu}_{\text{old}} - \hat{\mu}_{\text{placebo}}} = \frac{|\hat{\mu}_y|}{\hat{\mu}_z} = \frac{452.3}{6342} = 0.07$$

Can we say on the basis of this value that the new patch satisfies the FDA criterion in (6.2.1)? What about the accuracy, bias or variance of the estimator? The techniques introduced in the following sections may provide an answer to these questions.

•

6.3 Jackknife

The *jackknife* (or *leave-one-out*) resampling technique aims at providing a computational procedure to estimate the variance and the bias of a generic estimator $\hat{\boldsymbol{\theta}}$. The technique was first proposed by Quenouille in 1949 and is based on removing examples from the available dataset and recalculating the estimator. It is a general-purpose tool that is easy to implement and able to solve a number of estimation problems.

6.3.1 Jackknife estimation

In order to show the theoretical foundation of the jackknife, we first apply this technique to the estimator $\hat{\boldsymbol{\mu}}$ of the mean. Let $D_N = \{z_1, \ldots, z_N\}$ be the available dataset. Let us remove the ith example from D_N and let us calculate the *leave-one-out (l-o-o) mean estimate* from the $N-1$ remaining examples

$$\hat{\mu}_{(i)} = \frac{1}{N-1} \sum_{j \neq i}^{N} z_j = \frac{N\hat{\mu} - z_i}{N-1}$$

Observe from above that the following relation holds

$$z_i = N\hat{\mu} - (N-1)\hat{\mu}_{(i)} \tag{6.3.2}$$

that is, we can calculate the ith example z_i, $i = 1, \ldots, N$ if we know both $\hat{\mu}$ and $\hat{\mu}_{(i)}$. Suppose now we wish to estimate some parameter θ by using as estimator some complex statistic of the N data points

$$\hat{\theta} = g(D_N) = g(z_1, z_2, \ldots, z_N)$$

The jackknife procedure consists in first computing

$$\hat{\theta}_{(i)} = g(z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N), \quad i = 1, \ldots, N$$

which is called the ith *jackknife replication* of $\hat{\theta}$. Then by analogy with the relation (6.3.2) holding for the mean estimator, we define the i-th *pseudo value* by

$$\eta_{(i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(i)}. \tag{6.3.3}$$

These pseudo values assume the same role as the z_i in calculating the sample average (5.3.4). Hence the *jackknife estimate* of θ is given by

$$\hat{\theta}_{\text{jk}} = \frac{1}{N}\sum_{i=1}^{N}\eta_{(i)} = \frac{1}{N}\sum_{i=1}^{N}\left(N\hat{\theta} - (N-1)\hat{\theta}_{(i)}\right) = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)} \tag{6.3.4}$$

where

$$\hat{\theta}_{(\cdot)} = \frac{\sum_{i=1}^{N}\hat{\theta}_{(i)}}{N}.$$

The rationale of the jackknife technique is to use the quantity (6.3.4) in order to estimate the bias of the estimator. Since, according to (5.5.8), $\theta = E[\hat{\boldsymbol{\theta}}] - \text{Bias}[\hat{\boldsymbol{\theta}}]$, the jackknife approach consists in replacing θ by $\hat{\theta}_{\text{jk}}$ and $E[\hat{\boldsymbol{\theta}}]$ by $\hat{\theta}$, thus obtaining

$$\hat{\theta}_{\text{jk}} = \hat{\theta} - \text{Bias}_{\text{jk}}[\hat{\boldsymbol{\theta}}].$$

It follows that the *jackknife estimate of the bias of* $\hat{\boldsymbol{\theta}}$ is

$$\text{Bias}_{\text{jk}}[\hat{\boldsymbol{\theta}}] = \hat{\theta} - \hat{\theta}_{\text{jk}} = \hat{\theta} - N\hat{\theta} + (N-1)\hat{\theta}_{(\cdot)} = (N-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}).$$

Note that in the particular case of a mean estimator (i.e. $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$), we see that we obtain, as expected, $\text{Bias}_{\text{jk}}[\hat{\boldsymbol{\mu}}] = 0$.

A jackknife estimate of the variance of $\hat{\boldsymbol{\theta}}$ can be obtained from the sample variance of the pseudo-values. We define the *jackknife estimate of the variance of* $\hat{\boldsymbol{\theta}}$ as

$$\text{Var}_{\text{jk}}[\hat{\boldsymbol{\theta}}] = \text{Var}\left[\hat{\boldsymbol{\theta}}_{\text{jk}}\right] \tag{6.3.5}$$

Under the hypothesis of i.i.d. $\eta_{(i)}$

$$\text{Var}\left[\hat{\boldsymbol{\theta}}_{\text{jk}}\right] = \text{Var}\left[\frac{\sum_{i=1}^{N}\boldsymbol{\eta}_{(i)}}{N}\right] = \frac{\text{Var}\left[\boldsymbol{\eta}_{(i)}\right]}{N}$$

From (6.3.3) we have

$$\frac{\sum_{i=1}^{N}\eta_{(i)}}{N} = N\hat{\theta} - \frac{(N-1)}{N}\sum_{i=1}^{N}\hat{\theta}_{(i)}$$

Since

$$\eta_{(i)} = N\hat{\theta} - (N-1)\hat{\theta}_{(i)} \Leftrightarrow \eta_{(i)} - \frac{\sum_{i=1}^{N}\eta_{(i)}}{N} = -(N-1)\left(\hat{\theta}_{(i)} - \frac{\sum_{i=1}^{N}\hat{\theta}_{(i)}}{N}\right)$$

from (6.3.5) and (6.3.4) we obtain

$$\text{Var}_{\text{jk}}[\hat{\boldsymbol{\theta}}] = \frac{\sum_{i=1}^{N}\left(\eta_{(i)} - \hat{\theta}_{\text{jk}}\right)^2}{N(N-1)} = \left(\frac{N-1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right)$$

Note that in the case of the estimator of the mean (i.e. $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\mu}}$), since $\eta_{(i)} = z_i$ and $\hat{\theta}_{\text{jk}} = \hat{\mu}$, we find again the result (5.5.10)

$$\text{Var}_{\text{jk}}[\hat{\boldsymbol{\theta}}] = \frac{\sum_{i=1}^{N}(z_i - \hat{\mu})^2}{N(N-1)} = \frac{\hat{\sigma}^2}{N} = \text{Var}\left[\hat{\boldsymbol{\mu}}\right] \tag{6.3.6}$$

The major motivation for jacknife estimates is that they reduce bias. Also, it can be shown that under suitable conditions on the type of estimator $\hat{\boldsymbol{\theta}}$, the quantity (6.3.6) converges in probability to $\text{Var}\left[\hat{\boldsymbol{\theta}}\right]$. However, the jacknife can fail if the statistic $\hat{\boldsymbol{\theta}}$ is not smooth (i.e. small changes in data cause small changes in the statistic). An example of non-smooth statistic for which the jacknife works badly is the median.

6.4 Bootstrap

The method of *bootstrap* was proposed by Efron [61] as a computer-intensive technique to estimate the accuracy of a generic estimator $\hat{\boldsymbol{\theta}}$. Bootstrap relies on a data-based simulation method for statistical inference. The term *bootstrap* derives from the phrase *to pull oneself up by one's bootstrap* based on the fictional Adventures of Baron Munchausen. The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps. In general terms, to pull yourself up by your bootstraps means to succeed in something very difficult without any outside help[1].

The idea of statistical bootstrap is very simple, namely that in the absence of any other information, the sample itself offers the best guide of the sampling distribution. The method is completely automatic, requires no theoretical calculation, and is available no matter how mathematically complicated the estimator (5.4.6) is. By resampling with replacement from D_N we can build a set of B datasets $D_{(b)}, b = 1, \ldots, B$. From the empirical distribution of the statistics $g(D_{(b)})$ we can construct confidence intervals and tests for significance.

6.4.1 Bootstrap sampling

Consider a data set D_N. A *bootstrap data set* $D_{(b)}$, $b = 1, \ldots, B$ is created by randomly selecting N points from the original set D_N *with replacement* (Figure 6.1).

Since D_N itself contains N points, there is nearly always duplication of individual points in a bootstrap data set. Each point has an equal probability $1/N$ of being chosen on each draw. Hence, the probability that a point is chosen exactly k times is given by the binomial distribution (Section C.1.2)

$$\text{Prob}\{k\} = \frac{N!}{k!(N-k)!}\left(\frac{1}{N}\right)^k\left(\frac{N-1}{N}\right)^{N-k} \qquad 0 \le k \le N$$

Given a set of N distinct values, there is a total of $\binom{2N-1}{N}$ distinct bootstrap datasets. The number is quite large already for $N > 10$. For example, if $N = 3$ and $D_N = \{a, b, c\}$, we have 10 different bootstrap sets: {a,b,c}, {a,a,b}, {a,a,c}, {b,b,a}, {b,b,c}, {c,c,a}, {c,c,b}, {a,a,a}, {b,b,b}, {c,c,c}.

Under *balanced bootstrap sampling*, the B bootstrap sets are generated in such a way that each original data point is present exactly B times in the entire collection of bootstrap samples.

6.4.2 Bootstrap estimate of the variance

Given the estimator (5.4.6), for each bootstrap dataset $D_{(b)}$, $b = 1, \ldots, B$, we can define a *bootstrap replication*

$$\hat{\theta}_{(b)} = g(D_{(b)}) \qquad b = 1, \ldots, B$$

that is the value of the statistic for the specific bootstrap sample. The bootstrap approach computes the variance of the estimator $\hat{\boldsymbol{\theta}}$ through the variance of the set $\hat{\theta}_{(b)}$, $b = 1, \ldots, B$, given by

$$\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)} \quad \text{where} \quad \hat{\theta}_{(\cdot)} = \frac{\sum_{b=1}^{B}\hat{\theta}_{(b)}}{B} \tag{6.4.7}$$

[1] This term has not the same meaning (though the derivation is similar) as the one used in computer operating systems where bootstrap stands for starting a computer from an hardwired set of core instructions

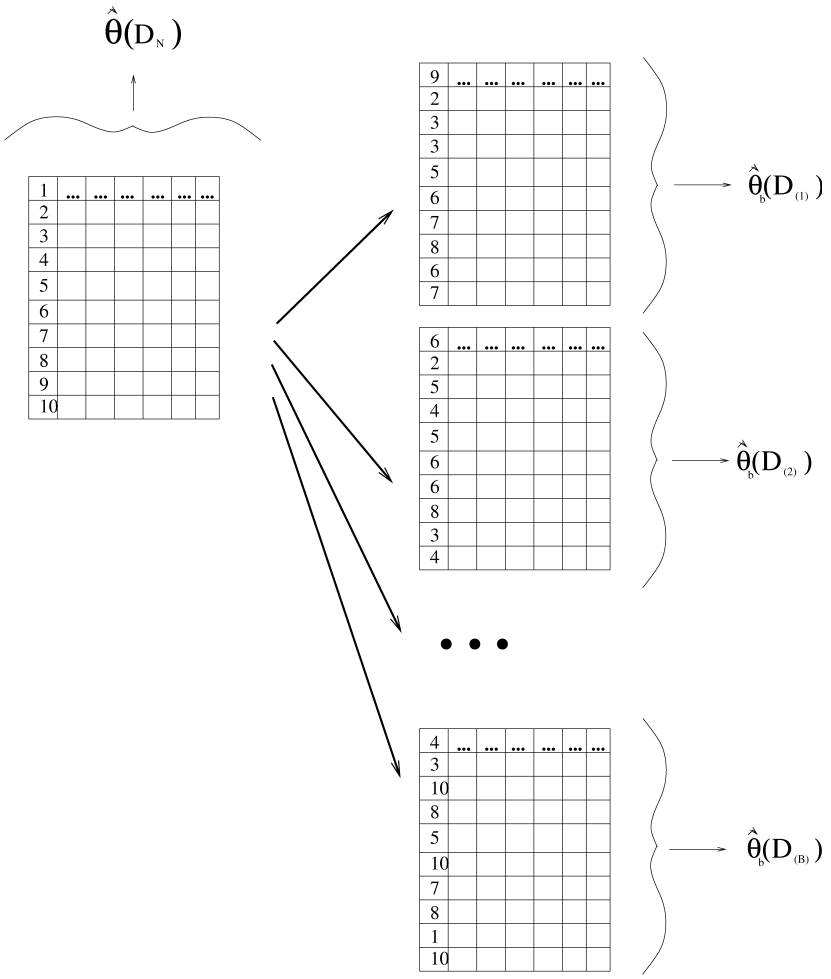


Figure 6.1: Bootstrap replications of a dataset and bootstrap statistic computation

It can be shown that if $\hat{\theta} = \hat{\mu}$, then for $B \to \infty$, the bootstrap estimate $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ converges to the variance $\text{Var}\,[\hat{\boldsymbol{\mu}}]$.

6.4.3 Bootstrap estimate of bias

Let $\hat{\boldsymbol{\theta}}$ be a plug-in estimator (Equation (5.3.3)) based on the sample D_N and

$$\hat{\theta}_{(\cdot)} = \frac{\sum_{b=1}^{B} \hat{\theta}_{(b)}}{B} \tag{6.4.8}$$

Since $\text{Bias}[\hat{\boldsymbol{\theta}}] = E[\hat{\boldsymbol{\theta}}] - \theta$, the *bootstrap estimate of the bias of the plug-in estimator* $\hat{\boldsymbol{\theta}}$ is obtained by replacing $E[\hat{\boldsymbol{\theta}}]$ with $\hat{\theta}_{(\cdot)}$ and θ with $\hat{\theta}$:

$$\text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\theta}_{(\cdot)} - \hat{\theta} \tag{6.4.9}$$

Then, since

$$\theta = E[\hat{\boldsymbol{\theta}}] - \text{Bias}[\hat{\boldsymbol{\theta}}]$$

the *bootstrap bias corrected* estimate is

$$\hat{\theta}_{\text{bs}} = \hat{\theta} - \text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\theta} - (\hat{\theta}_{(\cdot)} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{(\cdot)} \tag{6.4.10}$$

Note that if we want to estimate the bias of a a generic non plug-in estimator $g(D_N)$, the $\hat{\theta}$ term in the right-hand terms of (6.4.9) should anyway refer to the plug-in estimator $t(\hat{F})$ (Equation (5.3.3)).

R script

Run the R file `patch.R` for the estimation of bias and variance in the case of the patch data example.

•

6.4.4 Bootstrap confidence interval

Standard bootstrap confidence limits are based on the assumption that the estimator $\hat{\boldsymbol{\theta}}$ is normally distributed with mean θ and variance σ^2. Taking the bootstrap estimate of variance, an approximate $100(1-\alpha)\%$ confidence interval is given by

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]} = \hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)}} \tag{6.4.11}$$

An improved interval is given by using the bootstrap correction for bias

$$2\hat{\theta} - \hat{\theta}_{(\cdot)} \pm z_{\alpha/2}\sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)}} \tag{6.4.12}$$

Another bootstrap approach for constructing a $100(1-\alpha)\%$ confidence interval is to use the upper and lower $\alpha/2$ values of the bootstrap distribution. This approach is referred to as *bootstrap percentile confidence interval*. If $\hat{\theta}_{L,\alpha/2}$ denotes the value such that only a fraction $\alpha/2$ of all bootstrap estimates are inferior to it, and likewise $\hat{\theta}_{H,\alpha/2}$ is the value exceeded by only $\alpha/2$ of all bootstrap estimates, then the confidence interval is given by

$$[\hat{\theta}_{L,\alpha/2}, \hat{\theta}_{H,\alpha/2}] \tag{6.4.13}$$

where the two extremes are also called the *Efron's percentile confidence limits*.

6.4.5 The bootstrap principle

Given an unknown parameter θ of a distribution $F_{\mathbf{z}}$ and an estimator $\hat{\boldsymbol{\theta}}$, the goal of any estimation procedure is to derive or approximate the distribution of $\hat{\boldsymbol{\theta}} - \theta$. For example, the calculation of the variance of $\hat{\boldsymbol{\theta}}$ requires the knowledge of $F_{\mathbf{z}}$ and the computation of $E_{\mathbf{D}_N}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2]$. Now, in practical contexts, $F_{\mathbf{z}}$ is unknown, and the calculus of $E_{\mathbf{D}_N}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2]$ is not possible in an analytical way. The rationale of the bootstrap approach is (i) to replace $F_{\mathbf{z}}$ by the empirical counterpart (5.2.2) and (ii) to compute $E_{\mathbf{D}_N}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2]$ by a Monte Carlo simulation approach (Section 3.9) where several samples of size N are generated by resampling D_N.

The outcome of a bootstrap technique is a Monte Carlo approximation of the distribution $\hat{\boldsymbol{\theta}}_{(b)} - \hat{\theta}$. In other terms the variability of $\hat{\boldsymbol{\theta}}_{(b)}$ (based on the empirical distribution) around $\hat{\theta}$ is expected to be similar (or mimic) the variability of $\hat{\boldsymbol{\theta}}$ (based on the true distribution) around θ.

The bootstrap principle relies on the two following properties (i) as N gets larger and larger, the empirical distribution $\hat{F}_{\mathbf{z}}(\cdot)$ converges (almost surely) to $F_{\mathbf{z}}(\cdot)$ (Glivenko-Cantelli theorem (C.9.17)) and (ii) as B gets larger, the quantity (6.4.7) converges (in probability) to the variance of the estimator $\hat{\boldsymbol{\theta}}$ based on the empirical distribution (as stated in (C.8.14)). In other terms

$$\mathrm{Var}_{\mathrm{bs}}[\hat{\boldsymbol{\theta}}] \overset{B\to\infty}{\to} E_{\widehat{\mathbf{D}_N}}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2] \overset{N\to\infty}{\to} E_{\mathbf{D}_N}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2] \tag{6.4.14}$$

where $E_{\widehat{\mathbf{D}_N}}[(\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}}])^2]$ stands for the plug-in estimate of the variance of $\hat{\boldsymbol{\theta}}$ based on the empirical distribution.

In practice, for a small finite N, bootstrap estimation inevitably returns some error. This error is a combination of a *statistical error* and a *simulation error*. The statistical error component is due to the difference between the underlying distribution $F_{\mathbf{z}}(\cdot)$ and the empirical distribution $\hat{F}_{\mathbf{z}}(\cdot)$. The magnitude of this error depends on the choice of the estimator $\hat{\boldsymbol{\theta}}(\mathbf{D}_N)$ and decreases by increasing the number N of observations.

The simulation error component is due to the use of empirical (Monte Carlo) properties of $\hat{\boldsymbol{\theta}}(\mathbf{D}_N)$ rather than exact properties. Simulation error decreases by increasing the number B of bootstrap replications.

Unlike the jackknife method, in the bootstrap, the number of replicates B can be adjusted to the computer resources. In practice, two *rules of thumb* are typically used:

1. Even a small number of bootstrap replications, e.g. $B = 25$, is usually informative. $B = 50$ is often enough to give a good estimate of $\mathrm{Var}\left[\hat{\boldsymbol{\theta}}\right]$.

2. Very seldom are more than $B = 200$ replications needed for estimating $\mathrm{Var}\left[\hat{\boldsymbol{\theta}}\right]$. Much bigger values of B are required for bootstrap confidence intervals.

Note that the use of rough statistics $\hat{\boldsymbol{\theta}}$ (e.g. unsmooth or unstable) can make the resampling approach behave wildly. Examples of nonsmooth statistics are sample quantiles and the median.

In general terms, for i.i.d. observations, the following conditions are required for the convergence of the bootstrap estimate

1. the convergence of $\hat{F}$ to F (satisfied by the Glivenko-Cantelli theorem) for $N \to \infty$;

2. an estimator such that the estimate $\hat{\theta}$ is the corresponding functional of the empirical distribution.
$$\theta = t(F) \to \hat{\theta} = t(\hat{F})$$

This is satisfied for sample means, standard deviations, variances, medians and other sample quantiles.

3. a smoothness condition on the functional. This is not true for extreme order statistics such as the minimum and the maximum values.

But what happens when the dataset D_N is not i.i.d. sampled from a distribution F? In such non conventional configurations, the most basic version of bootstrap might fail. Examples are incomplete data (survival data, missing data), dependent data (e.g. variance of a correlated time series) and dirty data (outliers) configurations. In these cases, specific adaptations of the bootstrap procedure are required. For reason of space, we will not discuss them here. However, for a more exhaustive discussion on the limits of bootstrap, we invite the reader to refer to [121].

6.5 Randomisation tests

Randomisation tests were introduced by R.A. Fisher in 1935. The goal of a randomisation test is to help to discover some regularity (e.g. a *non random* property or pattern) in a *complicated* data set. A classic example is to take a pack of poker play-cards and check whether they were well shuffled by our poker opponent. According to the hypothesis testing terminology, randomisation tests make the null hypothesis of randomness and test this hypothesis against data. In order to test the randomness hypothesis, several random transformations of data are generated.

Suppose we are interested in some property which is related to the *order* of data. Let the original data set $D_N = \{x_1, \dots, x_N\}$ and $t(D_N)$ some statistic which is a function of the order in the data D_N. We want to test if the value of $t(D_N)$ is due only to randomness.

- An empirical distribution is generated by scrambling (or *shuffling*) R times the N elements at random. For example the jth, $j = 1, \dots, R$ scrambled data set could be $D_N^{(j)} = \{x_{23}, x_4, x_{343}, \dots\}$
- For each of the jth scrambled sets we compute a statistic $t^{(i)}$. The resulting distribution is called the *resampling distribution.*
- Suppose that the value of $t(D_N)$ is only exceeded by k of the R values of the resampling distribution.
- The probability of observing $t(D_N)$ under the null hypothesis (i.e. randomness) is only $p_t = k/R$. The null hypothesis can be accepted/rejected on the basis of p_t.

The quantity p_t plays the role of nonparametric p-value (Section 5.11.3) and it can be used, like its parametric counterpart, both to assess the evidence of the null hypothesis and to perform a decision test (e.g. refuse to play if we think cards were not sufficiently shuffled).

A bioinformatics example

Suppose we have a DNA sequence and we think that the number of repeated sequences (e.g. AGTAGTAGT) in the sample is greater than expected by chance. Let $t = 17$ be the number of repetitions. How to test this hypothesis? Let us formulate the null hypothesis that the base order is random. We can construct an empirical distribution under the null hypothesis by taking the original sample and randomly scrambling the bases $R = 1000$ times. This creates a sample with the same base frequencies as the original sample but where the order of bases is assigned at random.

Suppose that only 5 of the 1000 randomised samples has a number of repetition higher or equal than 17. The p-value (i.e. the probability of seeing $t = 17$ under the null hypothesis) which is returned by the randomisation test amounts to 0.005. You can run the randomisation test by using the R script file `randomiz.R`.

•

6.5.1 Randomisation and bootstrap

Both bootstrap and randomisation rely on resampling. But what are their peculiarities? A randomised sample is generated by scrambling the existing data (sampling without replacement) while a bootstrap sample is generated by sampling with replacement from the original sample. Also, randomisation tests are appropriate when the order or association between parts of data are assumed to convey important information. They test the null hypothesis that the order or the association is random. On the other side, bootstrap sampling aims to characterise the statistical distribution of some statistics $t(D_N)$ where the order makes no difference in the statistics (e.g. mean). Randomisation would be useless in that case since $t(D_N^{(1)}) = t(D_N^{(2)})$ if $D_N^{(1)}$ and $D_N^{(2)}$ are obtained by resampling D_N without replacement.

6.6 Permutation test

Permutation test is used to perform a nonparametric two-sample test. Consider a random sample $\{z_1, \ldots, z_M\}$ drawn from an unknown distribution $\mathbf{z} \sim F_{\mathbf{z}}(\cdot)$ and a random sample $\{y_1, \ldots, y_N\}$ from an unknown distribution $\mathbf{y} \sim F_{\mathbf{y}}(\cdot)$. For example, in a bioinformatics task the two datasets could be expression measures of a gene under M normal and N pathological conditions. Let the null hypothesis be that the two distributions are the same regardless of the analytical forms of the distributions.

Consider a (order-independent) test statistic for the observed data and call it $t(D_N, D_M)$. The rationale of the permutation test is to locate the statistic $t(D_N, D_M)$ with respect to the distribution which could be obtained if the null hypothesis were true. In order to build the null hypothesis distribution, all the possible $R = \binom{M+N}{M}$ partitionings of the $N + M$ observations in two subsets of size N and M are considered. If the null hypothesis were true, all the partitionings would be equally likely. Then for each i-th permutation ($i = 1, \ldots, R$) the permutation test computes the $t^{(i)}$ statistic. Eventually, the value $t(D_N, D_M)$ is compared with the set of values $t^{(i)}$. If the the value $t(D_N, D_M)$ falls in the $\alpha/2$ tails of the $t^{(i)}$ distribution, the null hypothesis is rejected with type I error α.

The permutation procedure will involve substantial computation unless M and N are small. When the number of permutations is too large a random sample of a large number R of permutations can be taken.

Note that when observations are drawn according to a normal distribution, it can be shown that the use of a permutation test gives results close to those obtained using the t test.

Example

Let us consider $D_4 = [74, 86, 98, 102, 89]$ and $D_3 = [10, 25, 80]$. We run a permutation test ($R = \binom{8}{4} = 70$ permutations) to test the hypothesis that the two sets belong to the same distribution (R script `s_perm.R`).

Let $t(D_N) = \hat{\mu}(D_4) - \hat{\mu}(D_3) = 51.46$. Figure 6.2 shows the position of $t(D_N)$ with respect to the null sampling distribution.

•

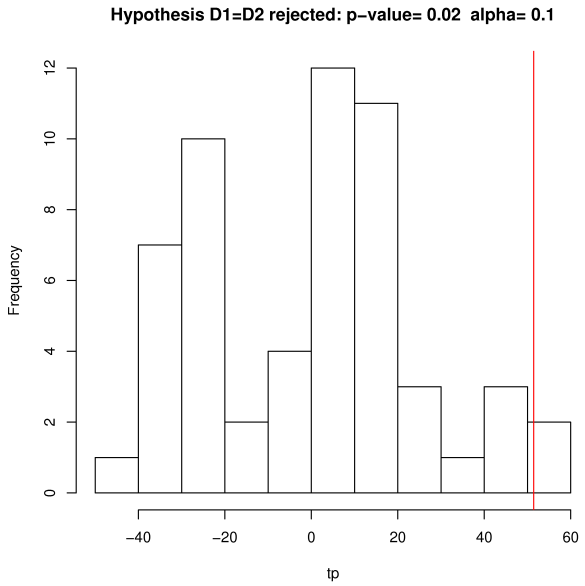


Figure 6.2: Null distribution returned by the permutation test and position (vertical red line) of the observed statistic

6.7 Considerations on nonparametric tests

Nonparametric tests are a worthy alternative to parametric approaches when no assumptions about the probability distribution may be made (e.g. in bioinformatics). It is risky, however, to consider them as a panacea, and a critical attitude towards them has to be preferred. In short terms, here you find some of the major advantages and disadvantages concerning the use of a nonparametric approach. Advantages:

- If the sample size is very small, there may be no alternative to using a nonparametric test unless the nature of the population distribution is *known exactly*.
- Nonparametric tests make fewer assumptions about the data.
- Nonparametric tests are available to analyse data that are inherently in ranks (e.g. taste of food), classificatory or categorical.
- Nonparametric tests are typically more intuitive and easier to implement.

Disadvantages:

- They involve high computational costs.
- The large availability of statistical software makes possible the potential misuse of statistical measures.
- A nonparametric test is less powerful than a parametric one when the assumptions of the parametric test are met.
- Assumptions are associated with most nonparametric statistical tests, namely, that the observations are independent.

6.8 Exercises

1. Suppose you want to estimate the skewness γ of the uniform r.v. $\mathbf{z} \sim \mathcal{U}[-2, 3]$ by using a dataset of size $N = 10$. By using R and its random generator, first plot the sampling distribution then estimate the bias and the variance of the following estimators:
   1. $\hat{\gamma} = \frac{\frac{1}{N}\sum_i (z_i - \hat{\mu})^3}{\hat{\sigma}^3}$
   2. $\hat{\gamma} = \frac{\frac{1}{N}\sum_i |z_i - \hat{\mu}|^3}{\hat{\sigma}^3}$
   3. $\hat{\gamma} = 1$

   Before each random generation set the seed to zero. Hint: the skewness of a uniform continuous variable is equal to 0.

2. Suppose you want to estimate the skewness γ of the uniform r.v. $\mathbf{z} \sim \mathcal{U}[-2, 3]$ by using a dataset of size $N = 10$. By using R and its random generator, first generate a dataset D_N with $N = 10$. By using the jacknife, plot the sampling distribution, then estimate the bias and the variance of the following estimators,
   1. $\hat{\gamma} = \frac{\frac{1}{N}\sum_i (z_i - \hat{\mu})^3}{\hat{\sigma}^3}$
   2. $\hat{\gamma} = \frac{\frac{1}{N}\sum_i |z_i - \hat{\mu}|^3}{\hat{\sigma}^3}$
   3. $\hat{\gamma} = 1$

   Compare the results with the ones of the exercise before. Before each random generation set the seed to zero.

3. Suppose you want to estimate the skewness γ of the uniform r.v. $\mathbf{z} \sim \mathcal{U}[-2, 3]$ by using a dataset of size $N = 10$. By using R and its random generator, first generate a dataset D_N with $N = 10$. By using the bootstrap method, plot the sampling distribution, then estimate the bias and the variance of the following estimators,
   1. $\hat{\gamma} = \frac{\frac{1}{N}\sum_i (z_i - \hat{\mu})^3}{\hat{\sigma}^3}$
   2. $\hat{\gamma} = \frac{\frac{1}{N}\sum_i |z_i - \hat{\mu}|^3}{\hat{\sigma}^3}$
   3. $\hat{\gamma} = 1$

   Compare the results with the ones of the two exercises before. Before each random generation set the seed to zero.

4. Let us consider a r.v. $\mathbf{z}$ such that $E[\mathbf{z}] = \mu$ and $\mathrm{Var}[\mathbf{z}] = \sigma^2$. Suppose we want to estimate from i.i.d. dataset D_N the parameter $\theta = \mu^2 = (E[\mathbf{z}])^2$. Let us consider three estimators:

$$\hat{\theta}_1 = \left(\frac{\sum_{i=1}^N z_i}{N} \right)^2$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^N z_i^2}{N}$$

$$\hat{\theta}_3 = \frac{(\sum_{i=1}^N z_i)^2}{N}$$

   - Are they unbiased?
   - Compute analytically the bias of the three estimators. Hint: use (3.3.30).
   - By using R, verify the result above by Monte Carlo simulation using different values of N.
   - By using R, estimate the bias of the three estimators by bootstrap.

   Solution: See the file `Exercise1.pdf` in the directory `gbcode/exercises` of the companion R package (Appendix F).

Chapter 7

A statistical framework of supervised learning

7.1 Introduction

A supervised learning problem can be described in statistical terms by the following elements:

1. A vector of n random input variables $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$, whose values are i.i.distributed according to an unknown probabilistic distribution $F_{\mathbf{x}}(\cdot)$.
2. A *target* operator which transforms the input values into *outputs* $\mathbf{y} \in \mathcal{Y}$ according to an unknown conditional probability distribution $F_{\mathbf{y}}(y|\mathbf{x} = x)$.
3. A collection D_N of N input/output data points $\langle x_i, y_i \rangle$, $i = 1, \ldots, N$, called the *training set* and drawn according to the joint input/output density $F_{\mathbf{x},\mathbf{y}}(x, y)$.
4. A *learning machine* or learning algorithm which, on the basis of the training set D_N, returns an estimation (or prediction) of the target for an input x. The input/output function estimated by the learning machine is called *hypothesis* or *model*.

Note that in this definition we encounter most of the notions presented in the previous chapters: probability distribution, conditional distribution, estimation.

Examples

Several practical problems can be seen as instances of a supervised learning problem:

- Predict whether a patient, hospitalised due to a heart attack, will have a second heart attack, on the basis of demographic, diet and clinical measurements.
- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
- Identify the risk factors for breast cancer, based on clinical, demographic and genetic variables.
- Classify the category of a text email (spam or not) on the basis of its text content.
- Characterise the mechanical property of a steel plate on the basis of its physical and chemical composition.

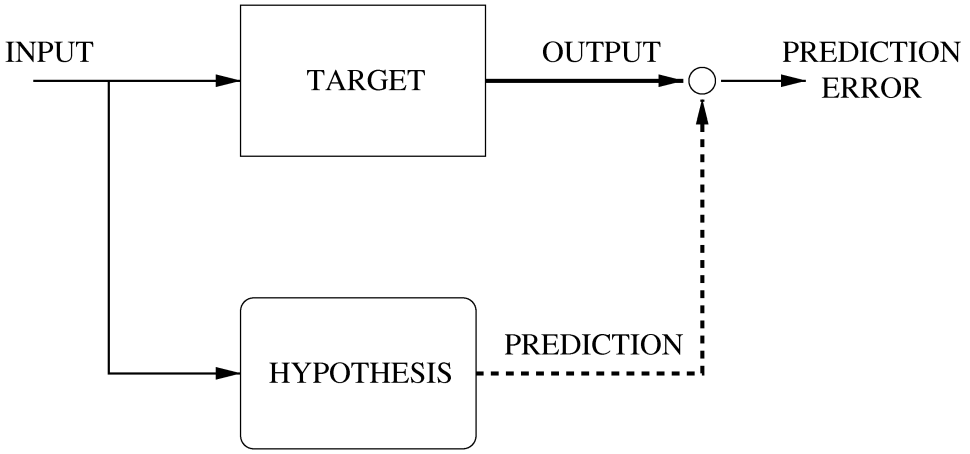


Figure 7.1: The supervised learning setting. The target operator returns an output for each input according to a fixed but unknown probabilistic law. The hypothesis predicts the value of the target output when entered with the same input.

In the case of the spam categorisation problem, the input vector may be a vector of size n where n is the number of the most used English words and the ith component of $\mathbf{x}$ represents the frequency of the ith word in the email text. The output $\mathbf{y}$ is a binary class which takes two values: {SPAM,NO.SPAM}. The training set is a set of emails previously labeled by the user as SPAM and NO.SPAM. The goal of the learning machine is to create a classification function which, once a vector x of word frequencies is presented, should be able to classify correctly the nature of the email.

•

A learning machine is nothing more than a particular instance of an estimator (5.4.7) whose goal is to estimate the parameters of the joint distribution $F_{\mathbf{x},\mathbf{y}}(y, x)$ (or sometimes of the conditional distribution $F_{\mathbf{y}}(y|\mathbf{x} = x)$) on the basis of a training set D_N, i.e. a set of i.i.d. realisations of the pair $\mathbf{x}$ and $\mathbf{y}$. The goal of a *learning machine* is to return a hypothesis with low prediction error, i.e. a hypothesis which computes an accurate estimate of the output of the target when the same test value is an input to the target and the predictor (Fig. 7.1). The prediction error is also usually called *generalisation error*, since it measures the capacity of the learned hypothesis to generalise to previously unseen test samples. *A learning algorithm generalises well if it returns an accurate prediction for i.i.d. test data, i.e. input/output pairs which are independent from the training set yet are generated by the same joint distribution* $F_{\mathbf{x},\mathbf{y}}(x, y)$. We insist on the importance of the two "i" in the i.i.d. assumption: test data are supposed i) to be generated by the same distribution underlying the training set but ii) to be independent from the training set.

We will only consider hypotheses in the form $h(\cdot, \alpha)$ where $\alpha \in \Lambda^*$ is a vector of model parameters[1] or weights. Therefore, henceforth, we will denote an hypothesis $h(\cdot, \alpha)$ by the corresponding vector $\alpha \in \Lambda^*$. As we will see later, examples of hypothesis are linear models $h(x, \alpha) = x^T\alpha$ (Section 9.1) where α represents the coefficients of the model, or feed-forward neural networks (Section 10.1.1) where α is the set of values taken by the weights of the neural architecture.

[1]It is important to remark that by model parameter we refer here to a tunable/trainable weight of the hypothesis function and not to the target of the estimation procedure as in Section 5.1.1

Let α_N be the hypothesis returned by the learning machine on the basis of the training set, and define G_N its generalisation error. The goal of the learning machine is then to seek the hypothesis α_N which minimises the value G_N.

In these terms, the learning problem could appear as a simple problem of optimisation which consists of searching the hypothesis α which yields the lowest generalisation error. Unfortunately the reality is not that simple, since the learning machine cannot measure directly G_N but only return an estimate of this quantity, denoted by $\hat{G}_N$. Moreover, what makes the problem still more complex is that the same finite training set is employed both to select α_N and to estimate G_N, thus inducing a strong correlation between these two quantities.

The common supervised learning practice to minimise the quantity G_N consists in

1. decomposing the set of hypothesis Λ^* into a nested sequence of hypothesis classes (or *model structures*) $\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \Lambda_S$ of increasing capacity (or expressiveness) s with $\Lambda^* = \cup_{s=1}^{S} \Lambda_s$

2. implementing a search procedure at two nested levels [123] (Fig. 7.2). The inner level, also known as *parametric identification*, considers a single class of hypotheses Λ_s and uses a method or *algorithm* to select a hypothesis $h(\cdot, \alpha_N^s)$ from this class. The algorithm typically implements a procedure of multivariate optimisation in the space of model parameters of the class Λ_s, which can be solved by (conventional) optimisation techniques. Examples of parametric identification procedures which will be presented in subsequent chapters are linear least-squares for linear models or back-propagated gradient-descent for feedforward neural networks [163]. The outer level, also called *structural identification*, ranges over nested classes of hypotheses Λ_s, $(s = 1, \ldots, S)$, and executes for each of them the parametric routine returning the vector α_N^s. The outcome of the parametric identification is used to assess the class Λ_s through a *validation* procedure which returns the estimate $\hat{G}_N^s$ on the basis of the finite training set. It is common to use nonparametric techniques to assess the quality of a predictor like the bootstrap (Section 6.4) or cross-validation [172] based on the jacknife strategy (Section 6.3).

3. selecting the best hypothesis in the set $\{\alpha_N^s\}$, with $s = 1, \ldots, S$, according to the assessments $\left\{\hat{G}_N^s\right\}$ produced by the validation step. This final step, which returns the model to be used for prediction, is usually referred to as the *model selection* procedure. Instances of model selection include the problem of choosing the degree of a polynomial model or the problem of determining the best number of hidden nodes in a neural network [25].

The outline of the chapter is as follows. Section 7.2 introduces the supervised learning problem in statistical terms. We will show that classification (Section 7.3) and regression (Section 7.4) can be easily cast in this framework. Section 7.5 introduces the statistical assessment of a learning machine while Section 7.6 reports some results from the work of Prof. Vapnik on statistical learning and in particular the formalisation of the notion of capacity of a learning machine. Section 7.7 discusses the notion of generalisation error and its bias/variance decomposition. Section 7.9 introduces the supervised learning procedure and its decomposition in structural and parametric identification. Model validation and in particular cross validation, a technique for estimating the generalisation error on the basis of a finite number of data, are introduced in Section 7.10.

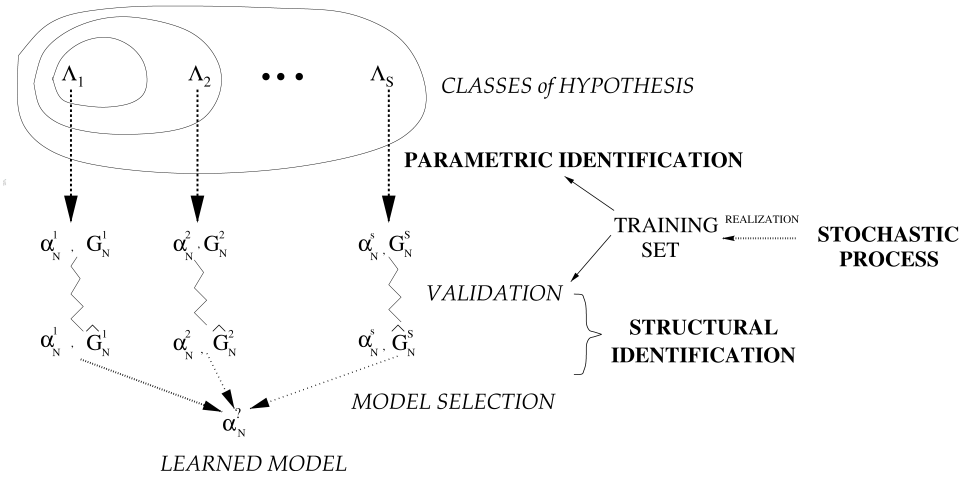


Figure 7.2: The learning problem and its decomposition in parametric and structural identification. The larger is the class of hypothesis Λ_s, the large is its expressive power in terms of functional relationships.

7.2 Estimating dependencies

This section details the main actors of the supervised learning problem:

- A data generator of random input vectors $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ independently and identically distributed (i.i.d) according to some unknown (but fixed) probability distribution $F_{\mathbf{x}}(x)$. The variable $\mathbf{x}$ is called the *independent variable*. It is helpful to distinguish between cases in which the experimenter has a complete control over the values of $\mathbf{x}$ and those cases in which she does not. When the nature of inputs is completely random, we consider x as a realisation of the random variable $\mathbf{x}$ having probability law $F_{\mathbf{x}}(\cdot)$. When the experimenter's control is complete, we can regard $F_{\mathbf{x}}(\cdot)$ as describing the relative frequencies with which different values for x are set.

- A *target* operator, which transforms the input $\mathbf{x}$ into the output value $\mathbf{y} \in \mathcal{Y}$ according to some unknown (but fixed) conditional distribution

$$F_{\mathbf{y}}(y|\mathbf{x} = x) \tag{7.2.1}$$

  (this includes the simplest case where the target implements some deterministic function $\mathbf{y} = f(\mathbf{x})$). The conditional distribution (7.2.1) formalizes the stochastic dependency between inputs and output.

- A *training set* $D_N = \{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_N, y_N\rangle\}$ made of N pairs (or training examples) $\langle x_i, y_i\rangle \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ independent and identically distributed (i.i.d) according to the joint distribution

$$F_{\mathbf{z}}(z) = F_{\mathbf{x},\mathbf{y}}(\langle x, y\rangle) \tag{7.2.2}$$

  Note that, as in Section 5.4, the observed training set $D_N \in \mathcal{Z}^N = (\mathcal{X} \times \mathcal{Y})^N$ is considered here as the realisation of a random variable $\mathbf{D}_N$.

- A *learning machine* having three components:

  1. A class of *hypothesis* functions $h(\cdot, \alpha)$ with $\alpha \in \Lambda$. We consider only the case where the functions $h(\cdot, \alpha) \in \mathcal{Y}$ are single valued mappings.

2. A *loss* function $L(\cdot,\cdot)$ associated with a particular y and a particular $h(x)$, whose value $L(y,h(x))$ measures the discrepancy between the output y and the prediction $h(x)$. For a given hypothesis $h(\cdot,\alpha)$, the *functional risk* is the loss average over the $\mathcal{X}\mathcal{Y}$-domain

$$R(\alpha) = E_{\mathbf{xy}}[\mathbf{L}] = \int_{\mathcal{X},\mathcal{Y}} L(y,h(x,\alpha))dF_{\mathbf{x},\mathbf{y}}(x,y) = \int_{\mathcal{X},\mathcal{Y}} L(y,h(x,\alpha))p(x,y)dxdy \quad (7.2.3)$$

Note that $\mathbf{L}$ is random since $\mathbf{x}$ and $\mathbf{y}$ are random test points (i.i.drawn from the same distribution (7.2.2) of the training set) while the hypothesis $h(\cdot,\alpha)$ is given. This is the expected loss if we test the hypothesis $h(\cdot,\alpha)$ over an infinite amount of i.i.d. input/output pairs generated by (7.2.2). For the class Λ of hypothesis we define

$$\alpha_0 = \arg\min_{\alpha\in\Lambda} R(\alpha) \quad (7.2.4)$$

as the hypothesis in the class Λ which has the lowest functional risk. Here, we assume for simplicity that there exists a minimum value of $R(\alpha)$ achievable by a function in the class Λ. We define with $R(\alpha_0)$ the *functional risk of the class* Λ *of hypotheses*.

3. If instead of a single class of hypothesis we consider the set Λ^* containing *all* possible single valued mappings $h : \mathcal{X} \to \mathcal{Y}$, we may define the quantity

$$\alpha^* = \arg\min_{\alpha\in\Lambda^*} R(\alpha) \quad (7.2.5)$$

and

$$R^* = R(\alpha^*) \quad (7.2.6)$$

as the absolute minimum rate of functional risk. Note that this quantity is ideal since it requires the complete knowledge of the distribution underlying the data. In a classification setting, the optimal model with parameters α^* is called the *Bayes classifier* and $R(\alpha^*)$ the *Bayes error* (Section 7.3.1). In a regression setting (Section 7.4) where $\mathbf{y} = f(x) + \mathbf{w}$ and the loss function is quadratic, $h(\cdot,\alpha^*) = f(\cdot)$ and $R(\alpha^*)$ amounts to the variance of $\mathbf{w}$.

4. An *algorithm* $\mathcal{L}$ of parametric identification which takes as input the training set D_N and returns as output one hypothesis function $h(\cdot,\alpha_N)$ with $\alpha_N \in \Lambda$. Here, we will consider only the case of *deterministic* and *symmetric* algorithms. This means respectively that they always give the same $h(\cdot,\alpha_N)$ for the same data set D_N and that they are insensitive to the ordering of the examples in D_N.

   The parametric identification of the hypothesis is done according to ERM (Empirical Risk Minimisation) inductive principle [182] where

$$\alpha_N = \alpha(D_N) = \arg\min_{\alpha\in\Lambda} R_{\text{emp}}(\alpha) \quad (7.2.7)$$

minimizes the *empirical risk* (also know as *training error* or *apparent error*)

$$R_{\text{emp}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,h(x_i,\alpha)) \quad (7.2.8)$$

constructed on the basis of the data set D_N.

This formulation of a supervised learning problem is quite general, given that it includes two basic statistical problems:

1. the problem of classification (also known as *pattern recognition*),
2. the problem of regression estimation.

These two problems and their link with supervised learning will be discussed in the following sections.

7.3 Dependency and classification

Classification is one of the most common problem in statistics. It consists in exploring the association between categorical dependent variables and independent variables which can take either continuous or discrete values. The problem of *classification* is formulated as follows: consider an input/output stochastic dependence which can be described by a joint distribution $F_{\mathbf{x},\mathbf{y}}(\cdot)$, such that once an input vector x is given, $\mathbf{y} \in \mathcal{Y} = \{c_1, \ldots, c_K\}$ takes a value among K different classes. In the example of spam email classification, $K = 2$ and c_1 =SPAM, c_2 =NO.SPAM. We assume that the dependence is described by a conditional discrete probability distribution $\text{Prob}\{\mathbf{y} = c_k | \mathbf{x} = x\}$ that satisfies

$$\sum_{k=1}^{K} \text{Prob}\{\mathbf{y} = c_k | x\} = 1$$

This means that observations are noisy and follow a probability distribution. In other terms, given an input x, $\mathbf{y}$ does not always take the same value. Pretending to have a zero-error classification in this setting is then completely unrealistic.

Example

Consider a stochastic dependence where $\mathbf{x}$ represents a year's month and $\mathbf{y}$ is a categorical variable representing the weather situation in Brussels. Suppose that $\mathbf{y}$ may take only the two values {RAIN,NO.RAIN}. The setting is stochastic since you might have rainy August and some rare sunny December days. Suppose that the conditional probability distribution of $\mathbf{y}$ is represented in Figure 7.3. This figure plots $\text{Prob}\{\mathbf{y} = \text{RAIN} | \mathbf{x} = \text{month}\}$ and $\text{Prob}\{\mathbf{y} = \text{NO.RAIN} | x = \text{month}\}$ for each month. Note that for each month the probability constraint is respected:

$$\text{Prob}\{\mathbf{y} = \text{RAIN} | \mathbf{x} = \text{month}\} + \text{Prob}\{\mathbf{y} = \text{NO.RAIN} | \mathbf{x} = \text{month}\} = 1$$

•

A classifier is a particular instance of estimator which for a given x is expected to return an estimate $\hat{y} = \hat{c} = h(x, \alpha)$ which takes a value in $\{c_1, \ldots, c_K\}$. Once a cost function is defined, the problem of classification can be expressed in terms of the formalism introduced in the previous section. An example of cost function is the indicator function (taking only two values: zero and one)

$$L(c, \hat{c}) = \begin{cases} 0 & \text{if } c = \hat{c} \\ 1 & \text{if } c \neq \hat{c} \end{cases} \tag{7.3.9}$$

also called the *0/1 loss*. However, we can imagine situations where some misclassifications are worse than others. In this case, it is better to introduce a *loss matrix* $L_{(K \times K)}$ where the element $L_{(jk)} = L_{(c_j, c_k)}$ denotes the cost of the misclassification

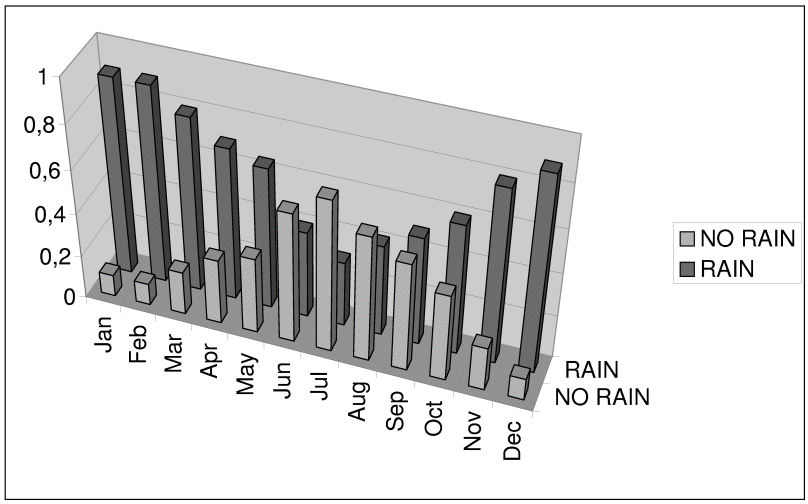


Figure 7.3: Conditional distribution Prob $\{\mathbf{y}|x\}$ where x is the current month and $\mathbf{y}$ is the random weather state. For example the column corresponding to x =Dec and y =RAIN returns the conditional probability of RAIN in December.

when the predicted class is $\hat{c}(x) = c_j$ and the correct class is c_k. This matrix must be null on the diagonal and non negative everywhere else. In practical cases the definition of a loss matrix could be quite challenging since it should take into account and combine several criteria, some easy to quantify (e.g. financial costs) and some much less (e.g. ethical considerations)[2]. Note that in the case of the 0-1 loss function (Equation 7.3.9) all the elements outside the diagonal are equal to one.

The goal of the classification procedure for a given x is to find the predictor $\hat{c}(x) = h(x, \alpha)$ that minimises the quantity

$$\sum_{k=1}^{K} L_{(\hat{c}(x),c_k)} \text{Prob}\{\mathbf{y} = c_k|x\} \tag{7.3.10}$$

which is an average of the $\hat{c}(x)$ row of the loss matrix weighted by the conditional probabilities of observing $\mathbf{y} = c_k$. Note that the average of the above quantity over the $\mathcal{X}$ domain

$$\int_{\mathcal{X}} \sum_{k=1}^{K} L_{(\hat{c}(x),c_k)} \text{Prob}\{\mathbf{y} = c_k|x\}\, dF_{\mathbf{x}} = \int_{\mathcal{X},\mathcal{Y}} L(y, h(x, \alpha)) dF_{\mathbf{x},\mathbf{y}} = R(\alpha) \tag{7.3.11}$$

corresponds to the functional risk (7.2.3).

The problem of classification can then be seen as a particular instance of the more general supervised learning problem described in Section 7.2.

[2]By default, any automatic classifier (and the associated decision maker) implicitly or explicitly embeds a loss function weighting often highly heterogeneous criteria. For instance, the Tesla automatic braking systems (implicitly or explicitly) assigns a cost to false positives (e.g. a bag wrongly identified as a pedestrian) and false negatives (e.g. a pedestrian mistaken for a bag).

7.3.1 The Bayes classifier

It can be shown that the optimal classifier $h(\cdot, \alpha_0)$ where α_0 is defined as in (7.2.4) is the one that returns for all x

$$c^*(x) = h(x, \alpha_0) = \arg \min_{c_j \in \{c_1, \ldots, c_K\}} \sum_{k=1}^{K} L_{(j,k)} \text{Prob}\{\mathbf{y} = c_k | x\} \tag{7.3.12}$$

The optimal classifier is also known as the *Bayes classifier*. In the case of a 0-1 loss function the optimal classifier returns

$$c^*(x) = \arg \min_{c_j \in \{c_1, \ldots, c_K\}} \sum_{k=1:K, k \neq j} \text{Prob}\{\mathbf{y} = c_k | x\} \tag{7.3.13}$$

$$= \arg \min_{c_j \in \{c_1, \ldots, c_K\}} (1 - \text{Prob}\{\mathbf{y} = c_j | x\}) \tag{7.3.14}$$

$$= \arg \min_{c_j \in \{c_1, \ldots, c_K\}} \text{Prob}\{\mathbf{y} \neq c_j | x\} = \arg \max_{c_j \in \{c_1, \ldots, c_K\}} \text{Prob}\{\mathbf{y} = c_j | x\} \tag{7.3.15}$$

The Bayes decision rule selects the j, $j = 1, \ldots, K$, that *maximizes* the posterior probability $\text{Prob}\{\mathbf{y} = c_j | x\}$.

Example

Consider a classification task where $\mathcal{X} = \{1, 2, 3, 4, 5\}$, $\mathcal{Y} = \{c_1, c_2, c_3\}$ and the loss matrix and the conditional probability values are given in the following figures.

REAL / PRED	C1	C2	C3
C1	0	1	5
C2	20	0	10
C3	2	1	0

LOSS MATRIX

x	C1	C2	C3
1	0.1	0.6	0.3
2	0.2	0.8	0.0
3	0.9	0.04	0.06
4	0.5	0.25	0.25
5	0.3	0.1	0.6

CONDITIONAL PROBABILITY

Let us focus on the optimal classification for $x = 2$. According to (7.3.12) the Bayes classification rule for $x = 2$ returns

$$c^*(2) = \arg \min_{k=1,2,3} \{L_{11}\text{Prob}\{\mathbf{y} = c_1 | \mathbf{x} = 2\} + L_{12}\text{Prob}\{\mathbf{y} = c_2 | \mathbf{x} = 2\} + L_{13}\text{Prob}\{\mathbf{y} = c_3 | \mathbf{x} = 2\},$$
$$L_{21}\text{Prob}\{\mathbf{y} = c_1 | \mathbf{x} = 2\} + L_{22}\text{Prob}\{\mathbf{y} = c_2 | \mathbf{x} = 2\} + L_{23}\text{Prob}\{\mathbf{y} = c_3 | \mathbf{x} = 2\},$$
$$L_{31}\text{Prob}\{\mathbf{y} = c_1 | \mathbf{x} = 2\} + L_{32}\text{Prob}\{\mathbf{y} = c_2 | \mathbf{x} = 2\} + L_{33}\text{Prob}\{\mathbf{y} = c_3 | \mathbf{x} = 2\}\}$$
$$= \arg \min_{k=1,2,3} \{0 * 0.2 + 1 * 0.8 + 5 * 0.0, 20 * 0.2 + 0 * 0.8 + 10 * 0.0,$$
$$2 * 0.2 + 1 * 0.8 + 0.0 * 0\} = \arg \min_{k=1,2,3} \{1, 4, 1.2\} = 1$$

What would have been the Bayes classification in the 0-1 case?

•

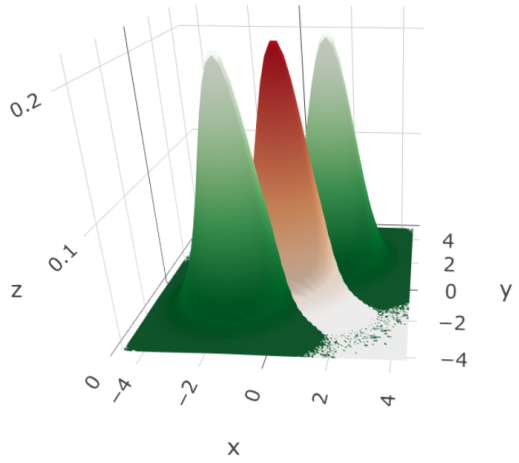


Figure 7.4: Class conditional distributions: the green class is distributed as a mixture of two gaussians while the red class as a gaussian.

7.3.2 Inverse conditional distribution

An important quantity, often used in classification algorithms, is the *inverse conditional distribution*. According to the Bayes theorem (3.1.20) we have that

$$\text{Prob}\{\mathbf{y} = c_k | \mathbf{x} = x\} = \frac{\text{Prob}\{\mathbf{x} = x | \mathbf{y} = c_k\}\,\text{Prob}\{\mathbf{y} = c_k\}}{\sum_{k=1}^{K} \text{Prob}\{\mathbf{x} = x | \mathbf{y} = c_k\}\,\text{Prob}\{\mathbf{y} = c_k\}} \tag{7.3.16}$$

and that

$$\text{Prob}\{\mathbf{x} = x | \mathbf{y} = c_k\} = \frac{\text{Prob}\{\mathbf{y} = c_k | \mathbf{x} = x\}\,\text{Prob}\{\mathbf{x} = x\}}{\sum_{x} \text{Prob}\{\mathbf{y} = c_k | \mathbf{x} = x\}\,\text{Prob}\{\mathbf{x} = x\}}. \tag{7.3.17}$$

The above relation means that by knowing the *a-posteriori* conditional distribution $\text{Prob}\{\mathbf{y} = c_k | \mathbf{x} = x\}$ and the *a-priori* distribution $\text{Prob}\{\mathbf{x} = x\}$, we can derive the *inverse conditional distribution* $\text{Prob}\{\mathbf{x} = x | \mathbf{y} = c_k\}$. This distribution is replaced by a density if $\mathbf{x}$ is continuous and is also known as the *class conditional density*. This distribution characterises the values of the inputs x for a given class c_k.

Shiny dashboard

The Shiny dashboard `classif2.R` illustrates a binary classification task where $\mathbf{x} \in \mathbb{R}^2$ and the two classes are *green* and *red*. The green and the class conditional distributions (7.3.17) are a mixture of two gaussians (Section 3.7.2) and a unimodal gaussian, respectively (Figure 7.4). Figure 7.5 illustrates the associated conditional distribution (7.3.16) if the two classes have an equal a-priori probability ($\text{Prob}\{\mathbf{y} = \text{red}\} = \text{Prob}\{\mathbf{y} = \text{green}\}$). Figure 7.6 shows the scattering of a set of $N = 500$ points sampled according to the class-conditional distributions in Figure 7.4.

•

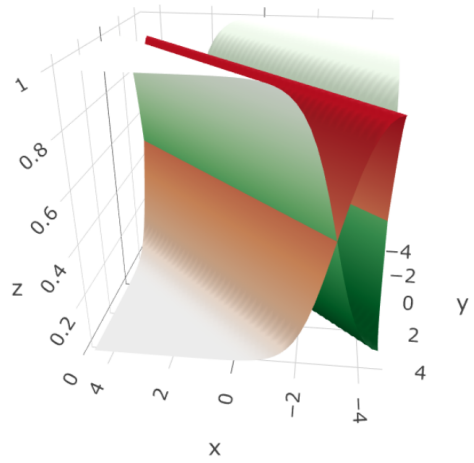


Figure 7.5: The a-posteriori conditional distribution associated to the class-conditional distributions (equal a-priori probability) in Figure 7.4.

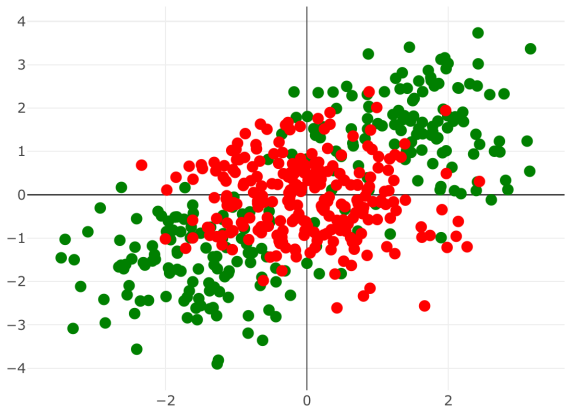


Figure 7.6: Dataset sampled according to the class-conditional distributions (equal a-priori probability) in Figure 7.4.

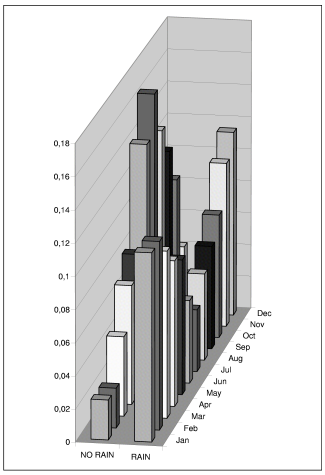


Figure 7.7: Inverse conditional distribution of the distribution in Figure 7.3

Example

Suppose we want to know during which months it is most probable to have rain. This boils down to have the distribution of x for $\mathbf{y} = RAIN$. Figure 7.7 plots the inverse conditional distributions $\text{Prob}\{\mathbf{x} = \text{month}|\mathbf{y} = \text{RAIN}\}$ and $\text{Prob}\{\mathbf{x} = \text{month}|\mathbf{y} = \text{NO.RAIN}\}$ according to (7.3.17) when we assume that the a priori distribution is uniform (i.e. $\text{Prob}\{\mathbf{x} = x\} = 1/12$ for all x).

Note that

$$\sum_{\text{month}} \text{Prob}\{\mathbf{x} = \text{month}|\mathbf{y} = \text{NO.RAIN}\} = \\ = \sum_{\text{month}} \text{Prob}\{\mathbf{x} = \text{month}|\mathbf{y} = \text{RAIN}\} = 1$$

•

7.4 Dependency and regression

Consider the stochastic relationship between two continuous random variables $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}$ described by

$$F_{\mathbf{x},\mathbf{y}}(x, y) \tag{7.4.18}$$

This means that to each vector x sampled according to the $F_x(x)$ there corresponds a scalar y sampled from $F_{\mathbf{y}}(y|\mathbf{x} = x)$. Assume that a set of N input/output observations is available. The estimation of the stochastic dependence on the basis of the empirical dataset requires the estimation of the conditional distribution $F_{\mathbf{y}}(y|x)$. This is known to be a difficult problem but for prediction purposes, most of the time, it is sufficient to estimate the conditional expectation

$$f(x) = E_{\mathbf{y}}[\mathbf{y}|x] = \int_{\mathcal{Y}} y dF_{\mathbf{y}}(y|x) \tag{7.4.19}$$

also known as the *regression function*.

The regression function is also related to the functional risk

$$R(\alpha) = \int L(y, h(x,\alpha)) dF_{\mathbf{x},\mathbf{y}}(x,y) = \int (y - h(x,\alpha))^2 dF_{\mathbf{x},\mathbf{y}}(x,y) \qquad (7.4.20)$$

for the quadratic loss $L(y,h) = (y-h)^2$. From (3.6.63) it can be shown that the minimum (7.2.4) is attained by the regression function $h(\cdot, \alpha_0) = f(\cdot)$ if the function f belongs to the set $h(x,\alpha)$, $\alpha \in \Lambda$.

Once defined the regression function f, the input/output stochastic dependency (7.4.18) is commonly represented in the *regression plus noise* form

$$\mathbf{y} = f(x) + \mathbf{w} = E_{\mathbf{y}}[\mathbf{y}|x] + \mathbf{w} \qquad (7.4.21)$$

where $\mathbf{w}$ denotes the noise term and satisfies $E[\mathbf{w}] = 0$ and $E[\mathbf{w}^2] = \sigma^2_{\mathbf{w}}$. The role of the noise is to make explicit that some variability of the target cannot be explained by the regression function f. Notice that the assumption of an additive noise $\mathbf{w}$ independent of x is common in statistical literature and is not overly restrictive. In fact, many other conceivable signal/noise models can be transformed into this form.

The problem of estimating the regression function (7.4.19) is then a particular instance of the supervised learning problem described in Section 7.2, where the learning machine is assessed by a quadratic cost function. Examples of learning algorithms for regression will be discussed in Section 9.1 and Section 10.1.

7.5 Assessment of a learning machine

A learning machine works well if it exhibits good generalisation, i.e. if it is able to perform good predictions for unseen input values, which are not part of the training set but that are generated by the same input/output distribution (7.2.2) underlying the training set. This ability is commonly assessed by the amount of bad predictions, measured by the *generalisation error*. The generalisation error of a learning machine can be evaluated at two levels:

Hypothesis: Let α_N be the hypothesis returned by a learning algorithm for a training set D_N according to the ERM principle (Eq. (7.2.7)). The functional risk $R(\alpha_N)$ in (7.2.3) represents the *generalisation error of the hypothesis* α_N. This quantity is also known as *conditional error rate* [96] since it is conditional on a given training set D_N.

Algorithm: Let us define the average of the loss L for a given input x over the ensemble of training sets of size N as

$$g_N(x) = E_{\mathbf{D}_N,\mathbf{y}}[\mathbf{L}|\mathbf{x} = x] = \int_{\mathcal{Z}^N,\mathcal{Y}} L(y, h(x,\alpha_N)) dF_{\mathbf{y}}(y|x) dF^N_{\mathbf{z}}(D_N) \qquad (7.5.22)$$

where $F^N_{\mathbf{z}}(D_N)$ is the distribution of the i.i.d. dataset D_N. In this expression $\mathbf{L}$ is a function of the random variables $\mathbf{D}_N$ (through h) and $\mathbf{y}$, while the test input x is fixed. In the case of a quadratic loss function, this quantity corresponds to the *mean squared error* (MSE) defined in Section 5.5.6. By averaging the quantity (7.5.22) over the $\mathcal{X}$ domain we have

$$G_N = \int_{\mathcal{X}} g_N(x) dF_{\mathbf{x}}(x) = E_{\mathbf{D}_N} E_{\mathbf{x},\mathbf{y}}[L(\mathbf{y}, h(\mathbf{x}, \boldsymbol{\alpha}_N))] \qquad (7.5.23)$$

that is the *generalisation error of the algorithm* $\mathcal{L}$ (also known as *expected error rate* [65] or *expected test error* [96]).

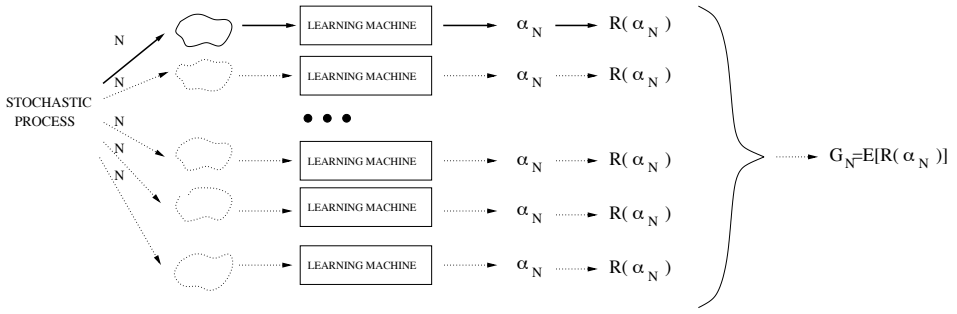


Figure 7.8: Functional risk vs. MISE

From (7.2.3) and (7.5.23) we obtain that

$$G_N = E_{\mathbf{D}_N}[R(\boldsymbol{\alpha}_N)]$$

where $R(\boldsymbol{\alpha}_N)$ is random because of the dependence on $\mathbf{D}_N$ (Figure 7.8).

In the case of a quadratic loss function, the quantity

$$\text{MISE} = E_{\mathbf{D}_N} E_{\mathbf{x},\mathbf{y}}[(\mathbf{y} - h(\mathbf{x}, \boldsymbol{\alpha}_N))^2] \tag{7.5.24}$$

takes the name of *mean integrated squared error* (MISE).

The two criteria correspond to two different ways of assessing the learning machine: the first is a measure to assess the specific hypothesis (7.2.7) chosen by ERM, the second assesses the average performance of the algorithm over training sets with N observations. According to the hypothesis-based approach the goal of learning is to find, on the basis of observations, the hypothesis that minimises the functional risk. According to the algorithmic-based approach the goal is to find, on the basis of observations, the algorithm which minimises the generalisation error. The two criteria will be detailed in Section 7.6 and 7.7, respectively. Note that both quantities requires the knowledge of $F_{\mathbf{x},\mathbf{y}}$ which is unfortunately unknown in real situations. A key issue in machine learning is then to take advantage of observable quantities, i.e. quantities that may be computed on the basis of the observed dataset, to estimate or approximate the measures discussed above. An important quantity in this sense is the empirical risk (7.2.8) which has however to be carefully considered in order to avoid too optimistic evaluations of the learning machine accuracy.

7.5.1 An illustrative example

The notation introduced in Section 7.2 and 7.5 is rigorous but it may appear hostile to the practitioner. In order to make the statistical concepts more affordable we

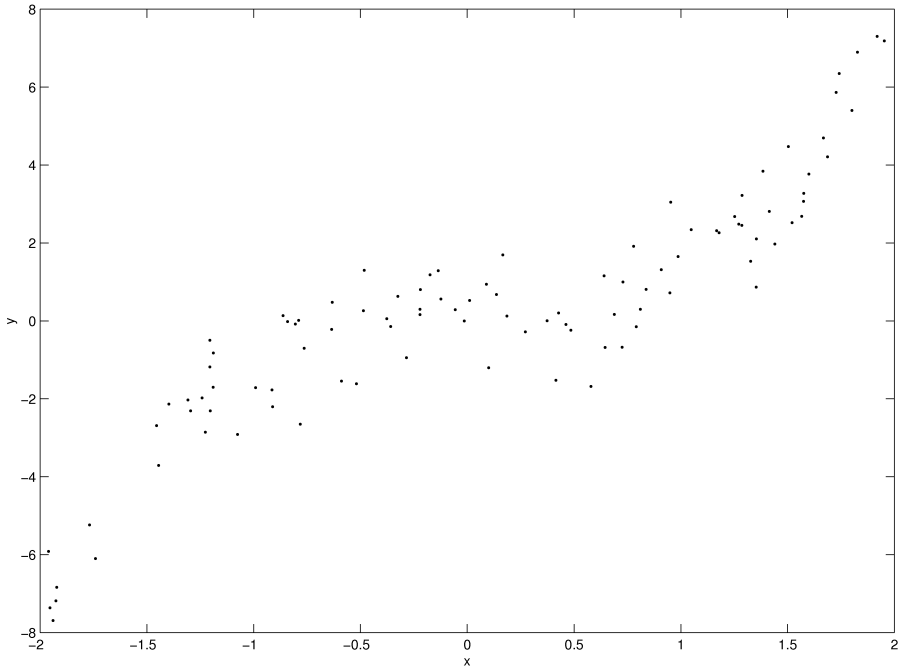


Figure 7.9: Training set (dots) obtained by sampling uniformly in the interval $[-2, 2]$ an input/output distribution with regression function $f(x) = x^3$ and unit variance.

present a simple example to illustrate these concepts. We consider a supervised learning regression problem where :

- The input is a scalar random variable $\mathbf{x} \in \mathbb{R}$ with a uniform probability distribution over the interval $[-2, 2]$.
- The target is distributed according to a conditional Gaussian distribution

$$p_{\mathbf{y}}(y|\mathbf{x} = x) = \mathcal{N}(x^3, 1) \tag{7.5.25}$$

  where the conditional expected value $E[\mathbf{y}|x]$ is the regression function $f(x) = x^3$ and the noise $\mathbf{w}$ has a unit variance.
- The training set $D_N = \{\langle x_i, y_i \rangle\}, i = 1, \ldots, N$ consists of $N = 100$ i.i.d. pairs (Figure 7.9) generated according to the distribution 7.5.25. Note that this training set can be easily generated with the following R commands

```
## script regr.R

N<-100
X<-runif(N,-2,2)
Y=X^3+rnorm(N)
plot(X,Y)
```

- The learning machine is characterised by the following three components:
  1. A class of hypothesis functions $h(x, \alpha) = \alpha x$ consisting of all the linear models passing through the origin. The class Λ is then the set of real numbers.

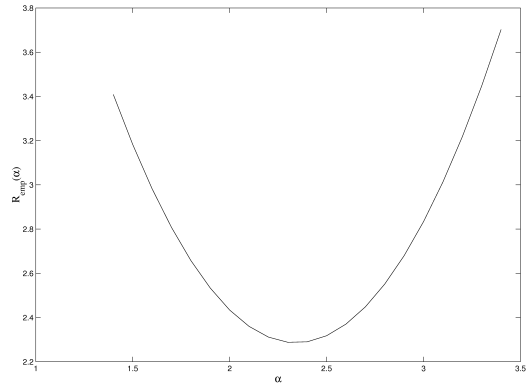


Figure 7.10: The empirical risk for the training set D_N vs. the model parameter value (x-axis). The minimum of the empirical risk is attained in $\alpha = 2.3272$.

2. A quadratic loss $L(y, h(x)) = (y - h(x))^2$.
3. An algorithm of parametric identification based on the least-squares technique, which will be detailed later in Section 9.1.2. The empirical risk is the quantity

$$R_{\text{emp}}(\alpha) = \frac{1}{100} \sum_{i=1}^{100} (y_i - \alpha x_i)^2 \tag{7.5.26}$$

The empirical risk is a function of α and the training set. For the given training set D_N, the empirical risk as a function of α is plotted in Fig. 7.10.

For the dataset D_N in Figure 7.9, it is possible to obtain α_N by minimising the empirical risk (7.5.26)

$$\alpha_N = \arg\min_{\alpha \in \Lambda} R_{\text{emp}}(\alpha) = \arg\min_{\alpha \in \Lambda} \frac{1}{100} \sum_{i=1}^{100} (y_i - \alpha x_i)^2 = 2.3272 \tag{7.5.27}$$

The selected hypothesis is plotted in the input/output domain in Fig. 7.11.

If the joint distribution (e.g. its conditional expectation and variance) were to be known, it would also be possible to compute the risk functional (7.2.3) as

$$R(\alpha) = \frac{1}{4} \int_{-2}^{2} (x^3 - \alpha x)^2 dx + 1 = 4\frac{\alpha^2}{3} - \frac{32}{5}\alpha + 71/7 \tag{7.5.28}$$

where the derivation of the equality is sketched in Appendix C.13. For the given joint distribution, the quantity $R(\alpha)$ is plotted as a function of α in Fig. 7.12. The function takes a global minimum in $\alpha_0 = 2.4$ as can be derived from the analytical expression in (7.5.28).

The computation of the quantity (7.5.22) requires however an average over all the possible realisations of the random variable $\boldsymbol{\alpha}_N$ for datasets of $N = 100$ points. Figure 7.13 shows 6 different realisations of the training set for the same conditional distribution (7.5.25) and the corresponding 6 values of α_N. Note that those six values may be considered as 6 different realisations of the sampling distribution (Section 5.4) of $\boldsymbol{\alpha}_N$.

It is important to remark that both the quantities (7.2.3) and (7.5.22) may be computed only if we know a priori the data joint distribution. Unfortunately, in real cases this knowledge is not accessible and the goal of learning theory is to study the problem of estimating these quantities from a finite set of data.

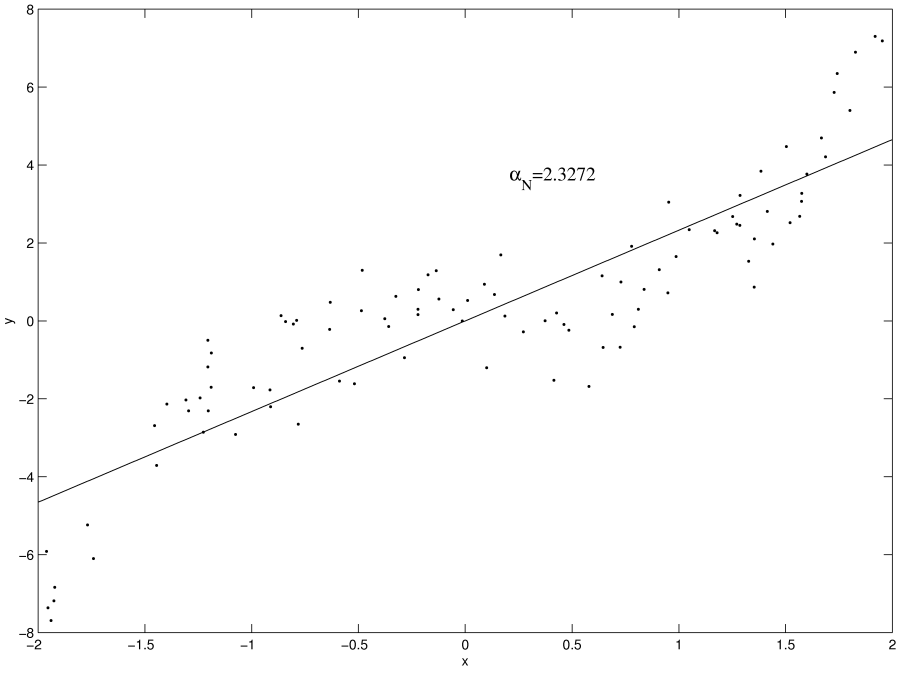


Figure 7.11: Training set (dotted points) and the linear hypothesis function $h(\cdot, \alpha_N)$ (straight line). The quantity α_N, which represents the slope of the straight line, is the value of the model parameter α which minimizes the empirical risk.

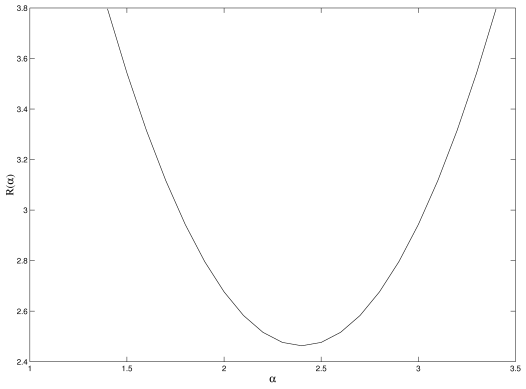


Figure 7.12: The functional risk (7.5.28) vs. the value of model parameter α (x-axis). The minimum of the functional risk is attained in $\alpha_0 = 2.4$.

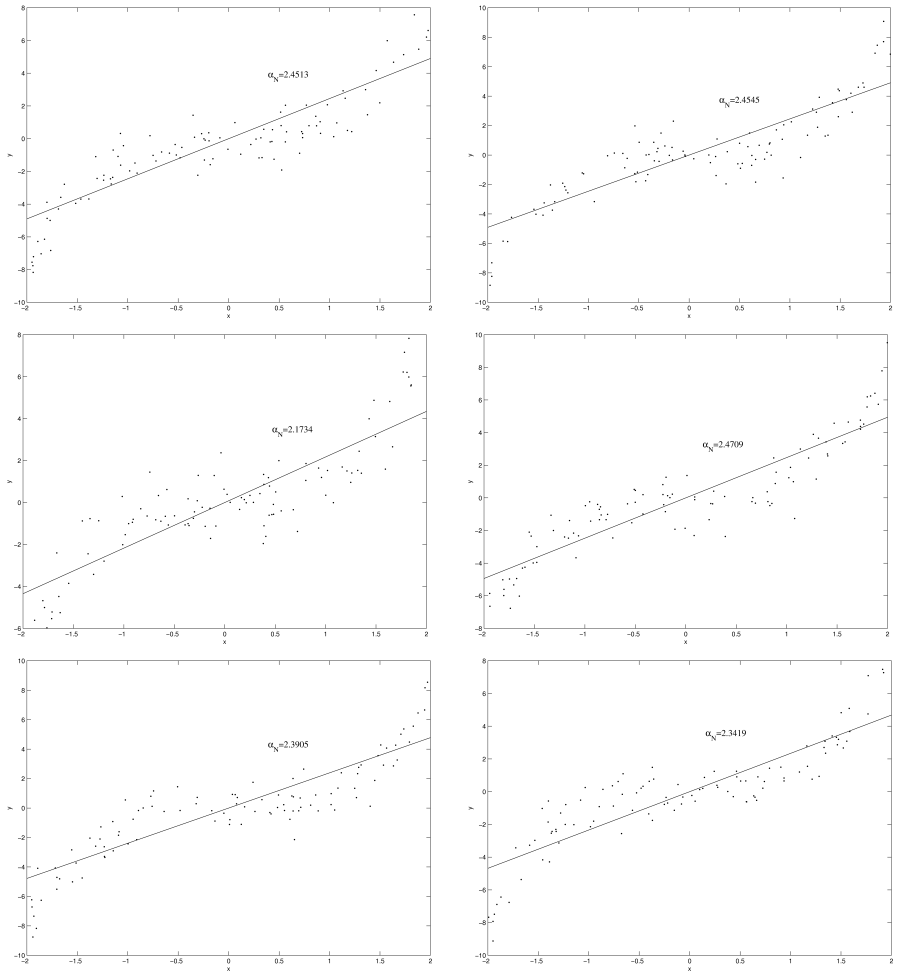


Figure 7.13: Six different realisations of a training set with $N = 100$ points (dots) and the relative hypotheses (solid straight lines) chosen according to the ERM principle (7.5.27).

Monte Carlo computation of generalisation error

The script `functRisk.R` computes by Monte Carlo the functional risk (7.5.28) for different values of α and returns the value of $\alpha_0 = 2.4$ which minimises it. Note that the functional risk is computed by generating a very large number of i.i.d. test examples.

The script `gener.R` computes by Monte Carlo the generalisation error (7.5.23). Unlike the previous script which considers only the predictive value of different hypothesis (with different α), this script assesses the average accuracy of the empirical risk minimisation strategy (7.5.27) for a finite number $N = 100$ of examples.

•

7.6 Functional and empirical risk

This section reports some results from the pioneering work of Prof. Vladimir Vapnik [184, 182, 183] on statistical learning. He defines the learning problem as the problem of finding the hypothesis which minimises the functional risk (7.2.3) on the basis of a finite set of observed data and without any specific assumption about the data distribution. For details and mathematical derivations, we refer the reader to his books [182, 183]. Here we will limit to report some of his most significant results. We start by rewriting the functional risk notation (7.2.3) as

$$R(\alpha) = \int L(y, h(x,\alpha) dF_{\langle \mathbf{x},\mathbf{y} \rangle}(x,y) = \int Q(z,\alpha) dF_{\mathbf{z}}(z) \qquad \alpha \in \Lambda \tag{7.6.29}$$

where $z = \langle x, y \rangle$, $Q(z,\alpha) = L(y, h(x,\alpha))$, the probability measure $F_{\mathbf{z}}(\cdot)$ is unknown but an i.i.d. sample $z_1, \ldots, z_N$ is given. Analogously, the empirical risk may be rewritten as

$$R_{\text{emp}}(\alpha_N) = \frac{1}{N} \sum_{i=1}^{N} Q(z_i, \alpha_N)$$

Let us define with Λ^* the set of *all* possible single valued mappings $f : \mathcal{X} \to \mathcal{Y}$ and consider the quantity

$$\alpha^* = \arg \min_{\alpha \in \Lambda^*} R(\alpha)$$

where $R(\alpha^*)$ is the absolute minimum rate of functional risk (7.2.6).

We can write the equality

$$R(\alpha_N) - R(\alpha^*) = (R(\alpha_N) - R(\alpha_0)) + (R(\alpha_0) - R(\alpha^*)) = \\ = \text{Err}_{\text{estim}}(\alpha_N) + \text{Err}_{\text{approx}}(\alpha_N)$$

where α_0 is the hypothesis with lowest risk in Λ (Equation (7.2.4)).

The first right-hand term is the *estimation error* while the second is the *approximation error* (Figure 7.14). The estimation error represents the discrepancy between the generalisation error of the best hypothesis in the class ($R(\alpha_0)$) and the one learned from D_N ($R(\alpha_N)$). The approximation error is non null when the best hypothesis in the class Λ ($h(\cdot, \alpha_0)$) is different from $h(\cdot, \alpha^*)$.

The trade-off between approximation and estimation error is controlled by the size of Λ: when the size of Λ is large, $R(\alpha_0)$ is close to $R(\alpha^*)$ but the estimation error could be large. On the other way round, if the size of Λ is small, the estimation error is limited but the approximation error could be non negligible.

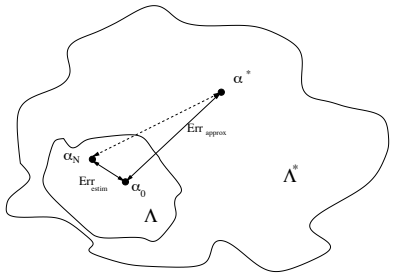


Figure 7.14: Decomposition of the functional risk into estimation and approximation error.

7.6.1 Consistency of the ERM principle

Functional and empirical risk are two key quantities in statistical learning (Figure 7.15). The functional risk represents the generalisation accuracy of the hypothesis once tested with new data while $R_{\text{emp}}(\cdot)$ measures the accuracy of the fitting to the training set. A main issue is that $R_{\text{emp}}(\cdot)$ could be a very bad estimator of the functional risk, e.g. when the class of hypothesis is too rich with respect to the size of the observed sample.

According to Vapnik it is important to characterise the relation between those two quantities, i.e. to define the (necessary and sufficient) conditions for the empirical risk $R_{\text{emp}}(\alpha_N)$ to converge for $N \to \infty$ to the best functional risk $R(\alpha_0)$ in the class Λ. This is known as the problem of consistency of the Empirical Risk Minimisation (ERM) principle.

In formal terms, the ERM principle is consistent for the set of functions $Q(z, \alpha)$ and for the probability distribution $P_{\mathbf{z}}(z)$ if the following two sequences converge in probability to the same limit

$$R(\alpha_N) \xrightarrow[N\to\infty]{P} R(\alpha_0)$$
$$R_{\text{emp}}(\alpha_N) \xrightarrow[N\to\infty]{P} R(\alpha_0)$$

The following lemma shows that both convergences may be studied by considering the quantity $\sup_{\alpha\in\Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$.

Lemma 3 (Devroye 1988)**.**

$$R(\alpha_N) - \inf_{\alpha\in\Lambda} R(\alpha) = R(\alpha_N) - R(\alpha_0) \leq 2 \sup_{\alpha\in\Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$$
$$|R_{\text{emp}}(\alpha_N) - R(\alpha_N)| \leq \sup_{\alpha\in\Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$$

Setting an upper bound for $\sup_{\alpha\in\Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$, we obtain an upper bound for three quantities:

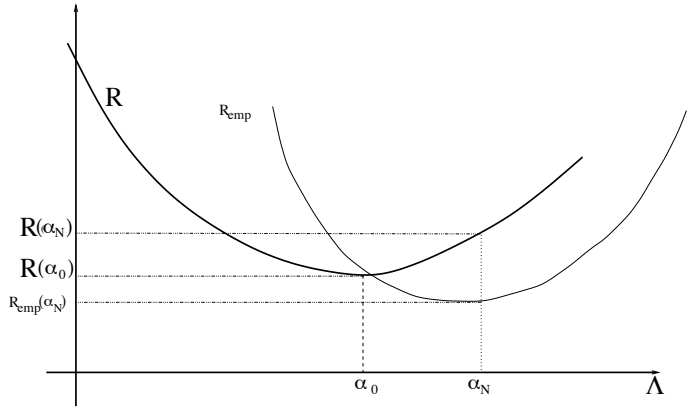


Figure 7.15: Functional and empirical risk.

1. the estimation error $R(\alpha_N) - R(\alpha_0)$ which returns the sub-optimality of the the model chosen by the ERM principle within the class $\alpha \in \Lambda$
2. $|R_{\text{emp}}(\alpha_N) - R(\alpha_N)|$ that is the error committed when the empirical risk is used to estimate the functional risk of the selected model
3. $|R_{\text{emp}}(\alpha_N) - R(\alpha_0)|$ that is the error made when the empirical risk is used to estimate the functional risk of the best model in the class Λ.

It can be shown that bounding $\sup_{\alpha\in\Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$ is not only a sufficient but also a necessary condition for consistency of the ERM principle.

7.6.2 Key theorem of learning

Theorem 6.1 (Vapnik,Chervonenkis, 1991)**.** *Let $Q(z,\alpha)$ $\alpha \in \Lambda$ be a set of functions that satisfy the condition*

$$a \leq \int Q(z,\alpha) dP(z) \leq b$$

Condition necessary and sufficient for the ERM principle to be consistent is that the empirical risk $R_{emp}(\alpha)$ converges uniformly to the actual risk $R(\alpha)$ over the set $Q(z,\alpha)$, $\alpha \in \Lambda$ that is

$$\lim_{N\to\infty} Prob\left\{\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right\} = 0 \qquad \forall\varepsilon > 0$$

This theorem rephrases the problem of ERM consistency as a problem of uniform convergence, which ensures that the empirical risk is a good approximation of the functional risk over all functions of Λ (i.e. including the worst-case).

The uniform convergence is trivial and guaranteed by the Law of Large Numbers if the set of functions $Q(z,\alpha)$ contains a single element: in fact, this is nothing more than the convergence of the average to expectation for increasing N. For a real-valued bounded function $a \leq Q(z,\alpha) \leq b$, by Hoeffding's inequalities (Section 5.6) we have

$$\text{Prob}\left\{\left|\int Q(z,\alpha)dP(z) - \frac{1}{N}\sum_{i=1}^{N} Q(z_i,\alpha)\right| > \varepsilon\right\} < \exp\left\{-\frac{2\varepsilon^2 N}{(b-a)^2}\right\}$$

Then the probability of a deviation between empirical and functional risk converges to zero for $N \to \infty$. It is easy to generalise to the case where $Q(z, \alpha)$ has a finite number K of elements:

$$\text{Prob}\left\{\sup_{1\leq k\leq K}\left|\int Q(z,\alpha)dP(z) - \frac{1}{N}\sum_{i=1}^{N} Q(z_i,\alpha)\right| > \varepsilon\right\} < K\exp\left\{-\frac{2\varepsilon^2 N}{(b-a)^2}\right\} = \exp\left\{\left(\frac{\ln K}{N} - \frac{2\varepsilon^2}{(b-a)^2}\right)N\right\}$$

In order to obtain uniform convergence for any ε, the expression

$$\lim_{N\to\infty}\frac{\ln K}{N} = 0 \tag{7.6.30}$$

has to be satisfied. A problem arises when the set of functions is infinite, like in machine learning where the most common classes of hypothesis are uncountable. In this case we need to generalise the classical law of large numbers to functional spaces. Consider the sequence of random variables

$$\xi_N = \sup_{\alpha\in\Lambda}\left(R(\alpha) - R_{\text{emp}}(\alpha)\right) = \sup_{\alpha\in\Lambda}\left(\int Q(z,\alpha)dF(z) - \frac{1}{N}\sum_{i=1}^{N} Q(z_i,\alpha)\right)$$

where the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, has an infinite number of elements. Unlike the finite case, the sequence ξ_N *does not necessarily converge to zero.* The problem of learning is then strongly related to the problem of defining which properties of the class of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, guarantee the convergence in probability of the sequence ξ_N to zero. In the following section we show some theoretical results from Vapnik about the relation between ERM consistency and the topological properties (notably the *diversity*) of the class of hypothesis.

7.6.2.1 Entropy of a set of functions

In what follows we limit to consider the binary classification setting though similar results can be shown for regression. In this setting the functions $Q(z, \alpha)$, $\alpha \in \Lambda$ are indicator functions since they may take only 0 or 1 values. In order to characterise the diversity of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, on the dataset D_N, let $\mathcal{N}^\Lambda(D_N)$ be the number of possible separations of D_N using the functions $Q(z, \alpha)$, $\alpha \in \Lambda$. Note that $\mathcal{N}^\Lambda(\mathbf{D}_N)$ is a random variable since $\mathbf{D}_N$ is a random variable.

An example of this concept is presented in Figure 7.16[3] where $N = 3$ and the functions $h(\cdot)$ implement linear separators of the 2D ($n = 2$) input space. This class of functions is able to perform all possible (i.e. $2^N = 8$) separations of the dataset. It is also said that the class Λ of functions *shatters* the dataset of size $N = 3$. In other words, a set of N points is said to be shattered by a class of hypothesis Λ if, no matter how we assign a binary label to each point, there exists a hypothesis in Λ that separates them. Note that a set of $N = 4$ points is not shattered by a class of linear separators.

The quantity

$$H^\Lambda(N) = E\ln\mathcal{N}^\Lambda(\mathbf{D}_N)$$

is called the *entropy* of the set of functions on the given data and measures the diversity of the class of hypothesis for a given number of observations.

The following theorem from Vapnik shows that this quantity is related to the consistency of the ERM principle.

[3]Taken from `https://datascience.stackexchange.com/questions/16140/how-to-calculate-vc-dimension/16146`

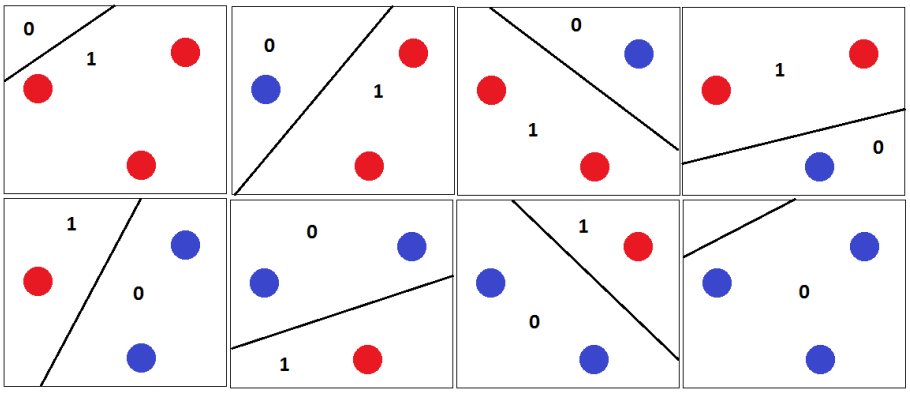


Figure 7.16: Number of linear separations of a dataset of $N = 3$ points.

Theorem 6.2. *A necessary and sufficient condition for the two-sided uniform convergence of the functional risk to the empirical risk is that*

$$\lim_{N\to\infty} \frac{H^{\Lambda}(N)}{N} = 0 \tag{7.6.31}$$

In other words, the ratio of the entropy to the number of observations should decrease to zero with increasing number of observations. Note that this condition depends on the underlying probability distribution $F_{\mathbf{z}}(\cdot)$ and that the entropy plays, for uncountable classes, the role played by the number of functions in the finite case (compare (7.6.30) with (7.6.31)).

7.6.2.2 Distribution independent consistency

Vapnik has been also able to extend the distribution-dependent result of the previous Section to a distribution-free setting.

Theorem 6.3. *Necessary and sufficient condition for consistency of ERM for any probability measure is*

$$\lim_{N\to\infty} \frac{G^{\Lambda}(N)}{N} = 0$$

where

$$G^{\Lambda}(N) = \ln \max_{D_N} \mathcal{N}^{\Lambda}(D_N)$$

is the growth function.

Vapnik proved that in the pattern recognition case

$$\text{Prob}\left\{\sup_{\alpha\in\Lambda}(R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon\right\} \leq 4\exp\left\{\left(\frac{G^{\Lambda}(2N)}{N} - \varepsilon^2\right)N\right\} \tag{7.6.32}$$

This means that, provided that $G^{\Lambda}(N)$ does not grow linearly in N, it is actually possible to bound the (unknown) functional risk $R(\alpha_N)$ on the basis of the (observable) empirical risk $R_{\text{emp}}(\alpha_N)$.

If we set the probability in (7.6.32) to $\delta > 0$ and we solve for ε, then the following inequality holds with probability $1 - \delta$:

$$R(\alpha_N) \leq R_{\text{emp}}(\alpha_N) + \frac{\sqrt{\mathcal{E}}}{2} \tag{7.6.33}$$

where the right-hand side is called the *guaranteed risk* and

$$\mathcal{E} = 4\frac{G^{\Lambda}(2N) - \ln(\delta/4)}{N}.$$

Several other bounds have been derived for different class of hypothesis in [182].

7.6.3 The VC dimension

Vapnik and Chervonenkis showed that either the relation $G^{\Lambda}(N) = N \ln 2$ holds true for all N, or there exists some maximal N for which this relation is satisfied. In this case, this maximal N is called the *VC (Vapnik and Chervonenkis) dimension* and denoted by $\mathcal{D}$. By construction, the VC dimension is the maximal number of points which can be shattered by functions in Λ.

Theorem 6.4. *Any growth function either satisfies the equality*

$$G^{\Lambda}(N) = N \ln 2$$

or is bounded by the inequality

$$G^{\Lambda}(N) \leq \mathcal{D}\left(\ln \frac{N}{\mathcal{D}} + 1\right)$$

where $\mathcal{D}$ is an integer such that when $N = \mathcal{D}$

$$G^{\Lambda}(\mathcal{D}) = \mathcal{D} \ln 2, G^{\Lambda}(\mathcal{D}+1) < (\mathcal{D}+1) \ln 2$$

The VC dimension of a set of indicator functions $Q(z, \alpha)$ is infinite if the growth function is linear. It is finite and equal to $\mathcal{D}$ if the growth function is bounded by a logarithmic function with coefficient $\mathcal{D}$.

The VC dimension quantifies the richness or capacity of a set of functions. If for any N an hypothesis function $h(\cdot, \alpha)$, $\alpha \in \Lambda$ can shatter N points (i.e. separate them in all 2^N possible ways) then $G_{\Lambda}(N) = N \ln 2$. In this case, the class of function has an infinite capacity and there is no ERM convergence (the empirical risk is always zero whatever is the functional risk): no learning from data is possible[4].

The finiteness of the the VC dimension is a necessary and sufficient condition for distribution independent consistency of ERM learning machines. The VC dimension of the set of *linear functions* with $n + 1$ model parameters is equal to $\mathcal{D} = n + 1$. Note that, though for this specific class the VC dimension equals the number of free parameters, this is not necessarily true for other family of functions. For instance, it can be shown that the VC dimension of the highly wiggly set of functions

$$h(x, \alpha) = \sin \alpha x, \qquad \alpha \in \mathbb{R}$$

[4]Note that, in Popper terminology (Section 2.6) this corresponds to a non scientific situation where no dataset may falsify the hypothesis, or equivalently it is always possible to find a hypothesis justifying what we observe. Since the class of hypothesis is too rich, no falsification (and then no generalisation or scientific discovery) is possible.

is infinite though it has a single parameter. At the same time, you can have set of functions with infinite number of parameters yet a finite VC dimension.

Generally speaking, the VC dimension of a set of functions can be either larger than or smaller than the number of parameters. The VC dimension of the set of functions (rather than the number of parameters) is responsible for the generalisation ability of learning machines.

Once defined $\mathcal{D}$, the relation between empirical and functional risk of a class of function with finite VC dimension is now made explicit by the bound (7.6.33) where the second summand is

$$\mathcal{E}(\mathcal{D}, N, \delta) = 4\frac{\mathcal{D}\left(\ln\frac{2N}{\mathcal{D}}+1\right) - \ln(\delta/4)}{N}$$

The reliability of the empirical risk as approximation of the functional risk depends on the ratio $N/\mathcal{D}$. If $N/\mathcal{D}$ is large (i.e. sample size much larger than the VC dimension), the $\mathcal{E}$ term is small and the empirical risk is a good approximation of the functional risk. In other terms, minimising the empirical risk guarantees a small value of the (expected) risk. On the contrary, if $N/\mathcal{D}$ is small (i.e. number of samples comparable to the VC dimension), a small empirical risk $R_{\text{emp}}(\alpha_N)$ does not guarantee a small value of the actual risk. In other terms a small empirical risk could be an optimistic (then biased) estimator of the associated functional risk. In those configurations, to minimise the actual risk $R(\alpha_N)$ it is recommended to address both terms of the confidence interval (e.g. by considering alternative classes of hypothesis).

7.7 Generalisation error

In the previous section we presented how Vapnik [181, 182, 183] formalised the learning task as the minimisation of functional risk $R(\alpha_N)$ in a situation where the joint distribution is unknown. This section focuses on the algorithm-based criterion G_N (Equation (7.5.24)) as a measure of the generalisation error of the learning machine.

In particular we will study how the generalisation error can be decomposed in the regression formulation and in the classification formulation.

7.7.1 The decomposition of the generalisation error in regression

Let us focus now on of the g_N measure (Equation (7.5.22)) of the generalisation error in the case of regression. In the case of a quadratic loss

$$L(y(x), h(x,\alpha)) = (y(x) - h(x,\alpha))^2 \tag{7.7.34}$$

the quantity g_N is often referred to as the *mean squared error* (MSE) and its marginal (7.5.24) as the *mean integrated squared error* (MISE). If the regression dependency is described in the *regression plus noise* form (7.4.21), the conditional target density can be written as

$$p_{\mathbf{y}}(y - f(x)|x) = p_{\mathbf{y}}(y - E_{\mathbf{y}}[\mathbf{y}|x]|x) = p_{\mathbf{w}}(w) \tag{7.7.35}$$

where $\mathbf{w}$ is a noisy random variable with zero mean and variance $\sigma^2_{\mathbf{w}}$.

This supervised learning problem can be seen as a particular instance of the estimation problem discussed in Chapter 5, where, for a given x, the unknown parameter θ to be estimated is the quantity $f(x)$ and the estimator based on the

training set is $\hat{\boldsymbol{\theta}} = h(x, \boldsymbol{\alpha}_N)$. The MSE quantity, defined in (5.5.14) coincides, apart from an additional term, with the term (7.5.22) since

$$
\begin{aligned}
g_N(x) &= E_{\mathbf{D}_N,\mathbf{y}}[\mathbf{L}|x] = && (7.7.36)\\
&= E_{\mathbf{D}_N,\mathbf{y}}\left[(\mathbf{y} - h(x,\boldsymbol{\alpha}_N))^2\right] = && (7.7.37)\\
&= E_{\mathbf{D}_N,\mathbf{y}}\left[(\mathbf{y} - E_{\mathbf{y}}\left[\mathbf{y}|x\right] + E_{\mathbf{y}}\left[\mathbf{y}|x\right] - h(x,\boldsymbol{\alpha}_N))^2\right] = && (7.7.38)\\
&= E_{\mathbf{D}_N,\mathbf{y}}\big[(\mathbf{y} - E_{\mathbf{y}}\left[\mathbf{y}|x\right])^2 + 2\mathbf{w}(E_{\mathbf{y}}\left[\mathbf{y}|x\right] - h(x,\boldsymbol{\alpha}_N)) + && (7.7.39)\\
&+ (E_{\mathbf{y}}\left[\mathbf{y}|x\right] - h(x,\boldsymbol{\alpha}_N))^2\big] = && (7.7.40)\\
&= E_{\mathbf{y}}\left[(\mathbf{y} - E_{\mathbf{y}}\left[\mathbf{y}|x\right])^2\right] + E_{\mathbf{D}_N}\left[(h(x,\boldsymbol{\alpha}_N) - E_{\mathbf{y}}\left[\mathbf{y}|x\right])^2\right] = && (7.7.41)\\
&= E_{\mathbf{y}}\left[\mathbf{w}^2\right] + E_{\mathbf{D}_N}\left[(h(x,\boldsymbol{\alpha}_N) - E_{\mathbf{y}}\left[\mathbf{y}|x\right])^2\right] && (7.7.42)\\
&= \sigma^2_{\mathbf{w}} + E_{\mathbf{D}_N}[(f(x) - h(x,\boldsymbol{\alpha}_N))^2] = \sigma^2_{\mathbf{w}} + E_{\mathbf{D}_N}[(\theta - \hat{\boldsymbol{\theta}})^2] = && (7.7.43)\\
&= \sigma^2_{\mathbf{w}} + \text{MSE} && (7.7.44)
\end{aligned}
$$

Note that $\mathbf{y} = f(x) + \mathbf{w} = E_{\mathbf{y}}\left[\mathbf{y}|x\right] + \mathbf{w}$, f is fixed but unknown and that the noise term $\mathbf{w}$ is independent of $\mathbf{D}_N$ and satisfies $E[\mathbf{w}] = 0$ and $E[\mathbf{w}^2] = \sigma^2_{\mathbf{w}}$

We can then apply *bias/variance* decomposition (5.5.14) to the regression problem where $\theta = f(x)$ and $\hat{\boldsymbol{\theta}} = h(x, \boldsymbol{\alpha}_N)$:

$$
\begin{aligned}
g_N(x) &= E_{\mathbf{D}_N,\mathbf{y}}\left[\mathbf{L}(x,y)\right] = \\
&= \sigma^2_{\mathbf{w}} + E_{\mathbf{D}_N}\left[(h(x,\boldsymbol{\alpha}_N) - E_{\mathbf{y}}\left[\mathbf{y}|x\right])^2\right] = \\
&= \sigma^2_{\mathbf{w}} + && \textit{noise variance}\\
&+ (E_{\mathbf{D}_N}\left[h(x,\boldsymbol{\alpha}_N)\right] - E_{\mathbf{y}}\left[\mathbf{y}|x\right])^2 + && \textit{squared bias}\\
&+ E_{\mathbf{D}_N}\left[(h(x,\boldsymbol{\alpha}_N) - E_{\mathbf{D}_N}\left[h(x,\boldsymbol{\alpha}_N)\right])^2\right] = && \textit{model variance}\\
&= \sigma^2_{\mathbf{w}} + B^2(x) + V(x)
\end{aligned}
\qquad (7.7.45)
$$

In a regression task, the bias $B(x)$ measures the difference in x between the average of the outputs of the hypothesis functions over the set of possible D_N and the regression function value $f(x) = E_{\mathbf{y}}[\mathbf{y}|x]$. The variance $V(x)$ reflects the variability of the guessed $h(x, \boldsymbol{\alpha}_N)$ as one varies over training sets of fixed dimension N. This quantity measures how sensitive the algorithm is to changes in the data set, regardless of the target. So by Eq. (7.5.24) by averaging (7.7.45) over $\mathcal{X}$ we obtain

$$
\text{MISE} = G_N = \sigma^2_{\mathbf{w}} + \int_{\mathcal{X}} B^2(x) dF_{\mathbf{x}} + \int_{\mathcal{X}} V(x) dF_{\mathbf{x}} \qquad (7.7.46)
$$

where the three terms are

1. the intrinsic noise term reflecting the target alone,
2. the integrated squared bias reflecting the target's relation with the learning algorithm and
3. the integrated variance term reflecting the learning algorithm alone.

As the aim of a learning machine is to minimise the quantity G_N and the computation of (7.7.46) requires the knowledge of the joint input/output distribution, this decomposition could appear as a useless theoretical exercise. In practical settings, the designer of a learning machine does not have access to the term G_N but can only estimate it on the basis of the training set. Nevertheless, the bias/variance decomposition is relevant in practical learning too since it provides a useful hint about how to control the error G_N. In particular, the bias term measures the lack of representational power of the class of hypotheses. This means that to reduce the

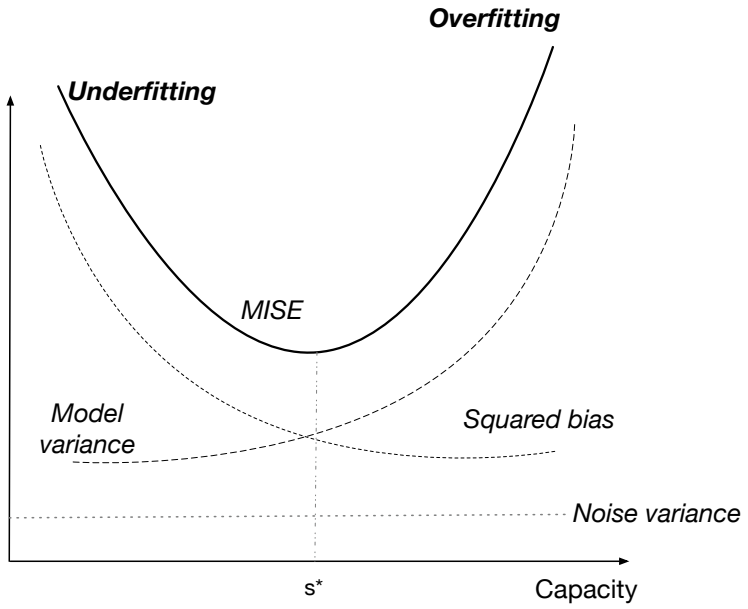


Figure 7.17: Bias/variance/noise tradeoff in regression: this is a qualitative representation of the relationship between the hypothesis' bias and variance and the capacity of the class of functions. The MISE generalisation error is the sum of the three terms (squared bias, hypothesis variance and noise variance) as shown in (7.7.46). Note that the variance of the noise is supposed to be target independent and then constant.

bias term of the generalisation error we should consider classes of hypotheses with a large capacity s, or in other words hypotheses which can approximate a large number of input/output mappings. On the other side, the variance term warns us against an excessive capacity (or complexity) s of the approximator. This means that a class of too powerful hypotheses runs the risk of being excessively sensitive to the noise affecting the training set; therefore, our class Λ_s could contain the target but it could be practically impossible to find it out on the basis of the available dataset.

In other terms, it is commonly said that an hypothesis with large bias but low variance *underfits* the data while an hypothesis with low bias but large variance *overfits* the data. In both cases, the hypothesis gives a poor representation of the target and a reasonable trade-off needs to be found.

A graphical illustration of the bias/variance/noise tradeoff (7.7.46) is made in Figure 7.17. The left side of the figure corresponds to an underfitting configuration where the model has too low capacity (i.e. high bias) to capture the nonlinearity of the regression function. The right side of the figure corresponds to an overfitting configuration where the model capacity is too large (i.e. high variance) leading then to high instability and poor generalisation. Note that Figure 7.17 requires a formal definition of the notion of capacity and that it is only a qualitative visualisation of the theoretical link between the hypothesis' properties and the capacity of the class of functions. Nevertheless it provides useful hints about the impact of the learning procedure on the final generalisation accuracy. The task of the model designer is to search for the optimal trade-off between the variance and the bias terms (ideally the capacity s^* in Figure 7.17), on the basis of the available training set. Section 7.9 will discuss how this search proceeds in practice in a real setting.

Two naive predictors

Consider a regression task $\mathbf{y} = f(\mathbf{x}) + \mathbf{w}$, where $\text{Var}[\mathbf{w}] = \sigma^2_{\mathbf{w}}$ and two naive predictors:

1. $h^{(1)}(x) = 0$
2. $h^{(2)}(x) = \frac{\sum_{i=1}^{N} y_i}{N}$

What about their generalisation errors in $x = \bar{x}$? By using (7.7.45) we obtain

1. $g_N^{(1)}(\bar{x}) = \sigma^2_{\mathbf{w}} + f(\bar{x})^2$
2. $g_N^{(2)}(\bar{x}) = \sigma^2_{\mathbf{w}} + (f(\bar{x}) - E[\mathbf{y}])^2 + \text{Var}[\mathbf{y}]/N$

The script `naive.R` executes a Monte Carlo validation of the formulas above.

•

7.7.2 The decomposition of the generalisation error in classification

Let us consider a classification task with K output classes and a loss function L. For a given input x, we denote by $\hat{\mathbf{y}}$ the class predicted by the classifier $h(x, \boldsymbol{\alpha}_N)$ trained with a dataset D_N. We derive the analytical expression of $g_N(x)$, usually referred to as the *mean misclassification error* (MME).

$$\text{MME}(x) = E_{\mathbf{y},\mathbf{D}_N}[L(\mathbf{y}, h(x, \boldsymbol{\alpha}_N))|x] = E_{\mathbf{y},\mathbf{D}_N}[L(\mathbf{y}, \hat{\mathbf{y}})] = \tag{7.7.47}$$

$$= E_{\mathbf{y},\mathbf{D}_N}[\sum_{k,j=1}^{K} L_{(j,k)} 1(\hat{\mathbf{y}} = c_j|x) 1(\mathbf{y} = c_k|x)] = \tag{7.7.48}$$

$$= \sum_{k,j=1}^{K} L_{(j,k)} E_{\mathbf{D}_N}[1(\hat{\mathbf{y}} = c_j|x)] E_{\mathbf{y}}[1(\mathbf{y} = c_k|x)]] = \tag{7.7.49}$$

$$= \sum_{k,j=1}^{K} L_{(j,k)} \text{Prob}\{\hat{\mathbf{y}} = c_j|x\} \text{Prob}\{\mathbf{y} = c_k|x\} \tag{7.7.50}$$

where $1(\cdot)$ is the indicator function which returns zero when the argument is false and one otherwise. Note that the distribution of $\hat{\mathbf{y}}$ depends on the training set $\mathbf{D}_N$ while the distribution of $\mathbf{y}$ is the distribution of a test set (independent of $\mathbf{D}_N$). For zero-one loss function, since $\mathbf{y}$ and $\hat{\mathbf{y}}$ are independent, the MME expression simplifies to

$$\begin{aligned}
\text{MME}(x) &= \sum_{k,j=1}^{K} 1(c_j \neq c_k) \text{Prob}\{\hat{\mathbf{y}} = c_j|x\} \text{Prob}\{\mathbf{y} = c_k|x\} = \\
&= 1 - \sum_{k,j=1}^{K} 1(c_j = c_k) \text{Prob}\{\hat{\mathbf{y}} = c_j|x\} \text{Prob}\{\mathbf{y} = c_k|x\} = \\
&= 1 - \sum_{k}^{K} \text{Prob}\{\hat{\mathbf{y}} = c_k|x\} \text{Prob}\{\mathbf{y} = c_k|x\} = \text{Prob}\{\mathbf{y} \neq \hat{\mathbf{y}}\}
\end{aligned} \tag{7.7.51}$$

A decomposition of a related quantity was proposed in [194]. Let us consider the squared sum:

$$\frac{1}{2}\sum_{j=1}^{K}\left(\text{Prob}\left\{\mathbf{y}=c_j\right\}-\text{Prob}\left\{\hat{\mathbf{y}}=c_j\right\}\right)^2=$$

$$\frac{1}{2}\left(\sum_{j=1}^{K}\text{Prob}\left\{\mathbf{y}=c_j\right\}^2\right)+\frac{1}{2}\left(\sum_{j=1}^{K}\text{Prob}\left\{\hat{\mathbf{y}}=c_j\right\}^2\right)-\sum_{j=1}^{K}\text{Prob}\left\{\mathbf{y}=c_j\right\}\text{Prob}\left\{\hat{\mathbf{y}}=c_j\right\}$$

By adding one to both members and by using (7.7.47) we obtain a decomposition analogous to the one in (7.7.45)

$$\begin{aligned} g_N(x)=\text{MME}(x)= \\ &=\frac{1}{2}\left(1-\left(\sum_{j=1}^{K}\text{Prob}\left\{\mathbf{y}=c_j|x\right\}^2\right)\right)+ \quad \textit{“noise”} \\ &+\frac{1}{2}\sum_{j=1}^{K}\left(\text{Prob}\left\{\mathbf{y}=c_j|x\right\}-\text{Prob}\left\{\hat{\mathbf{y}}=c_j|x\right\}\right)^2+ \quad \textit{“squared bias”} \\ &+\frac{1}{2}\left(1-\left(\sum_{j=1}^{K}\text{Prob}\left\{\hat{\mathbf{y}}=c_j|x\right\}^2\right)\right) \quad \textit{“variance”} \end{aligned} \tag{7.7.52}$$

The noise term measures the degree of uncertainty of $\mathbf{y}$ and consequently the degree of stochasticity of the dependance. It equals zero if and only if there exists a class c such that $\text{Prob}\left\{\mathbf{y}=c|x\right\}=1$ and zero otherwise. Note that this quantity does not depend on the learning algorithm nor on the training set.

The variance term measures how variant the classifier prediction $\hat{\mathbf{y}}=h(x,\boldsymbol{\alpha}_N)$ is. This quantity is zero if the predicted class is always the same regardless of the training set.

The squared bias term measures the squared difference between the $\mathbf{y}$ and the $\hat{\mathbf{y}}$ probability distributions on the domain $\mathcal{Y}$.

7.8 The hypothesis-based vs the algorithm-based approach

In the previous sections we introduced two different manners of assessing the accuracy of a learning machine. The reader could logically raise the following question: which approach is the most adequate in practice?

Instead of providing a direct answer to such question, we prefer to conduct a short comparison of the assumptions and limitations related to the two approaches.

The hypothesis-based approach formulates learning as the problem of finding the hypothesis which minimises the functional risk. Vapnik reformulates this problem into the problem of consistency of a learning process based on ERM. The main result is that it is possible to define a probabilistic distribution-free bound on the functional risk which depends on the empirical risk and the VC dimension of the class of hypothesis. Though this achievement is impressive from a theoretical and scientific perspective (it was published in a Russian book in the 60s), its adoption in practical settings is not always easy for several reasons: results derive from asymptotic considerations though learning by definition deals with finite samples, the computation of the VC dimension is explicit only for specific classes of hypothesis functions and the bound, derived from worst-case analysis, is not always tight enough for practical purposes.

The algorithm-based approach relies on the possibility of emulating the stochastic process underlying the dataset by means of resampling procedures like cross-validation or bootstrap. Note that this approach is explicitly criticised by Vapnik and others who consider it inappropriate to reason in terms of data generation once a single dataset is available. According to [57] "averaging over the data would be unnatural, because in a given application, one has to live with the data at hand. It would be marginally useful to known the number G_N as this number would indicate the quality of an average data sequence, not *your* data sequence". Nevertheless, though it is hard to guarantee formally the accuracy of a resampling strategy, its general-purpose nature, simplicity and ease of implementation have been, along years, key ingredients of its success.

Whatever the degree of realism of the hypothesis made by the two approaches is, it is worth making a pragmatic and historical consideration. Though the Vapnik results represent a major scientific success and underlie the design of powerful learning machines (notably SVM), in a wider perspective it is fair to say that cross-validation is the most common and successful workhorse of practical learning applications. This means that, though most data scientists have been eager to formalise the consistency of their algorithms in terms of Vapnik bounds, in practice they had recourse to intensive cross-validation tricks to make it work in the real world. Now, more than 60 years after the first computational version of learning processes, we have enough evidence to say that cross-validation is a major element of the machine learning success story. This is the reason why in the following sections we will focus on an algorithm-based approach aiming to assess (and minimise) the generalisation error by means of a resampling strategy.

7.9 The supervised learning procedure

The goal of supervised learning is to return the hypothesis with the lowest generalisation error. Since we assume that data samples are generated in a random way, there is no hypothesis which gives a null generalisation error. Therefore, the generalisation error G_N of the hypothesis returned by a learning machine has to be compared to the minimal generalisation error that can be attained by the best single-valued mapping. Let us define by Λ^* the set of all possible single valued mappings $h : \mathcal{X} \to \mathcal{Y}$ and consider the hypothesis

$$\alpha^* = \arg \min_{\alpha \in \Lambda^*} R(\alpha) \tag{7.9.53}$$

where $R(\alpha)$ has been defined in (7.2.3).

Thus, $R(\alpha^*)$ represents the absolute minimum rate of error obtainable by a single valued approximator of the unknown target. For maintaining a simple notation, we put $G^* = R(\alpha^*)$. For instance, in our illustrative example in Section 7.5.1, α^* denotes the parameters of the cubic function and G^* amounts to the unit variance of the Gaussian noise.

In theoretical terms, a relevant issue is to demonstrate that the generalisation error G_N of the model with parameters α_N learned from the dataset D_N converges to the minimum G^* for N going to infinity. Unfortunately, in real learning settings, two problems must be dealt with. The first is that the error G_N cannot be computed directly but has to be estimated from data. The second is that a single class Λ could not be large enough to contain the hypothesis α^*.

A common practice to handle these problems is to decompose the learning procedure in the following sequence of steps:

1. A nested sequence of classes of hypotheses

$$\Lambda_1 \subseteq \cdots \subseteq \Lambda_s \subseteq \ldots \Lambda_S \tag{7.9.54}$$

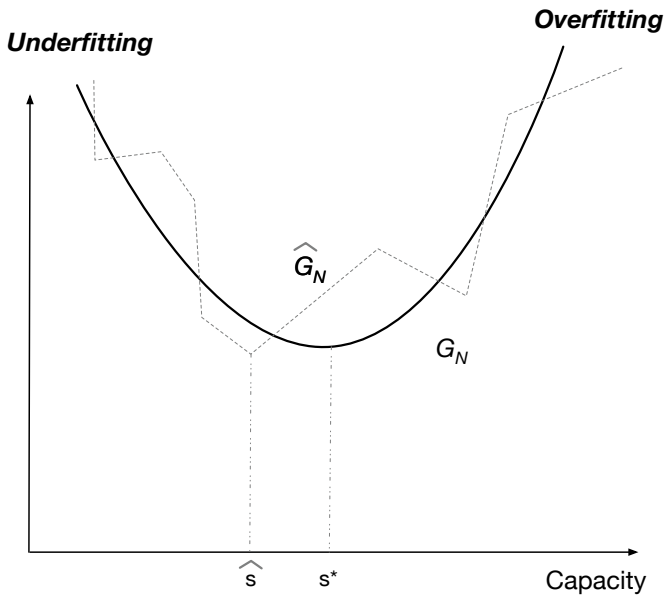


Figure 7.18: Bias/variance/noise tradeoff and model selection: since the generalisation error (e.g. MISE) is not accessible in practical settings, model selection is performed on the basis of an estimation (dotted line) which may induce an error (and a variability) in the selection (7.9.55) of the best capacity.

is defined so that $\Lambda^* = \cup_{s=1}^{S} \Lambda_s$ where s denotes the capacity of the class. This guarantees that the set of hypotheses taken into consideration will necessarily contain the best hypothesis α^*.

A priori informations as well as considerations related to the bias/variance dilemma can help in the design of this sequence.

2. For each class in the sequence, a hypothesis $h(\cdot, \alpha_N^s)$, $s = 1, \ldots, S$, is selected by minimising the empirical risk (7.2.8). This step is defined as the *parametric identification* step of the learning procedure.

3. For each class in the sequence, a validation procedure returns $\hat{G}_N^s$ which estimates the generalisation error G_N^s of the hypothesis α_N^s. This step is called the *validation* step of the learning procedure.

4. The hypothesis $h(\cdot, \alpha_N^{\bar{s}}) \in \Lambda_{\bar{s}}$ with

$$\bar{s} = \arg\min_s \hat{G}_N^s \tag{7.9.55}$$

is returned as the final outcome. This final step is called the *model selection* step.

In order to accomplish the learning procedure, and specifically the selection in (7.9.55), we need an estimation of the generalisation error (Section 7.10). However, since the estimator of the generalisation error may be affected by an error (as any estimator), this may induce an error and a variability in the model selection step (7.9.55) (Figure 7.18).

7.10 Validation techniques

This section discusses validation methods to estimate the generalisation error G_N from a finite set of N observations.

The empirical risk (also called *apparent error*) $R_{\text{emp}}(\alpha_N)$ introduced in (7.2.7) could be the most intuitive estimator of G_N. However, it is generally known that the empirical risk is a biased (and optimistic) estimate of G_N and that $R_{\text{emp}}(\alpha_N)$ tends to be smaller than G_N, because the same data have been used both to construct and to evaluate $h(\cdot, \alpha_N)$. A demonstration of the biasedness of the empirical risk for a quadratic loss function in a regression setting is available in Appendix C.14. In Section 9.1.16 we will analytically derive the biasedness of the empirical risk in case of linear regression models.

The study of error estimates other than the apparent error is of significant importance if we wish to obtain results applicable to practical learning scenarios. There are two main ways to obtain better, i.e. unbiased, estimates of G_N: the first requires some knowledge on the distribution underlying the data set, the second makes no assumptions on the data. As we will see later, an example of the first approach is the FPE criterion (presented in Section 9.1.16.2) while examples of the second approach are the resampling procedures.

7.10.1 The resampling methods

Cross-validation [172] is a well-known method in sampling statistics to circumvent the limits of the apparent error estimate. The basic idea of cross-validation is that one builds a model from one part of the data and then uses that model to predict the rest of the data. The dataset D_N is split l times in a training and a test subset, the first containing N_{tr} examples, the second containing $N_{ts} = N - N_{tr}$ examples. Each time, N_{tr} examples are used by the parametric identification algorithm $\mathcal{L}$ to select a hypothesis $\alpha^i_{N_{tr}}$, $i = 1, \ldots, l$, from Λ and the remaining N_{ts} examples are used to estimate the error of $h(\cdot, \alpha^i_{N_{tr}})$ (Fig. 7.19)

$$\hat{R}_{ts}(\alpha^i_{N_{tr}}) = \sum_{j=1}^{N_{ts}} L(y_j, h(x_j, \alpha^i_{N_{tr}})) \tag{7.10.56}$$

The resulting average of the l errors $\hat{R}_{ts}(\alpha^i_{N_{tr}})$, $i = 1, \ldots, l$, is the cross-validation estimate

$$\hat{G}_{\text{cv}} = \frac{1}{l} \sum_{i=1}^{l} \hat{R}_{ts}(\alpha^i_{N_{tr}}) \tag{7.10.57}$$

A common form of cross-validation is the "leave-one-out" (l-o-o). Let $D_{(i)}$ be the training set with z_i removed, and $h(x, \alpha_{N(i)})$ be the corresponding prediction rule. The l-o-o cross-validated error estimate is

$$\hat{G}_{\text{loo}} = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, h(x_i, \alpha_{N(i)})\right) \tag{7.10.58}$$

In this case l equals the number of training points and $N_{ts} = 1$.

Bootstrap (Section 6.4) is also used to return a nonparametric estimate of G_N, by repeatedly sampling the training cases *with replacement*. Since empirical risk is a biased optimistic estimation of generalisation error and bootstrap is an effective method to assess bias (Section 6.4.3), it follows that bootstrap plays a role in a validation strategy.

A bootstrap sample $D_{(b)}$ is a "fake" dataset $\{z_{1b}, z_{2b}, \ldots, z_{Nb}\}$, $b = 1, \ldots, B$ randomly selected from the training set $\{z_1, z_2, \ldots, z_N\}$ with replacement.

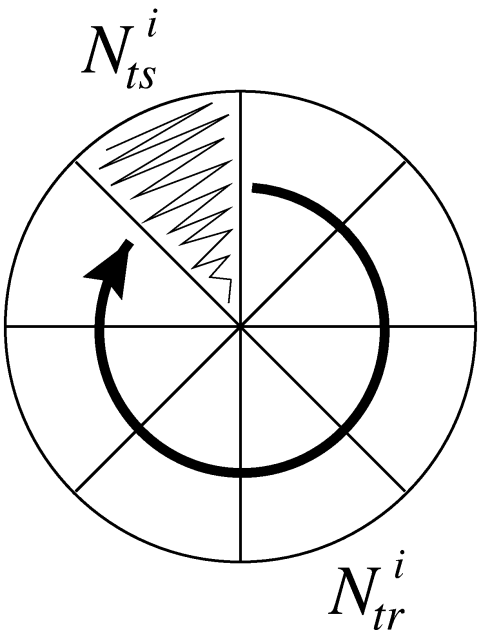


Figure 7.19: Partition of the training dataset in the i^{th} fold of cross-validation. The quantity N_{tr}^{i} is the number of training points while N_{ts}^{i} is the number of test points.

Efron and Tibshirani [64] proposed to use bootstrap to correct the bias (or optimism) of empirical risk by adopting a strategy similar to Section 6.4.3. Equation (6.4.9) estimates the bias of an estimator by computing the gap between the average bootstrap (6.4.8) estimate and the sample estimation. In the case of generalisation, the sample estimation $\hat{\theta}$ is the empirical risk and the bootstrap estimate $\hat{\theta}_{(\cdot)}$ may be computed as follows

$$\hat{G}_{(\cdot)} = \frac{1}{B}\left[\sum_{i=1}^{N}(P_{ib}L\left(y_i, h(x_i, \alpha_{(b)})\right)\right] \tag{7.10.59}$$

where P_{ib} indicates the proportion of the bootstrap sample $D_{(b)}$, $b = 1, \ldots, B$ containing the ith training point z_i,

$$P_{ib} = \frac{\#_{j=1}^{N}(z_{jb} = z_i)}{N} \tag{7.10.60}$$

and $\alpha_{(b)}$ is the output of the parametric identification performed on the set $D_{(b)}$.

The difference between empirical risk and (7.10.59)

$$\text{Bias}_{\text{bs}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N}\left((P_{ib} - \frac{1}{N})L\left(y_i, h(x_i, \alpha_{(b)})\right)\right) \tag{7.10.61}$$

is the bias correction term to be subtracted to empirical risk to obtain a *bootstrap bias corrected* estimate (6.4.10) of the generalisation error.

An alternative consists in using the holdout principle in combination with the bootstrap one [64]. Since each bootstrap set is a resampling of the original training set, it may happen that some of the original examples (called *out-of-bag*) do not belong to it: we can then use them to have an independent holdout set to be used

for generalisation assessment. The bootstrap estimation of the generalisation error (also known as E0) is then

$$\hat{G}_{\text{bs}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|B^{(i)}|}\sum_{b\in B^{(i)}} L\left(y_i, h(x_i, \alpha_{(b)})\right) \tag{7.10.62}$$

where $B^{(i)}$ is the set of bootstrap samples which do not contain the ith point and $|B^{(i)}|$ is its size. The terms where $|B^{(i)}| = 0$ are discarded.

7.11 Concluding remarks

The goal of a learning procedure is to return a hypothesis which is able to predict accurately the outcome of an input/output probabilistic mapping on the basis of past observations. In order to achieve this goal, the learning procedure has to deal with three major difficulties.

Minimisation of the empirical risk: in a general case finding the global minimum of the empirical risk as in (7.2.7) demands the resolution of a multivariate and nonlinear optimisation problem for which no analytical solution could exist. Some heuristics to address this issue are discussed in Section 8.6.

Finite number of data: in real problems, a single random realisation of the statistical process, made of a finite number of input/output pairs, is accessible to the learning machine. This means, that the hypothesis generated by a learning algorithm is a random variable as well. In theory, it would be required to have access to the underlying process and to generate several times the training set, in order to have a reliable assessment of the learning algorithm. In practice, the use of repeated realisations is not viable in a real learning problem.

The validation procedure copes with this problem by trying to assess a random variable on the basis of a single realisation. In particular we focused on cross-validation, a resampling method which works by simulating the stochastic process underlying the data.

No a priori knowledge: we consider a setting where no knowledge about the process underlying the data is available. This lack of a priori knowledge puts no constraints on the complexity of the class of hypotheses to consider, with the consequent risk of using an inadequate type of approximator. The model selection deals with this issue by considering classes of hypotheses of increasing complexity and selecting the one which behaves the best according to the validation criteria. This strategy ensures the covering of the whole spectrum of approximators, ranging from low bias/high variance to high bias/low variance models, making easier the selection of a good trade-off on the basis of the available data.

So far, the learning problem has been introduced and discussed for a generic class of hypotheses, and we did not distinguish on purpose between different learning machines. The following chapter will show the parametric and the structural identification procedure as well as the validation phase for some specific learning approaches.

7.12 Exercises

1. Consider an input/output regression task where $n = 1$, $E[\mathbf{y}|x] = \sin(\pi x/2)$ and $p(y|x) = \mathcal{N}(\sin(\pi x/2), \sigma^2)$, $\sigma = 0.1$ and $\mathbf{x} \sim \mathcal{U}(-2, 2)$. Let N be the size of the

training set and consider a quadratic loss function.

Let the class of hypothesis be $h_M(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m x^m$ with $\alpha_j \in [-2, 2], j = 0, \ldots, M$.

For $N = 20$ generate $S = 50$ replicates of the training set. For each replicate, estimate the value of the parameters that minimise the empirical risk, compute the empirical risk and the functional risk.

1. Plot the evolution of the distribution of the empirical risk for $M = 0, 1, 2$.
2. Plot the evolution of the distribution of the functional risk for $M = 0, 1, 2$.

Hints: to minimise the empirical risk, perform a grid search in the space of parameter values, i.e. by sweeping all the possible values of the parameters in the set $[-1, -0.9, -0.8, \ldots, 0.8, 0.9, 1]$. To compute the functional risk generate a set of $N_{ts} = 10000$ i.i.d. input/output testing examples.

Solution: See the file `Exercise6.pdf` in the directory `gbcode/exercises` of the companion R package (Appendix F).