Preface

How does one decide among competing explanations of data given limited observations? This is the problem of *model selection*. A central concern in model selection is the danger of *overfitting*: the selection of an overly complex model that, while fitting observed data very well, predicts future data very badly. Overfitting is one of the most important issues in inductive and statistical inference: besides model selection, it also pervades applications such as prediction, pattern classification and parameter estimation.

The minimum description length (MDL) principle is a relatively recent method for inductive inference that provides a generic solution to the model selection problem, and, more generally, to the overfitting problem. MDL is based on the following insight: any regularity in the data can be used to *compress* the data, i.e. to describe it using fewer symbols than the number of symbols needed to describe the data literally. The more regularities there are, the more the data can be compressed. Equating "learning" with "finding regularity," we can therefore say that the more we are able to compress the data, the more we have *learned* about the data. Formalizing this idea leads to a general theory of inductive inference with several attractive properties:

1. **Occam's razor.** MDL chooses a model that trades off goodness-of-fit on the observed data with "complexity" or "richness" of the model. As such, MDL embodies a form of Occam's razor, a principle that is both intuitively appealing and informally applied throughout all the sciences.

2. **No overfitting, *automatically*.** MDL methods *automatically* and *inherently* protect against overfitting and can be used to estimate both the parameters and the structure (e.g., number of parameters) of a model. In contrast, to avoid overfitting when estimating the structure of a model, traditional

methods such as maximum likelihood must be *modified* and extended with additional, typically ad hoc principles.

3. **Bayesian interpretation.** Some (not all) MDL procedures are closely related to Bayesian inference. Yet they avoid some of the interpretation difficulties of the Bayesian approach, especially in the realistic case when it is known a priori to the modeler that none of the models under consideration is true. In fact:

4. **No need for "underlying truth."** In contrast to other statistical methods, MDL procedures have a clear interpretation independent of whether or not there exists some underlying "true" model.

5. **Predictive interpretation.** Because data compression is formally equivalent to a form of probabilistic prediction, MDL methods can be interpreted as searching for a model with good predictive performance on *unseen* data. This makes MDL related to, yet different from, data-oriented model selection techniques such as cross-validation.

This Book

This book provides an extensive, step-by-step introduction to the MDL principle, with an emphasis on conceptual issues. From the many talks that I have given on the subject, I have noticed that the same questions about MDL pop up over and over again. Often, the corresponding answers can be found only — if at all — in highly technical journal articles. The main aim of this book is to serve as a reference guide, in which such answers can be found in a much more accessible form. There seems to be a real need for such an exposition because, quoting Lanterman (2001), of "the challenging nature of the original works and the preponderance of misinterpretations and misunderstandings in the applied literature." Correcting such misunderstandings is the second main aim of this book.

First Aim: Accessibility I first learned about MDL in 1993, just before finishing my master's in computer science. As such, I knew some basic probability theory and linear algebra, but I knew next to nothing about advanced measure-theoretic probability, statistics, and information theory. To my surprise, I found that to access the MDL literature, I needed substantial knowledge about all three subjects! This experience has had a profound influence on this book: in a way, I wanted to write a book which I would have been

able to understand when I was a beginning graduate student. Therefore, since with some difficulty its use can be avoided, there is no measure theory whatsoever in this book. On the other hand, this book is full of statistics and information theory, since these are essential to any understanding of MDL. Still, both subjects are introduced at a very basic level in Part I of the book, which provides an initial introduction to MDL. At least this part of the book should be readable without any prior exposure to statistics or information theory.

If my main aim has succeeded, then this book should be accessible to (a) researchers from the diverse areas dealing with inductive inference, such as statistics, pattern classification, and branches of computer science such as machine learning and data mining; (b) researchers from biology, econometrics, experimental psychology, and other applied sciences that frequently have to deal with inductive inference, especially model selection; and (c) philosophers interested in the foundations of inductive inference. This book should enable such readers to understand what MDL is, how it can be used, and what it does.

Second Aim: A Coherent, Detailed Overview In the year 2000, when I first thought about writing this book, the field had just witnessed a number of advances and breakthroughs, involving the so-called *normalized maximum likelihood code*. These advances had not received much attention outside of a very small research community; most practical applications and assessments of MDL were based on "old" (early 1980s) methods and ideas. At the time, some pervasive myths were that "MDL is just two-part coding", "MDL is BIC" (an asymptotic Bayesian method for model selection), or "MDL is just Bayes." This prompted me and several other researchers to write papers and give talks about the new ideas, related to the normalized maximum likelihood. Unfortunately, this may have had somewhat of an adverse effect: I now frequently talk to people who think that MDL is just "normalized maximum likelihood coding." This is just as much of a myth as the earlier ones! In reality, MDL in its modern form is based on a general notion known in the information-theoretic literature as *universal coding*. There exist many types of universal codes, the main four types being the Bayesian, two-part, normalized maximum likelihood, and prequential plug-in codes. All of these can be used in MDL inference, and which one to use depends on the application at hand. While this emphasis on universal codes is already present in the overview (Barron, Rissanen, and Yu 1998), their paper requires substan-

tial knowledge of information theory and statistics. With this book, I hope to make the universal coding-based MDL theory accessible to a much wider audience.

A Guide for the Reader

This book consists of four parts. Part I is really almost a separate book. It provides a very basic introduction to MDL, as well as an introductory overview of the statistical and information-theoretic concepts needed to understand MDL. Part II is entirely devoted to universal coding, the information-theoretic notion on which MDL is built. Universal coding is really a theory about data compression. It is easiest to introduce without directly connecting it to inductive inference, and this is the way we treat it in Part II. In fact though, there is a very strong relation between universal coding and inductive inference. This connection is formalized in Part III, where we give a detailed treatment of MDL theory as a theory of inductive inference based on universal coding. Part IV can once again be read separately, providing an overview of the statistical theory of *exponential families*. It provides background knowledge needed in the proofs of theorems in Part II.

The Fast Track — How to Avoid Reading Most of This Book I do not suppose that any reader will find the time to read all four parts in detail. Indeed, for readers with prior exposure to MDL, this book may serve more like a reference guide than an introduction in itself. For the benefit of readers with no such prior knowledge, each chapter in Part I and Part II starts with a brief list of its contents as well as a *fast track*–paragraph, which indicates the parts that should definitely be read, and the parts that can be skipped at first reading. This allows a "fast track" through Part I and Part II, so that the reader can quickly reach Part III, which treats state-of-the-art MDL inference. Additionally, some sections are marked with an asterisk (*). Such sections contain advanced material and may certainly be skipped at first reading.

> Also, the reader will frequently find paragraphs such as the present one, which are set in smaller font. These provide additional, more detailed discussion of the issues arising in the main text, and may also be skipped at first reading.

Also, at several places, the reader will find boxes like the one below:

Boxes Contain the Most Important Ideas
Each chapter contains several boxes like this one. These contain the most important insights. Together, they form a summary of the chapter.

To further benefit the hurried reader, we now give a brief overview of each part:

Part I Chapter 1 discusses some of the basic ideas underlying MDL in a mostly nonmathematical manner. Chapter 2 briefly reviews general mathematical and probabilistic preliminaries. Chapter 3 gives a detailed discussion of some essential information-theoretic ideas. Chapter 4 applies these notions to statistical models. This chapter gives an extensive analysis of the log-likelihood function and its expectation. It may be of interest for teachers of introductory statistics, since the treatment emphasizes some, in my view, quite important aspects usually not considered in statistics textbooks. For example, we consider in detail what happens if we vary the data, rather than the parameters. Chapter 5 then gives a first mathematically precise implementation of MDL. This is the so-called crude two-part code MDL. I call it "crude" because it is suboptimal, and not explicitly based on universal coding. I included it because it is easy to explain — especially the fact that it has obvious defects raises some serious questions, and thinking about these questions seems the perfect introduction to the "refined" MDL that we introduce in Part III of the book.

> Although some basic familiarity with elementary probability theory is assumed throughout the text, all probabilistic concepts needed are briefly reviewed in Chapter 2. They are typically taught in undergraduate courses and can be found in books such as (Ross 1998). Strictly speaking, the text can be read without any prior knowledge of statistics or information theory — all concepts and ideas are introduced in Chapters 3 and 4. Nevertheless, some prior exposure to these subjects is probably needed to fully appreciate the developments in Part II and Part III. More extensive introductions to the statistical concepts needed can be found in, for example (Bain and Engelhardt 1989; Casella and Berger ; Rice 1995).

Part II Part II then treats the general theory of universal coding, with an emphasis on issues that are relevant to MDL. It starts with a brief introduction which gives a high-level overview of the chapters contained in Part II. Its first chapter, Chapter 6, then contains a detailed introduction to the main

ideas, in the restricted context of countable model classes. Each of the four subsequent chapters gives a detailed discussion of one of the four main types of universal codes, in the still restricted context of "parametric models" with (essentially) compact parameter spaces. Chapters 11, 12, and 13 deal with general parametric models — including linear regression models — as well as nonparametric models.

Part III Part III gives a detailed treatment of refined MDL. We call it "refined" so as to mark the contrast with the "crude" form of MDL of Chapter 5. It starts with a brief introduction which gives a high-level overview of refined MDL. Chapter 14 deals with refined MDL for model selection. Chapter 15 is about its other two main applications: hypothesis selection (a basis for parametric and nonparametric density estimation) and prediction. Consistency and rate-of-convergence results for refined MDL are detailed in Chapter 16. Refined MDL is placed in its proper context in Chapter 17, in which we discuss its underlying philosophy and compare it to various other approaches.

Compared to Part I, Part II and Part III contain more advanced material, and some prior exposure to statistics may be needed to fully appreciate the developments. Still, all required information-theoretic concepts — invariably related to *universal coding* — are once again discussed at a very basic level. These parts of the book mainly serve as a reference guide, providing a detailed exposition of the main topics in MDL inference. The discussion of each topic includes details which are often left open in the existing literature, but which are important when devising practical applications of MDL. When pondering these details, I noticed that there are several open questions in MDL theory which previously have not been explicitly posed. We explicitly list and number such open questions in Part II and Part III. These parts also contain several new developments: in order to tell a coherent story about MDL, I provide some new results — not published elsewhere — that connect various notions devised by different authors.

> The main innovations are the "distinguishability" interpretation of MDL for finite models in Chapter 6, the "phase transition" view on two-part coding in Chapter 10, the luckiness framework as well as the CNML-1 and CNML-2 extensions of the normalized maximum likelihood code in Chapter 11, and the connections between Césaro and standard KL risk and the use of redundancy rather than resolvability in the convergence theorem for two-part MDL in Chapter 15.

I also found it useful to rephrase and re-prove existing mathematical theorems in a unified way. The many theorems in Part II and Part III usually express results that are similar to existing theorems by various authors, mainly Andrew Barron, Jorma Rissanen, and Bin Yu. Since these theorems were often stated in slightly different contexts, they are hard to compare. In our version, they become easily comparable. Specifically, in Part II, we restrict the treatment to so-called *exponential families* of distributions, which is a weakening of existing results. Yet, the theorems invariably deal with uniform convergence, which is often a strengthening of existing results.

Part IV: Exponential Family Theory The theorems in Part II make heavy use of the general and beautiful theory of *exponential* or, relatedly, *maximum entropy* families of probability distributions. Part IV is an appendix that contains an overview of these families and their mathematical properties. When writing the book, I found that most existing treatments are much too restricted to contain the results that we need in this book. The only general treatments I am aware of (Barndorff-Nielsen 1978; Brown 1986) use measure theory, and give a detailed treatment of behavior at parameters tending to the boundaries of the parameter space. For this reason, they are quite hard to follow. Thus, I decided to write my own overview, which avoids measure theory and boundary issues, but otherwise contains most essential ideas such as sufficiency, mean-value and canonical parameterizations, duality, and maximum entropy interpretations.

Acknowledgments

Tim van Erven, Peter Harremoës, Wouter Koolen, In Jae Myung, Mark Pitt, Teemu Roos, Steven de Rooij, and Tomi Silander read and commented on parts of this text. I would especially like to thank Tim, who provided comments on the entire manuscript.

Mistakes
Of course, the many mistakes which undoubtedly remain in this text are all my (the author's) sole responsibility. I welcome all emails that point out mistakes in the text!

Among those who have helped shape my views on statistical inference, two people stand out: Phil Dawid and Jorma Rissanen. Other people who have

strongly influenced my thinking on these matters are Vijay Balasubramanian, Andrew Barron, Richard Gill, Teemu Roos, Paul Vitányi, Volodya Vovk, and Eric-Jan Wagenmakers. My wife Louise de Rooij made a very visible and colourful contribution. Among the many other people who in some way or other had an impact on this book I should mention Petri Myllymäki, Henry Tirri, Richard Shiffrin, Johan van Benthem, and, last but not least, Herbert, Christa and Wiske Grünwald. As leaders of our research group at CWI (the National Research Institute for Mathematics and Computer Science in the Netherlands), Harry Buhrman and Paul Vitányi provided the pleasant working environment in which this book could be written. The initial parts of this book were written in 2001, while I was visiting the University of California at Santa Cruz. I would like to thank Manfred Warmuth and David Draper for hosting me. Finally and most importantly, I would like to thank my lovely wife Louise for putting up with my foolishness for so long.