1
Why We Wrote This Book and How You Should Read It

The advent of high-speed computing has enabled a transformation in practical statistical methodology. We have entered the age of "machine learning". Roughly, this means that we replace assumptions and approximations by computing power in order to derive statements about characteristics of a dataset. Statisticians and computer scientists have produced an enormous literature on this subject. The jargon includes bootstrap, cross-validation, learning curves, receiver operating characteristic, decision trees, neural nets, boosting, bagging, and so on. These algorithms are becoming increasingly popular in analysis of particle and astrophysics data. The idea of this book is to provide an introduction to these tools and methods in language and context appropriate to the physicist.

Machine learning may be divided into two broad types, "supervised learning" and "unsupervised learning", with due caution toward oversimplifying. Supervised learning can be thought of as fitting a function in a multivariate domain over a set of measured values. Unsupervised learning is exploratory analysis, used when you want to discover interesting features of a dataset. This book is focused on the supervised learning side.

Supervised learning comes in two flavors: classification and regression. Classification aims at separating observations of different kinds such as signal and background. The fitted function in this case takes categorical values. Fitting a function with continuous values is addressed by regression. Fitting a scalar function of a scalar argument by least squares is a well-known regression tool. In case of a vector argument, classification appears to be used more often in modern physics analysis. This is why we focus on classification.

This book is not an introductory probability and statistics text. We assume our readers have been exposed to basic probability theory and to basic methods in parameter estimation such as maximum likelihood. We do not ignore these basic tools, but aim our discussion past the elementary development. Solid understanding of linear algebra, familiarity with multivariate calculus and some exposure to set theory are required as well.

Chapter 2 reviews techniques for parametric likelihood, a subject familiar to most physicists. We include discussion of practical issues such as fits for small statistics and fits near the boundary of a physical region, as well as advanced topics such as sPlots. Goodness of fit measures for univariate and multivariate data

Statistical Analysis Techniques in Particle Physics, First Edition. Ilya Narsky and Frank C. Porter.

are reviewed in Chapter 3. These measures can be applied to distribution fitting by parametric likelihood described in Chapter 2 and nonparametric density estimation described in Chapter 5. Chapter 4 introduces resampling techniques such as the bootstrap, in the context of parameter estimation. Chapter 5 overviews techniques for nonparametric density estimation by histograms and kernel smoothing. The subjects reviewed in these chapters are part of the traditional statistician's toolkit.

In Chapter 6 we turn attention to topics in machine learning. Chapter 6 introduces basic concepts and gives a cursory survey of the material that largely remains beyond the scope of this book. Before learning can begin, data need to be cleaned up and organized in a suitable way. These basic processing steps, in particular treatment of categorical variables, missing values, and outliers, are discussed in Chapter 7. Other important steps, optionally taken before supervised learning starts, are standardizing variable distributions and reducing the data dimensionality. Chapter 8 reviews simple techniques for univariate transformations and advanced techniques such as principal and independent component analysis.

In Chapter 9 we shift the focus to classification. Chapters 9 and 10 are essential for understanding the material in Chapters 11–18. Chapter 9 formalizes the problem of classification and lays out the common workflow for solving this problem. Resampling techniques are revisited here as well, with an emphasis on their application to supervised learning. Chapter 10 explains how the quality of a classification model can be judged and how two (or more) classification models can be compared by a formal statistical test. Some topics in these chapters can be skipped at first reading. In particular, analysis of data with class imbalance, although an important practical issue, is not required to enjoy the rest of the book.

Chapters 11–15 review specific classification techniques. Although many of them can be used for two-class (binary) and multiclass learning, binary classification has been studied more actively than multiclass algorithms. Chapter 16 describes a framework for reducing multiclass learning to a set of binary problems.

Chapter 17 provides a summary of the material learned in Chapters 11–16. Summaries oversimplify and should be interpreted with caution. Bearing this in mind, use this chapter as a practical guide for choosing a classifier appropriate for your analysis.

Chapter 18 reviews methods for selecting the most important variables from all inputs in the data. The importance of a variable is measured by its effect on the predictive power of a classification model.

Bump hunting in multivariate data may be an important component of searches for new physics processes at the Large Hadron Collider, as well as other experiments. We discuss appropriate techniques in Chapter 19. This discussion is focused on multivariate nonparametric searches, in a setting more complex than univariate likelihood fits.

Throughout the book, we illustrate the application of various algorithms to data, either simulated or borrowed from a real physics analysis, using examples of MATLAB code. These examples could be coded in another language. We have chosen MATLAB for two reasons. First, one of the authors, employed by MathWorks, has

been involved in design and implementation of the MATLAB utilities supporting various algorithms described in this book. We thus have intimate knowledge of how these utilities work. Second, MATLAB is a good scripting language. If we used tools developed in the particle physics community, these code snippets would be considerably longer and less transparent.

There are many software suites that provide algorithms described here besides MATLAB. In Chapter 20 we review several software toolkits, whether developed by physicists or not.

We hope that this book can serve both pedagogically and as a reference. If your goal is to learn the broad arsenal of statistical tools for physics analysis, read Chapters 2–9. If you are interested primarily in classification, read Chapters 9 and 10; then choose one or more chapters from Chapters 11–15 for in-depth reading. If you wish to learn a specific classification technique, read Chapters 9 and 10 before digging into the respective chapter or section.

Sections labeled with ✂ present advanced material and can be cut at first reading.

In various places throughout the book we use datasets, either simulated or measured. Some of these datasets can be downloaded from the Machine Learning Repository maintained by University of California in Irvine, http://www.ics.uci.edu/~mlearn.

A fraction of this book is posted at the Caltech High Energy Physics site, http://www.hep.caltech.edu/~NarskyPorter. The posted material includes code for all MATLAB examples, without comments or graphs. Sections not included in the book are posted there as well.