Part I

Introduction

1 Prologue

There is nothing like returning to a place that remains unchanged to find the ways in which you yourself have altered.

Nelson Mandela[1]

That is what learning is. You suddenly understand something you've understood all your life, but in a new way.

Doris Lessing

If we are always arriving and departing, it is also true that we are eternally anchored. One's destination is never a place but rather a new way of looking at things.

Henry Miller[2]

The only real voyage of discovery consists not in seeking new landscapes but in having new eyes.

Marcel Proust

1.1 Machines that learn – some recent history

Statistical learning machines arose as a branch of computer science. These intriguing computer-intensive methods are now being applied to extract useful knowledge from an increasingly wide variety of problems involving oceans of information, heterogeneous variables, and analytically recalcitrant data. Such problems have included:

- predicting fire severity in the US Pacific Northwest (Holden *et al.*, 2008);
- predicting rainfall in Northeastern Thailand (Ingsrisawang *et al.*, 2008);

[1] Nelson Mandela, *A Long Walk to Freedom* © Nelson Rolihlahla Mandela 1994. Reprinted by permission of Little, Brown and Company.
[2] Reproduced with permission of Curtis Brown Group Ltd, London on behalf of the Estate of Henry Miller © Henry Miller 1957. All rights reserved.

- handwriting recognition (Schölkopf and Smola, 2002);
- speech emotion classification (Casale *et al.*, 2008).

Learning machines have also been applied to biomedical problems, such as:

- colon cancer detection derived from 3-D virtual colonoscopies (Jerebko *et al.*, 2003, 2005);
- detecting differential gene expression in microarrays, for data that can involve more than a million single nucleotide polymorphisms (SNPs), in addition to clinical information, over several thousand patients (Díaz-Uriarte and Alvarez de Andrés, 2006);
- predicting short-term hospital survival in lupus patients (Ward *et al.*, 2006);
- finding the most predictive clinical or demographic features for patients having spinal arthritis (Ward *et al.*, 2009).

It is specifically toward this large class of clinical, biomedical, and biological problems that this book is directed. For example, for many important medical problems it is often the case that a large collection of heterogeneous data is available for each patient, including several forms of brain and neuroimaging data, a long stream of clinical values, such as blood pressure, age, cholesterol values, detailed patient histories, and much more. We will see that learning machines can be especially effective for analysis of such heterogeneous data where conventional parametric models are clearly not suitable, that is, where the standard assumptions for those models don't apply, can't easily be verified, or simply don't exist yet.

We alert the reader to the fact that there is a wide and growing range of terminology describing more or less the same area studied here. Prominent among these terms are *data mining*, *pattern recognition*, *knowledge discovery*, and our preferred choice, *machine learning*. Any such method, computer algorithm, or scheme can also be called a *prediction engine* if it claims to predict an outcome, discrete or continuous. From our perspective these mean the same thing.

In this framework, classical statistical procedures such as *logistic regression* or *linear regression* are simply classical examples of learning machines. These are parametric and model-based, as they ask the data to estimate some small number of coefficients (terms, parameters) in the model and often under specific assumptions about the probability distribution of the data (it's normally distributed, say). For example, given a set of features (clinical values, patient

characteristics, numerous expensive measurements), a logistic regression model could be used to estimate the probability that a critically ill lupus patient will not survive the first 72 hours of an initial emergency hospital visit; see Ward *et al.* (2006).

On the other hand, most of the prediction engines we study are nonlinear and nonparametric models. The classic troika of data analysis assumptions – linearity, normality, and equal variance – is usually not invoked for learning machines. So, while the success of many statistical learning machine methods *does* rely on mathematical and statistical ideas (ranging from elementary to profound), the methods typically do not depend on, or derive their inspiration from, classical statistical outlooks.

That they often work as well as they do (or, notably, do not) is systematically studied under the heading of *probabilistic learning theory*; see Devroye *et al.* (1996). This is a rapidly developing area of research but this technical, foundational material is neither elementary nor easily summarized. Hence not even key conclusions are being widely taught to data analysts, and this has inhibited the adoption and understanding of learning machines. In this text we attempt to provide a summary, and also attempt to demystify another aspect of learning machines: the results of a given learning machine may be quite respectable in terms of low prediction error, say, but how the machine got to such a good prediction is often hard to sort out. Here, even a strong background in classical data analysis methods such as regression, correlation, and linear models, while a good thing in itself, may not be of much help, or more likely is just the beginning of understanding.

In dealing systematically with all these factors that have seemingly argued against broad acceptance of learning machines, we hope to keep the discussion relatively uncomplicated. Instead of making each page optically dense with equations we point the reader to some of the mathematical conclusions, but make an effort to keep the details out of sight. As such, the equations appear as heavily redacted and in translation from the original.

By traversing this territory we attempt to make another point: that working with statistical learning machines can push us to think about novel structures and functional connections in our data. This awareness is often counterintuitive, and familiar methods such as simple correlations, or the slightly more evolved partial correlations, are often not sufficient to pin

down these deeper connections. We discuss these problems systematically in Part III: Analysis fundamentals.

Would that we could be as inventive and resourceful in our understanding as Nature is in setting the evidence before us each day.

There is another aspect of this journey that is important to mention up front. That is, many topics will reappear in different settings throughout this book. Among these are overfitting, interactions among predictor variables, measuring variable importances, classification vs. group probability membership, and resampling methods (such as the bootstrap, cross-validation). The reader can expect to uncover other repetitions. We think this reoccurrence of key topics is unavoidable and perhaps a good thing. How each topic arises and is, or is not, resolved in each instance, we hope will be reinforcing threads of the discussion.

Since the validity of the machine predictions do not typically derive mathematical comfort from standard parametric statistical methods (analysis of variance, correlations, regression analysis, normal distribution theory), we will also find that the predictions of statistical learning machines will push us to think carefully about the problem of *validating* those predictions.

There are two points here regarding validation. First, since the machines, these newer models, do not start from assumptions about the data – that it's normally distributed, say – a consequence is that devising procedures for testing the model is harder. Second, most of these procedures are reinforcing in their conclusions and are (often) nearly trivial to implement. Which is all to say that the work put into understanding unfamiliar connections in the data, and validating the machine predictions or estimations, is in itself a good thing and can have lasting scientific merit beyond the mechanics of the statistical learning methods we discuss.

Our first words of caution: we don't propose to locate any shining true model in the biological processes we discuss, nor claim that learning machines are, as a class of statistical procedures, uniquely wonderful (well, often wonderful, but not uniquely so). A statistical learning machine is a procedure that can potentially make good use of difficult data. And *if* it is shown to be effective in making a good prediction about a clinical process or outcome, it can lead us to refined understanding of that process: it can help us *learn* about the process.

Let's see what forms this learning can take . . .

1.2 Twenty canonical questions

Some highlights of this textbook, in the form of key questions, are presented. Some of these questions arise naturally in classical statistical analysis, and some arise specifically when using learning machines. Some of the questions are very open-ended, having multiple answers, some are themselves compound questions, while Question 20 is a trick question. Citations to chapters and sections in the book that provide solutions and discussions are given for each of the *Twenty Canonical Questions*:

(1) Are there any over-arching insights as to why (or when) one learning machine or method might work, or when it might not?

(Section 2.11)

(2) How do we conduct robust selection of the most predictive features, and how do we compare different lists of important features?

(Sections 6.6, 7.3, 9.7)

(3) How do we generate a simple, smaller, more interpretable model given a well fitting but much larger one that fits the data very well? That is, how do we move from an inscrutable, but highly predictive learning machine to a tiny, familiar kind of model? How do we move from an efficient black box to a simple open book?

(Section 5.12)

(4) Why does the use of a very large number of features, say 500K genetic expression values, or a million SNPs, very likely lead to massive overfitting? What exactly is overfitting and why is it a bad thing? How is it possible to do *any* data analysis when we are given 550,009 features (or, 1.2 million features) and only 132 subjects?

(Sections 2.6, 7.8)

(5) Related to (4), how do we deal with the fact that a very large feature list (500K SNPs) may also nicely predict entirely random (permuted) outcomes of the original data; how can we evaluate any feature selection in this situation?

(Sections 6.12, 10.8, 11.10)

(6) How do we interpret or define interactions in a model? What are the unforeseen, or the unwanted consequences of introducing interactions?

(Section 5.6)

(7) How should we do prediction error analysis, and get means or variances of those error estimates, for any single machine?

(Section 11.4)

(8) Related to (7), given that the predictions made by a pair of machines on each subject are correlated, how do we construct error estimate comparisons for the two machines?

(Section 11.12)

(9) Since unbalanced groups are a routine in biology, for example with 25 patients and 250 controls, what are the hazards and remedies?

(Sections 4.5, 10.6, 11.11)

(10) Can data consisting of a circle of red dots (one group) inside an outer circle of green dots (a second group) ever derive from a biologically relevant event? What does this say about simulated data?

(Section 2.11)

(11) How much data do we need in order to state that we have a good model and error estimates? Can classical statistics help with determining sample sizes for obtaining good models with learning machines, given that learning machines are often nonparametric and nonlinear?

(Section 2.11)

(12) It is common that features are quite entangled, having high-order correlation structure among themselves. Dropping features that correlated with each other can be quite mistaken. Why is that? How can very weak predictors, acting jointly, still be highly predictive when acting together?

(Chapter 9)

(13) Given that several models can look very different but lead to nearly identical predictions, does it matter which model is chosen in the end, or, is it necessary to choose any single model?

(Sections 9.4, 12.3)

(14) Closely related to (13), distinct learning machines, and statistical models in general, can be combined into a single, larger and often better model, so what combining methods are mathematically known to be better than others?

(Chapter 12)

(15) How do we estimate the probability of any single subject being in one group or another, rather than making a pure (0,1) prediction: "Doctor, you say I will

probably survive five years, but do I have a 58% chance of survival or an 85% chance?"

(Section 2.8)

(16) What to do with missing data occurring in 3% of a 500K SNP dataset? In 15.4% of the data? Must we always discard cases having missing bits here and there? Can we fill in the missing bits and still get to sound inference about the underlying biology?

(Section 7.8)

(17) How can a learning machine be used to detect – or define – biologically or clinically important subsets?

(Section 2.3)

(18) Suppose we want to make predictions for continuous outcomes, like temperature. Can learning machines help here, too?

(Section 2.2)

(19) How is it that using parts of a good composite clinical score can be more predictive than the score itself?

(Sections 4.3, 6.9)

(20) What are the really clever methods that work really well on *all* data?

(Sections 2.11, 11.14)

1.3 Outline of the book

Given the varied statistical or mathematical background the reader may have, and the relative novelty of learning machine methods, we divide the book into four parts.

Part I Introduction

A survey of the landscape of statistical learning machines is given, followed by a more detailed discussion of some specific machines. Then, we offer a chapter in which examples are given applying several learning machines to three datasets. These datasets include a simple computer-generated one that is still indicative of real-world data, one from published work on clinical predictions for lupus patients, and another also from published work related to a large study of stroke patient functional outcomes.

Part II A machine toolkit

Here we discuss three machines: logistic regression, decision trees, and Random Forests™. Logistic regression is a basic tool in biomedical data analysis, though usually not thought of as a learning machine. It is a parametric prediction scheme, and can give us a small, usually well-understood reference model. Single decision trees are but one instance of a very primitive learning machine, devices that often do surprisingly well, and yet also might not do so well, depending on the data structure, sample size, and more. We then discuss Random Forests, which as the name suggests are built from many (many!) single decision trees. Since its introduction by Leo Breiman and Adele Cutler (Breiman, 2004), it now has its own growing literature showing how it can be improved upon, and how it might not succeed. It also has two new and important incarnations, *Random Jungles* and *Conditional Inference Forests*; see Note 1. Random Forests is guaranteed to be good (under certain reasonable assumptions, according to theory), and can predict (estimate) continuous outcomes as well. Of course, as we will often point out, many other machines may do much better, or just as well, and the practical equivalence of various machines is a conclusion we usually see when there is at least a moderately strong signal in the data.

Part III Analysis fundamentals

No special statistical or mathematical or computer programming maturity is strictly required for this book. We assume the reader's statistical portfolio contains only some awareness of *correlations*, *regression analysis*, and *analysis of variance*. Very good reviews and discussions of these topics can be found, for example, in Agresti and Franklin (2009), Glantz (2002), Sprent and Smeeton (2001), and of course in many other excellent texts. However, we will also include some of the basics about correlation and regression, and offer a full discussion of the problems with both.

Given these analysis basics, we assert that statistical learning machines can push us toward new and unexpected appreciation of how data works and how structure can reveal itself. The Twenty Canonical Questions point in this direction, and as well, crafting the right question can be better than offering a simple answer.

We next discuss the problem of error analysis: estimates of how well a machine is doing. Finally, we address the problem of evaluating the predictive accuracy of two machines, by seeing how they do on each case (subject), one at a time. It is essential to witness that these outcomes on each case (subject) will generate a pair of results, and typically these will be correlated. Therefore, confidence intervals of the difference in error rates need to include this non-independence. While often overlooked in some technical studies of learning machines, this subject has recently (Tango, 1998, 1999, 2000; Newcombe, 1998a–c) been thoroughly upgraded in the statistical literature, and newer methods have demonstrated remarkable improvements over very classical methods. These improved methods, and the need for the improvements, are covered in Chapter 10.

Part IV Machine strategies

The final section of this book outlines a specific, practical strategy: use a learning machine for intensive, thorough processing of the data, and then if the error values seem promising, proceed using some subset of the features in a much smaller, more familiar model, such as logistic regression. Often the initial choice of learning machine (for example, a support vector machine instead of a Random Forest) is not too critical, if there is significant signal in the data, but there are important exceptions. And the choice of a small, interpretable model is not so important either, given a central interest in predictive accuracy. That is, as mentioned above and as discussed frequently below, if any good signal is present in the data then most learning machines will find it, and then many smaller models will fit equally well. There are many paths through the wilderness of validation. The important point is that a learning machine is only a single, initial element in thorough data analysis.

1.4 A comment about example datasets

As mentioned above, we study three biomedical datasets in Chapter 4. One of these, *Cholesterol*, is a simulated dataset, based on two measures of serum cholesterol. The second example, the *Lupus* study, appeared in a published collaboration (Ward *et al.*, 2006). The third example, the large *Stroke* study, also appeared in a published collaboration (König *et al.*, 2007, 2008). The