Appendix

Appendix A
A Summary of Various Models

Table A.1 shows a short summary of several characteristics of the models discussed here. These properties generally hold, but are not always true for every problem. For example, linear discriminant analysis models do not perform feature selection, but there are specialized versions of the model that use regularization to eliminate predictors. Also, the interpretability of a model is subjective. A single tree might be understandable if it is not excessively large and the splits do not involve a large number of categories.

As stated in Chap. 2, no one model is uniformly better than the others. The applicability of a technique is dependent on the type of data being analyzed, the needs of the modeler, and the context of how the model will be used.

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3,

Table A.1: A summary of models and some of their characteristics

Model	Allows $n < p$	Pre-processing	Interpretable	Automatic feature selection	# Tuning parameters	Robust to predictor noise	Computation time
Linear regression†	×	CS, NZV, Corr	✓	×	0	×	✓
Partial least squares	✓	CS	✓	○	1	×	✓
Ridge regression	×	CS, NZV	✓	×	1	×	✓
Elastic net/lasso	✓	CS, NZV	✓	✓	1–2	×	✓
Neural networks	✓	CS, NZV, Corr	×	×	2	×	×
Support vector machines	✓	CS	×	×	1–3	×	×
MARS/FDA	✓		○	✓	1–2	○	○
K-nearest neighbors	✓	CS, NZV	×	×	1	○	✓
Single trees	✓		○	✓	1	✓	✓
Model trees/rules†	✓		○	✓	1–2	✓	✓
Bagged trees	✓		×	✓	0	✓	○
Random forest	✓		×	○	0–1	✓	×
Boosted trees	✓		×	✓	3	✓	×
Cubist†	✓		×	○	2	✓	×
Logistic regression*	×	CS, NZV, Corr	✓	×	0	×	✓
{LQRM}DA*	×	NZV	○	×	0–2	×	✓
Nearest shrunken centroids*	✓	NZV	○	✓	1	×	✓
Naïve Bayes*	✓	NZV	×	×	0–1	○	○
C5.0*	✓		○	✓	0–3	✓	×

†regression only *classification only

Symbols represent affirmative (✓), negative (×), and somewhere in between (○)

- CS = centering and scaling
- NZV = remove near-zero predictors
- Corr = remove highly correlated predictors

Appendix B
An Introduction to R

The R language (Ihaka and Gentleman 1996; R Development Core Team 2010) is a platform for mathematical and statistical computations. It is free in two senses. First, R can be obtained free of charge (although commercial versions exist). Second, anyone can examine or modify the source code. R is released under the *General Public License* (Free Software Foundation June 2007), which outlines how the program can be redistributed.

R is used extensively in this book for several reasons. As just mentioned, anyone can download and use the program. Second, R is an extremely powerful and flexible tool for data analysis, and it contains extensive capabilities for predictive modeling.

The Comprehensive R Archive Network (CRAN) web site contains the source code for the program, as well as compiled versions that are ready to use:

http://cran.r-project.org/

This appendix is intended to be a crash course in basic concepts and syntax for R. More in-depth guides to the language basics are Spector (2008) and Gentleman (2008). The software development life cycle is detailed in R Development Core Team (2008).

B.1 Start-Up and Getting Help

CRAN contains pre-compiled versions of R for Microsoft Windows, Apple OS X, and several versions of Linux. For Windows and OS X, the program comes with a graphical user interface (GUI). When installing complied versions of R for these two operating systems, an icon for R is installed on the computer. To start an interactive session, launch the program using the icon. Alternatively, R can be started at the command line by typing `R`.

Once the program is started, the `q` function (for quit) ends the session.

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3,

```
> # Comments occur after '#' symbols and are not executed
> # Use this command to quit
> q()
```

When quitting, the user will be prompted for options for saving their current work. Note the language is case-sensitive: `Q` could not be used to quit the session.

To get help on a specific topic, such as a function, put a question mark before the function and press enter:

```
> # Get help on the Sweave function
> ?Sweave
```

This opens the `Sweave` help page. One common challenge with R is finding an appropriate function. To search within all the local R functions on your computer, `apropos` will match a keyword against the available functions:

```
> apropos("prop")
 [1] "apropos"                          "getProperties"
 [3] "pairwise.prop.test"               "power.prop.test"
 [5] "prop.table"                       "prop.test"
 [7] "prop.trend.test"                  "reconcilePropertiesAndPrototype"
```

Alternatively, the `RSiteSearch` function conducts an online search of all functions, manuals, contributed documentation, the `R-Help` newsgroup, and other sources for a keyword. For example, to search for different methods to produce ROC curves,

```
> RSiteSearch("roc")
```

will open a web browser and show the matches. The `restrict` argument of this function widens the search (see `?RSiteSearch` for more details).

B.2 Packages

Base R is the nominal system that encompasses the core language features (e.g., the executable program, the fundamental programming framework). Most of the actual R code is contained in distinct modules called *packages*. When R is installed, a small set of core packages is also installed (see R Development Core Team (2008) for the definitive list). However, a large number of packages exist outside of this set. The CRAN web site contains over 4,150 packages for download while the Bioconductor project (Gentleman et al. 2004), an R-based system for computational biology, includes over 600 R packages.

To load a package, the `library` function is used:

```
> # Load the random forests package
> library(randomForest)
> # Show the list of currently loaded packages and other information
> sessionInfo()
```

```
R version 2.15.2 (2012-10-26)
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

locale:
[1] C

attached base packages:
[1] splines   tools     stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] randomForest_4.6-7  BiocInstaller_1.8.3 caret_5.15-045
 [4] foreach_1.4.0       cluster_1.14.3      reshape_0.8.4
 [7] plyr_1.7.1          lattice_0.20-10     Hmisc_3.10-1
[10] survival_2.36-14    weaver_1.24.0       codetools_0.2-8
[13] digest_0.6.0

loaded via a namespace (and not attached):
[1] grid_2.15.2     iterators_1.0.6
```

The function `install.packages` can be used to install additional modules. For example, to install the rpart package for classification and regression trees discussed in Sects. 8.1 and 14.1, the code

```
> install.packages("rpart")
```

can be used. Alternatively, the CRAN web site includes "task views" which group similar packages together. For example, the task view for "Machine Learning" would install a set of predictive modeling packages:

```
> # First install the task view package
> install.packages("ctv")
> # Load the library prior to first use
> library(ctv)
> install.views("MachineLearning")
```

Some packages depend on other packages (or specific versions). The functions `install.packages` and `install.views` will determine additional package requirements and install the necessary dependencies.

B.3 Creating Objects

Anything created in R is an *object*. Objects can be assigned values using "`<-`". For example:

```
> pages <- 97
> town <- "Richmond"
> ## Equals also works, but see Section B.9 below
```

To see the value of an object, simply type it and hit enter. Also, you can explicitly tell R to print the value of the object

```
> pages
  [1] 97
> print(town)
  [1] "Richmond"
```

Another helpful function for understanding the contents of an object is `str` (for structure). As an example, R automatically comes with an object that contains the abbreviated month names.

```
> month.abb
   [1] "Jan" "Feb" "Mar" "Apr" "May" "Jun" "Jul" "Aug" "Sep" "Oct" "Nov"
  [12] "Dec"
> str(month.abb)
   chr [1:12] "Jan" "Feb" "Mar" "Apr" "May" "Jun" "Jul" ...
```

This shows that `month.abb` is a character object with twelve elements. We can also determine the structure of objects that do not contain data, such as the `print` function discussed earlier:

```
> str(print)
  function (x, ...)
> str(sessionInfo)
  function (package = NULL)
```

This is handy for looking up the names of the function arguments. Functions will be discussed in more detail below.

B.4 Data Types and Basic Structures

There are many different core data types in R. The relevant types are numeric, character, factor, and logical types. Logical data can take on value of `TRUE` or `FALSE`. For example, these values can be used to make comparisons or can be assigned to an object:

```
> if(3 > 2) print("greater") else print("less")
  [1] "greater"
> isGreater <- 3 > 2
> isGreater
  [1] TRUE
> is.logical(isGreater)
  [1] TRUE
```

Numeric data encompass integers and double precision (i.e., decimal valued) numbers. To assign a single numeric value to an R object:

```
> x <- 3.6
> is.numeric(x)
```

```
  [1] TRUE
> is.integer(x)
  [1] FALSE
> is.double(x)
  [1] TRUE
> typeof(x)
  [1] "double"
```

Character strings can be created by putting text inside of quotes:

```
> y <- "your ad here"
> typeof(y)
  [1] "character"
> z <- "you can also 'quote' text too"
> z
  [1] "you can also 'quote' text too"
```

Note that R does not restrict the length of character strings.

There are several helpful functions that work on strings. First, `char` counts the number of characters:

```
> nchar(y)
  [1] 12
> nchar(z)
  [1] 29
```

The `grep` function can be used to determine if a substring exists in the character string

```
> grep("ad", y)
  [1] 1
> grep("my", y)
  integer(0)
> # If the string is present, return the whole value
> grep("too", z, value = TRUE)
  [1] "you can also 'quote' text too"
```

So far, the R objects shown have a single value or element. The most basic data structure for holding multiple values of the same type of data is a vector. The most basic method of creating a vector is to use the `c` function (for *combine*). To create a vector of numeric data:

```
> weights <- c(90, 150, 111, 123)
> is.vector(weights)
  [1] TRUE
> typeof(weights)
  [1] "double"
> length(weights)
  [1] 4
```

```
> weights + .25
  [1]  90.25 150.25 111.25 123.25
```

Note that the last command is an example of *vector operations*. Instead of looping over the elements of the vector, vector operations are more concise and efficient operations.

Many functions work on vectors:

```
> mean(weights)
  [1] 118.5
> colors <- c("green", "red", "blue", "red", "white")
> grep("red", colors)
  [1] 2 4
> nchar(colors)
  [1] 5 3 4 3 5
```

An alternate method for storing character data in a vector is to use *factors*. Factors store character data by first determining all unique values in the data, called the factor levels. The character data is then stored as integers that correspond to the factor levels:

```
> colors2 <- as.factor(colors)
> colors2
  [1] green red   blue  red   white
  Levels: blue green red white
> levels(colors2)
  [1] "blue"  "green" "red"   "white"
> as.numeric(colors2)
  [1] 2 3 1 3 4
```

There are a few advantages to storing data in factors. First, less memory is required to store the values since potentially long character strings are saved only once (in the levels) and their occurrences are saved as vectors. Second, the factor vector "remembers" all of the possible values. Suppose we subset the factor vector by removing the first value using a negative integer value:

```
> colors2[-1]
  [1] red   blue  red   white
  Levels: blue green red white
```

Even though the element with a value of "green" was removed, the factor still keeps the same levels. Factors are the primary means of storing discrete variables in R and many classification models use them to specify the outcome data.

To work with a subset of a vector, single brackets can be used in different ways:

```
> weights
  [1]  90 150 111 123
```

```
> # positive integers indicate which elements to keep
> weights[c(1, 4)]
  [1]  90 123
> # negative integers correspond to elements to exclude
> weights[-c(1, 4)]
  [1] 150 111
> # A vector of logical values can be used also but there should
> # be as many logical values as elements
> weights[c(TRUE, TRUE, FALSE, TRUE)]
  [1]  90 150 123
```

Vectors must store the same type of data. An alternative is a list; this is a type of vector that can store objects of any type as elements:

```
> both <- list(colors = colors2, weight = weights)
> is.vector(both)
  [1] TRUE
> is.list(both)
  [1] TRUE
> length(both)
  [1] 2
> names(both)
  [1] "colors" "weight"
```

Lists can be filtered in a similar manner as vectors. However, double brackets return only the element, while single brackets return another list:

```
> both[[1]]
  [1] green red   blue  red   white
  Levels: blue green red white
> is.list(both[[1]])
  [1] FALSE
> both[1]
  $colors
  [1] green red   blue  red   white
  Levels: blue green red white
> is.list(both[1])
  [1] TRUE
> # We can also subset using the name of the list
> both[["colors"]]
  [1] green red   blue  red   white
  Levels: blue green red white
```

Missing values in R are encoded as `NA` values:

```
> probabilities <- c(.05, .67, NA, .32, .90)
> is.na(probabilities)
  [1] FALSE FALSE  TRUE FALSE FALSE
```

```
> # NA is not treated as a character string
> probabilities == "NA"
  [1] FALSE FALSE    NA FALSE FALSE
> # Most functions propagate missing values...
> mean(probabilities)
  [1] NA
> # ... unless told otherwise
> mean(probabilities, na.rm = TRUE)
  [1] 0.485
```

B.5 Working with Rectangular Data Sets

Rectangular data sets usually refer to situations where samples are in rows of a data set while columns correspond to variables (in some domains, this convention is reversed). There are two main structures for rectangular data: matrices and data frames. The main difference between these two types of objects is the type of data that can be stored within them. A matrix can only contain data of the same type (e.g., character or numeric) while data frames must contain columns of the same data type. Matrices are more computationally efficient but are obviously limited.

We can create a matrix using the `matrix` function. Here, we create a numeric vector of integers from one to twelve and use three rows and four columns:

```
> mat <- matrix(1:12, nrow = 3)
> mat
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

The rows and columns can be given names:

```
> rownames(mat) <- c("row 1", "row 2", "row 3")
> colnames(mat) <- c("col1", "col2", "col3", "col4")
> mat
      col1 col2 col3 col4
row 1    1    4    7   10
row 2    2    5    8   11
row 3    3    6    9   12
> rownames(mat)
  [1] "row 1" "row 2" "row 3"
```

Matrices can be subset using method similar to vectors, but rows and columns can be subset separately:

```
> mat[1, 2:3]
```

```
  col2 col3
     4    7
> mat["row 1", "col3"]
  [1] 7
> mat[1,]
  col1 col2 col3 col4
     1    4    7   10
```

One difficulty with subsetting matrices is that dimensions can be *dropped*; if either a single row or column is produced by subsetting a matrix, then a vector is the result:

```
> is.matrix(mat[1,])
  [1] FALSE
> is.vector(mat[1,])
  [1] TRUE
```

One method for avoiding this is to pass the `drop` option to the matrix when subsetting:

```
> mat[1,]
  col1 col2 col3 col4
     1    4    7   10
> mat[1,,drop = FALSE]
       col1 col2 col3 col4
  row 1    1    4    7   10
> is.matrix(mat[1,,drop = FALSE])
  [1] TRUE
> is.vector(mat[1,,drop = FALSE])
  [1] FALSE
```

Data frames can be created using the `data.frame` function:

```
> df <- data.frame(colors = colors2,
+                  time = 1:5)
> df
    colors time
  1  green    1
  2    red    2
  3   blue    3
  4    red    4
  5  white    5
> dim(df)
  [1] 5 2
> colnames(df)
  [1] "colors" "time"
> rownames(df)
  [1] "1" "2" "3" "4" "5"
```

In addition to the subsetting techniques previously shown for matrices, the `$` operator can be used to return single columns while the `subset` function can be used to return more complicated subsets of rows:

```
> df$colors
  [1] green red   blue  red   white
  Levels: blue green red white
> subset(df, colors %in% c("red", "green") & time <= 2)
    colors time
  1  green    1
  2    red    2
```

A helpful function for determining if there are any missing values in a row of a matrix or data frame is the `complete.cases` function, which returns `TRUE` if there are no missing values:

```
> df2 <- df
> # Add missing values to the data frame
> df2[1, 1] <- NA
> df2[5, 2] <- NA
> df2
    colors time
  1   <NA>    1
  2    red    2
  3   blue    3
  4    red    4
  5  white   NA
> complete.cases(df2)
  [1] FALSE  TRUE  TRUE  TRUE FALSE
```

B.6 Objects and Classes

Each object has at least one type or *class* associated with it. The class of an object declares what it is (e.g., a character string, linear model, web site URL). The class defines the structure of an object (i.e., how it is stored) and the possible operations associated with this type of object (called *methods* for the class). For example, if some sort of model object is created, it may be of interest to:

- *Print* the model details for understanding
- *Plot* the model for visualization, or
- *Predict* new samples

In this case, `print`, `plot`, and `predict` are some of the possible methods for that particular type of model (as determined by its class). This paradigm is called object-oriented programming.

We can quickly determine the class of the previous objects:

```
> pages
  [1] 97
> class(pages)
  [1] "numeric"
> town
  [1] "Richmond"
> class(town)
  [1] "character"
```

When the user directs R to perform some operation, such as creating predictions from a model object, the class determines the specific code for the prediction equation. This is called *method dispatch*. There are two main techniques for object-oriented programming in R: S3 classes and S4 classes. The S3 approach is more simplistic than S4 and is used by many packages. S4 methods are more powerful than S3 methods but are too complex to adequately describe in this overview. Chambers (2008) describes these techniques in greater detail.

With S3 methods, the naming convention is to use dots to separate classes and methods. For example, `summary.lm` is the function that is used to compute summary values for an object that has the `lm` class (this class is to fit linear models, such as linear regression analysis). Suppose a user created an object called `myModel` using the `lm` function. The command

```
modelSummary <- summary(myModel)
```

calculates the common descriptive statistics for the model. R sees that `myModel` has class `lm`, so it executes the code in the function `summary.lm`.

For this text, it is important to understand the concept of objects, classes, and methods. However, these concepts will be used at a high level; the code contained in the book rarely delves into the technical minutia "under the hood." For example, the `predict` function will be used extensively, but the use will not be required to know which specific method is executed.

B.7 R Functions

In R, modular pieces of code can be collected in functions. Many functions have already been used in this section, such as the `library` function that loads packages. Functions have *arguments*: specific slots that are used to pass objects into the function. In R, arguments are named (unlike other languages, such as matlab). For example, the function for reading data stored in comma delimited format (CSV) into an R object has these arguments:

```
> str(read.csv)
  function (file, header = TRUE, sep = ",", quote = "¨", dec = ".",
      fill = TRUE, comment.char = "", ...)
```

where `file` is a character string that points to the CSV file and `header` indicates whether the initial row corresponds to variable names. The `file` argument has no default value and the function will result in an error if no file name is specified. Since these functions are named, they can be called in several different ways:

```
> read.csv("data.csv")
> read.csv(header = FALSE, file = "data.csv")
```

Notice that the `read.csv` function has an argument at the end with three dots. This means that other arguments can be added to the `read.csv` function call that are passed to a specific function within the code for `read.csv`. In this case, the code uses another function called `read.table` that is more general. The `read.table` contains an argument called `na.strings` that is absent from `read.csv`. This argument tells R which character values indicate a missing value in the file. Using

```
> read.csv("data.csv", na.strings = "?")
```

has the effect of passing the argument `na.strings = "?"` from the `read.csv` function to the `read.table` function. Note that this argument must be named if it is to be passed through. The three dots are used extensively in the computing sections of each chapter.

B.8 The Three Faces of `=`

So far, the `=` symbol has been used in several different contexts:

1. Creating objects, such as `x = 3`
2. Testing for equivalence: `x == 4`
3. Specifying values to function arguments: `read.csv(header = FALSE)`

This can be confusing for newcomers. For example:

```
> new = subset(old, subset = value == "blue", drop = FALSE)
```

uses the symbol four times across all three cases. One method for avoiding confusion is to use `<-` as the assignment operator.

B.9 The **AppliedPredictiveModeling** Package

This package serves as a companion to the book and includes many of the data sets used here that are not already available in other R packages. It also includes the R code used throughout the chapters and R functions. The package is available on CRAN.

Table B.1: A survey of commands to produce class probabilities across different packages

Object class	Package	`predict` Function syntax
`lda`	MASS	`predict(object)` (no options needed)
`glm`	stats	`predict(object, type = "response")`
`gbm`	gbm	`predict(object, type = "response", n.trees)`
`mda`	mda	`predict(object, type = "posterior")`
`rpart`	rpart	`predict(object, type = "prob")`
`Weka_classifier`	RWeka	`predict(object, type = "probability")`
`LogitBoost`	caTools	`predict(object, type = "raw", nIter)`

The `train` function in the caret package uses a common syntax of `predict(object, type = "prob")`

B.10 The caret Package

The caret package (short for **C**lassification **A**nd **RE**gression **T**raining) was created to streamline the process for building and evaluating predictive models. Using the package, a practitioner can quickly evaluate many different types of models to find the more appropriate tool for their data.

The beauty of R is that it provides a large and diverse set of modeling packages. However, since these packages are created by many different people over time, there are a minimal set of conventions that are common to each model. For example, Table B.1 shows the syntax for calculating class probabilities for several different types of classification models. Remembering the syntactical variations can be difficult and this discourages users from evaluating a variety of models. One method to reduce this complexity is to provide a unified interface to functions for model building and prediction. caret provides such an interface for across a wide vary of models (over 140). The package also provides many options for data pre-processing and resampling-based parameter tuning techniques (Chaps. 3 and 4).

In this text, resampling is the primary approach for optimizing predictive models with tuning parameters. To do this, many alternate versions of the training set are used to train the model and predict a holdout set. This process is repeated many times to get performance estimates that generalize to new data sets. Each of the resampled data sets is independent of the others, so there is no formal requirement that the models must be run sequentially. If a computer with multiple processors or cores is available, the computations could be spread across these "workers" to increase the computational efficiency. caret leverages one of the parallel processing frameworks in R to do just this. The foreach package allows R code to be run either sequentially or in parallel using several different technologies, such as the multicore or Rmpi packages (see Schmidberger et al. (2009) for summaries and descriptions of the available options). There are several R packages that work with foreach

to implement these techniques, such as doMC (for multicore) or doMPI (for Rmpi).

To tune a predictive model using multiple workers, the syntax in the caret package functions (e.g., `train`, `rfe` or `sbf`) does not change. A separate function is used to "register" the parallel processing technique and specify the number of workers to use. For example, to use the multicore package (not available on Windows) with five cores on the same machine, the package is loaded and then registered:

```
> library(doMC)
> registerDoMC(cores = 5)
> ## All subsequent models are then run in parallel
> model <- train(y ~ ., data = training, method = "rf")
```

The syntax for other packages associated with foreach is very similar. Note that as the number of workers increases, the memory required also increases. For example, using five workers would keep a total of six versions of the data in memory. If the data are large or the computational model is demanding, performance can be affected if the amount of required memory exceeds the physical amount available.

Does this help reduce the time to fit models? The job scheduling data (Chap. 17) was modeled multiple times with different number of workers for several models. Random forest was used with 2,000 trees and tuned over 10 values of m_{try}. Variable importance calculations were also conducted during each model fit. Linear discriminant analysis was also run, as was a cost-sensitive radial basis function support vector machine (tuned over 15 cost values). All models were tuned using five repeats of 10-fold cross-validation. The results are shown in Fig. B.1. The y-axis corresponds to the total execution time (encompassing model tuning and the final model fit) versus the number of workers. Random forest clearly took the longest to train and the LDA models were very computationally efficient. The total time (in minutes) decreased as the number of workers increase but stabilized around seven workers. The data for this plot were generated in a randomized fashion so that there should be no bias in the run order. The bottom right panel shows the *speedup* which is the sequential time divided by the parallel time. For example, a speedup of three indicates that the parallel version was three times faster than the sequential version. At best, parallelization can achieve linear speedups; that is, for M workers, the parallel time is $1/M$. For these models, the speedup is close to linear until four or five workers are used. After this, there is a small improvement in performance. Since LDA is already computationally efficient, the speed-up levels off more rapidly than the other models. While not linear, the decrease in execution time is helpful—a nearly 10 h model fit was decreased to about 90 min.

Note that some models, especially those using the RWeka package, may not be able to be run in parallel due to the underlying code structure.

One additional "trick" that `train` exploits to increase computational efficiency is to use sub-models; a single model fit can produce predictions for

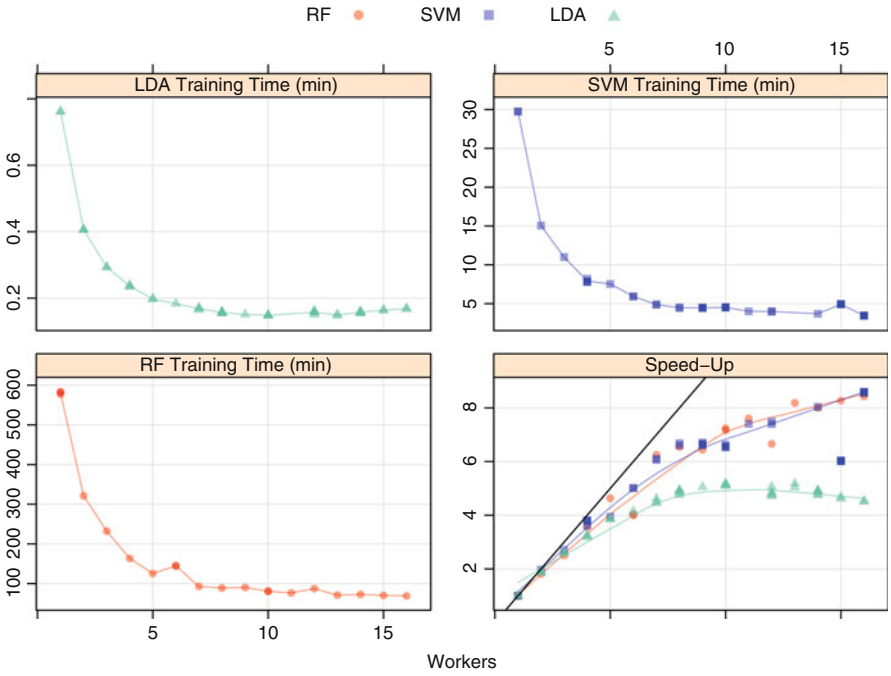

Fig. B.1: Three models run using different numbers of workers. The y-axis is either the execution time in minutes or the speed-up (in the *bottom right panel*)

multiple tuning parameters. For example, in most implementations of boosted models, a model trained on B boosting iterations can produce predictions for models for iterations less than B. For the grant data, a `gbm` model was fit that evaluated 200 distinct combinations of the three tuning parameters (see Fig. 14.10). In reality, `train` only created objects for 40 models and derived the other predictions from these objects.

More detail on the caret package can be found in Kuhn (2008) or the four extended manuals (called "vignettes") on the package web site (Kuhn 2010).

B.11 Software Used in this Text

The excellent `Sweave` function (Leisch 2002a,b) in R enabled data analysis code to be integrated within the content of this text. The function executed the R code and replaced the code with the items produced by the code, such as text, figures, and tables. All the software and data used here are publicly available at the time of this writing. The R packages AppliedPredictiveModeling and

caret include many of the data sets. For data that could not be included, the AppliedPredictiveModeling package includes R code to recreate the data set used in this text. An extensive list of packages and functions in R related to reproducible research can be found on CRAN:

http://cran.r-project.org/web/views/ReproducibleResearch.html

Version 2.15.2 (2012-10-26) of R was used in conjunction with the following package versions: AppliedPredictiveModeling (1.01), arules (1.0-12), C50 (0.1.0-013), caret (5.15-045), coin (1.0-21), CORElearn (0.9.40), corrplot (0.70), ctv (0.7-4), Cubist (0.0.12), desirability (1.05), DMwR (0.2.3), doBy (4.5-5), doMC (1.2.5), DWD (0.10), e1071 (1.6-1), earth (3.2-3), elasticnet (1.1), ellipse (0.3-7), gbm (1.6-3.2), glmnet (1.8-2), Hmisc (3.10-1), ipred (0.9-1), kernlab (0.9-15), klaR (0.6-7), lars (1.1), latticeExtra (0.6-24), lattice (0.20-10), MASS (7.3-22), mda (0.4-2), minerva (1.2), mlbench (2.1-1), nnet (7.3-5), pamr (1.54), partykit (0.1-4), party (1.0-3), pls (2.3-0), plyr (1.7.1), pROC (1.5.4), proxy (0.4-9), QSARdata (1.02), randomForest (4.6-7), RColorBrewer (1.0-5), reshape2 (1.2.1), reshape (0.8.4), rms (3.6-0), rpart (4.0-3), RWeka (0.4-12), sparseLDA (0.1-6), subselect (0.12-2), svmpath (0.952), and tabplot (0.12). Some of these packages are not directly related to predictive modeling but were used to compile or format the content or for visualization.

Appendix C
Interesting Web Sites

Software

http://www.r-project.org
This is the main R web site with links to announcements, manuals, books, conference, and other information.

http://cran.r-project.org
CRAN, or the Comprehensive R Archive Network, is the primary repository for R and numerous add-on packages.

http://cran.r-project.org/web/views/MachineLearning.html
The machine learning Task View is a list of many predictive modeling packages in R.

http://caret.r-forge.r-project.org
The caret package is hosted here.

http://www.rulequest.com
RuleQuest releases commercial and open-source versions of Cubist and C5.0.

http://rattle.togaware.com
Rattle (Williams 2011) is a graphical user interface for R predictive models.

http://www.cs.waikato.ac.nz/ml/weka/
Weka is collection of Java programs for data mining.

http://orange.biolab.si
Orange is an open-source, cross-platform graphical user interface to many machine learning tools. The interface is a "pipeline" where users piece together components to create a workflow.

http://www.knime.org
"KNIME (Konstanz Information Miner) is a user-friendly and comprehensive open-source data integration, processing, analysis, and exploration platform."

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3,

http://www.spss.com/software/modeler
The IBM SPSS Modeler, formerly called Clemintine, is a visual platform for model building.

http://www.sas.com/technologies/analytics/datamining/miner
A SAS product for data mining.

Other programs are listed at http://www.kdnuggets.com/software/suites.html.

Competitions

http://www.kaggle.com

http://tunedit.org

Data Sets

http://archive.ics.uci.edu/ml
The University of California (Irvine) is a well-known location for classification and regression data sets.

http://www.kdnuggets.com/datasets
The Association For Computing Machinery (ACM) has a special interest group on Knowledge Discovery in Data (KDD). The KDD group organizes annual machine learning competitions.

http://fueleconomy.gov
A web site run by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency that lists different estimates of fuel economy for passenger cars and trucks.

http://www.cheminformatics.org
This web site contains many examples of computational chemistry data sets.

http://www.ncbi.nlm.nih.gov/geo
The NCBI GEO web site is "a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community."

References

Abdi H, Williams L (2010). "Principal Component Analysis." *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(4), 433–459.

Agresti A (2002). *Categorical Data Analysis*. Wiley–Interscience.

Ahdesmaki M, Strimmer K (2010). "Feature Selection in Omics Prediction Problems Using CAT Scores and False Nondiscovery Rate Control." *The Annals of Applied Statistics*, **4**(1), 503–519.

Alin A (2009). "Comparison of PLS Algorithms when Number of Objects is Much Larger than Number of Variables." *Statistical Papers*, **50**, 711–720.

Altman D, Bland J (1994). "Diagnostic Tests 3: Receiver Operating Characteristic Plots." *British Medical Journal*, **309**(6948), 188.

Ambroise C, McLachlan G (2002). "Selection Bias in Gene Extraction on the Basis of Microarray Gene–Expression Data." *Proceedings of the National Academy of Sciences*, **99**(10), 6562–6566.

Amit Y, Geman D (1997). "Shape Quantization and Recognition with Randomized Trees." *Neural Computation*, **9**, 1545–1588.

Armitage P, Berry G (1994). *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford, 3rd edition.

Artis M, Ayuso M, Guillen M (2002). "Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims." *The Journal of Risk and Insurance*, **69**(3), 325–340.

Austin P, Brunner L (2004). "Inflation of the Type I Error Rate When a Continuous Confounding Variable Is Categorized in Logistic Regression Analyses." *Statistics in Medicine*, **23**(7), 1159–1178.

Ayres I (2007). *Super Crunchers: Why Thinking–By–Numbers Is The New Way To Be Smart*. Bantam.

Barker M, Rayens W (2003). "Partial Least Squares for Discrimination." *Journal of Chemometrics*, **17**(3), 166–173.

Batista G, Prati R, Monard M (2004). "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data." *ACM SIGKDD Explorations Newsletter*, **6**(1), 20–29.

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3,

Bauer E, Kohavi R (1999). "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants." *Machine Learning*, **36**, 105–142.

Becton Dickinson and Company (1991). *ProbeTec ET Chlamydia trachomatis and Neisseria gonorrhoeae Amplified DNA Assays (Package Insert)*.

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000). "Tissue Classification with Gene Expression Profiles." *Journal of Computational Biology*, **7**(3), 559–583.

Bentley J (1975). "Multidimensional Binary Search Trees Used for Associative Searching." *Communications of the ACM*, **18**(9), 509–517.

Berglund A, Kettaneh N, Uppgård L, Wold S, DR NB, Cameron (2001). "The GIFI Approach to Non–Linear PLS Modeling." *Journal of Chemometrics*, **15**, 321–336.

Berglund A, Wold S (1997). "INLR, Implicit Non–Linear Latent Variable Regression." *Journal of Chemometrics*, **11**, 141–156.

Bergmeir C, Benitez JM (2012). "Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS." *Journal of Statistical Software*, **46**(7), 1–26.

Bergstra J, Casagrande N, Erhan D, Eck D, Kégl B (2006). "Aggregate Features and AdaBoost for Music Classification." *Machine Learning*, **65**, 473–484.

Berntsson P, Wold S (1986). "Comparison Between X-ray Crystallographic Data and Physiochemical Parameters with Respect to Their Information About the Calcium Channel Antagonist Activity of 4-Phenyl-1,4-Dihydropyridines." *Quantitative Structure-Activity Relationships*, **5**, 45–50.

Bhanu B, Lin Y (2003). "Genetic Algorithm Based Feature Selection for Target Detection in SAR Images." *Image and Vision Computing*, **21**, 591–608.

Bishop C (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Bishop C (2006). *Pattern Recognition and Machine Learning*. Springer.

Bland J, Altman D (1995). "Statistics Notes: Multiple Significance Tests: The Bonferroni Method." *British Medical Journal*, **310**(6973), 170–170.

Bland J, Altman D (2000). "The Odds Ratio." *British Medical Journal*, **320**(7247), 1468.

Bohachevsky I, Johnson M, Stein M (1986). "Generalized Simulated Annealing for Function Optimization." *Technometrics*, **28**(3), 209–217.

Bone R, Balk R, Cerra F, Dellinger R, Fein A, Knaus W, Schein R, Sibbald W (1992). "Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis." *Chest*, **101**(6), 1644–1655.

Boser B, Guyon I, Vapnik V (1992). "A Training Algorithm for Optimal Margin Classifiers." In "Proceedings of the Fifth Annual Workshop on Computational Learning Theory," pp. 144–152.

Boulesteix A, Strobl C (2009). "Optimal Classifier Selection and Negative Bias in Error Rate Estimation: An Empirical Study on High–Dimensional Prediction." *BMC Medical Research Methodology*, **9**(1), 85.

Box G, Cox D (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252.

Box G, Hunter W, Hunter J (1978). *Statistics for Experimenters*. Wiley, New York.

Box G, Tidwell P (1962). "Transformation of the Independent Variables." *Technometrics*, **4**(4), 531–550.

Breiman L (1996a). "Bagging Predictors." *Machine Learning*, **24**(2), 123–140.

Breiman L (1996b). "Heuristics of Instability and Stabilization in Model Selection." *The Annals of Statistics*, **24**(6), 2350–2383.

Breiman L (1996c). "Technical Note: Some Properties of Splitting Criteria." *Machine Learning*, **24**(1), 41–47.

Breiman L (1998). "Arcing Classifiers." *The Annals of Statistics*, **26**, 123–140.

Breiman L (2000). "Randomizing Outputs to Increase Prediction Accuracy." *Mach. Learn.*, **40**, 229–242. ISSN 0885-6125.

Breiman L (2001). "Random Forests." *Machine Learning*, **45**, 5–32.

Breiman L, Friedman J, Olshen R, Stone C (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Bridle J (1990). "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition." In "Neurocomputing: Algorithms, Architectures and Applications," pp. 227–236. Springer–Verlag.

Brillinger D (2004). "Some Data Analyses Using Mutual Information." *Brazilian Journal of Probability and Statistics*, **18**(6), 163–183.

Brodnjak-Vonina D, Kodba Z, Novi M (2005). "Multivariate Data Analysis in Classification of Vegetable Oils Characterized by the Content of Fatty Acids." *Chemometrics and Intelligent Laboratory Systems*, **75**(1), 31–43.

Brown C, Davis H (2006). "Receiver Operating Characteristics Curves and Related Decision Measures: A Tutorial." *Chemometrics and Intelligent Laboratory Systems*, **80**(1), 24–38.

Bu G (2009). "Apolipoprotein E and Its Receptors in Alzheimer's Disease: Pathways, Pathogenesis and Therapy." *Nature Reviews Neuroscience*, **10**(5), 333–344.

Buckheit J, Donoho DL (1995). "WaveLab and Reproducible Research." In A Antoniadis, G Oppenheim (eds.), "Wavelets in Statistics," pp. 55–82. Springer-Verlag, New York.

Burez J, Van den Poel D (2009). "Handling Class Imbalance In Customer Churn Prediction." *Expert Systems with Applications*, **36**(3), 4626–4636.

Cancedda N, Gaussier E, Goutte C, Renders J (2003). "Word–Sequence Kernels." *The Journal of Machine Learning Research*, **3**, 1059–1082.

Caputo B, Sim K, Furesjo F, Smola A (2002). "Appearance–Based Object Recognition Using SVMs: Which Kernel Should I Use?" In "Proceedings of NIPS Workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision," .

Strobl C, Carolin C, Boulesteix A-L, Augustin T (2007). "Unbiased Split Selection for Classification Trees Based on the Gini Index." *Computational Statistics & Data Analysis*, **52**(1), 483–501.

Castaldi P, Dahabreh I, Ioannidis J (2011). "An Empirical Assessment of Validation Practices for Molecular Classifiers." *Briefings in Bioinformatics*, **12**(3), 189–202.

Chambers J (2008). *Software for Data Analysis: Programming with R*. Springer.

Chan K, Loh W (2004). "LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees." *Journal of Computational and Graphical Statistics*, **13**(4), 826–852.

Chang CC, Lin CJ (2011). "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology*, **2**, 27: 1–27:27.

Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002). "SMOTE: Synthetic Minority Over–Sampling Technique." *Journal of Artificial Intelligence Research*, **16**(1), 321–357.

Chun H, Keleş S (2010). "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1), 3–25.

Chung D, Keles S (2010). "Sparse Partial Least Squares Classification for High Dimensional Data." *Statistical Applications in Genetics and Molecular Biology*, **9**(1), 17.

Clark R (1997). "OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets." *Journal of Chemical Information and Computer Sciences*, **37**(6), 1181–1188.

Clark T (2004). "Can Out–of–Sample Forecast Comparisons Help Prevent Overfitting?" *Journal of Forecasting*, **23**(2), 115–139.

Clemmensen L, Hastie T, Witten D, Ersboll B (2011). "Sparse Discriminant Analysis." *Technometrics*, **53**(4), 406–413.

Cleveland W (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association*, **74**(368), 829–836.

Cleveland W, Devlin S (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association*, pp. 596–610.

Cohen G, Hilario M, Pellegrini C, Geissbuhler A (2005). "SVM Modeling via a Hybrid Genetic Strategy. A Health Care Application." In R Engelbrecht, AGC Lovis (eds.), "Connecting Medical Informatics and Bio–Informatics," pp. 193–198. IOS Press.

Cohen J (1960). "A Coefficient of Agreement for Nominal Data." *Educational and Psychological Measurement*, **20**, 37–46.

Cohn D, Atlas L, Ladner R (1994). "Improving Generalization with Active Learning." *Machine Learning*, **15**(2), 201–221.

Cornell J (2002). *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*. Wiley, New York, NY.

Cortes C, Vapnik V (1995). "Support–Vector Networks." *Machine Learning*, **20**(3), 273–297.

Costa N, Lourenco J, Pereira Z (2011). "Desirability Function Approach: A Review and Performance Evaluation in Adverse Conditions." *Chemometrics and Intelligent Lab Systems*, **107**(2), 234–244.

Cover TM, Thomas JA (2006). *Elements of Information Theory*. Wiley–Interscience.

Craig-Schapiro R, Kuhn M, Xiong C, Pickering E, Liu J, Misko TP, Perrin R, Bales K, Soares H, Fagan A, Holtzman D (2011). "Multiplexed Immunoassay Panel Identifies Novel CSF Biomarkers for Alzheimer's Disease Diagnosis and Prognosis." *PLoS ONE*, **6**(4), e18850.

Cruz-Monteagudo M, Borges F, Cordeiro MND (2011). "Jointly Handling Potency and Toxicity of Antimicrobial Peptidomimetics by Simple Rules from Desirability Theory and Chemoinformatics." *Journal of Chemical Information and Modeling*, **51**(12), 3060–3077.

Davison M (1983). *Multidimensional Scaling*. John Wiley and Sons, Inc.

Dayal B, MacGregor J (1997). "Improved PLS Algorithms." *Journal of Chemometrics*, **11**, 73–85.

de Jong S (1993). "SIMPLS: An Alternative Approach to Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems*, **18**, 251–263.

de Jong S, Ter Braak C (1994). "Short Communication: Comments on the PLS Kernel Algorithm." *Journal of Chemometrics*, **8**, 169–174.

de Leon M, Klunk W (2006). "Biomarkers for the Early Diagnosis of Alzheimer's Disease." *The Lancet Neurology*, **5**(3), 198–199.

Defernez M, Kemsley E (1997). "The Use and Misuse of Chemometrics for Treating Classification Problems." *TrAC Trends in Analytical Chemistry*, **16**(4), 216–221.

DeLong E, DeLong D, Clarke-Pearson D (1988). "Comparing the Areas Under Two Or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics*, **44**(3), 837–45.

Derksen S, Keselman H (1992). "Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables." *British Journal of Mathematical and Statistical Psychology*, **45**(2), 265–282.

Derringer G, Suich R (1980). "Simultaneous Optimization of Several Response Variables." *Journal of Quality Technology*, **12**(4), 214–219.

Dietterich T (2000). "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." *Machine Learning*, **40**, 139–158.

Dillon W, Goldstein M (1984). *Multivariate Analysis: Methods and Applications*. Wiley, New York.

Dobson A (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC.

Drucker H, Burges C, Kaufman L, Smola A, Vapnik V (1997). "Support Vector Regression Machines." *Advances in Neural Information Processing Systems*, pp. 155–161.

Drummond C, Holte R (2000). "Explicitly Representing Expected Cost: An Alternative to ROC Representation." In "Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining," pp. 198–207.

Duan K, Keerthi S (2005). "Which is the Best Multiclass SVM Method? An Empirical Study." *Multiple Classifier Systems*, pp. 278–285.

Dudoit S, Fridlyand J, Speed T (2002). "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, **97**(457), 77–87.

Duhigg C (2012). "How Companies Learn Your Secrets." *The New York Times*. URL http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Dunn W, Wold S (1990). "Pattern Recognition Techniques in Drug Design." In C Hansch, P Sammes, J Taylor (eds.), "Comprehensive Medicinal Chemistry," pp. 691–714. Pergamon Press, Oxford.

Dwyer D (2005). "Examples of Overfitting Encountered When Building Private Firm Default Prediction Models." *Technical report*, Moody's KMV.

Efron B (1983). "Estimating the Error Rate of a Prediction Rule: Improvement on Cross–Validation." *Journal of the American Statistical Association*, pp. 316–331.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). "Least Angle Regression." *The Annals of Statistics*, **32**(2), 407–499.

Efron B, Tibshirani R (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science*, pp. 54–75.

Efron B, Tibshirani R (1997). "Improvements on Cross–Validation: The 632+ Bootstrap Method." *Journal of the American Statistical Association*, **92**(438), 548–560.

Eilers P, Boer J, van Ommen G, van Houwelingen H (2001). "Classification of Microarray Data with Penalized Logistic Regression." In "Proceedings of SPIE," volume 4266, p. 187.

Eugster M, Hothorn T, Leisch F (2008). "Exploratory and Inferential Analysis of Benchmark Experiments." *Ludwigs-Maximilians-Universität München, Department of Statistics, Tech. Rep*, **30**.

Everitt B, Landau S, Leese M, Stahl D (2011). *Cluster Analysis*. Wiley.

Ewald B (2006). "Post Hoc Choice of Cut Points Introduced Bias to Diagnostic Research." *Journal of clinical epidemiology*, **59**(8), 798–801.

Fanning K, Cogger K (1998). "Neural Network Detection of Management Fraud Using Published Financial Data." *International Journal of Intelligent Systems in Accounting, Finance & Management*, **7**(1), 21–41.

Faraway J (2005). *Linear Models with R*. Chapman & Hall/CRC, Boca Raton.

Fawcett T (2006). "An Introduction to ROC Analysis." *Pattern Recognition Letters*, **27**(8), 861–874.

Fisher R (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, **7**(2), 179–188.

Forina M, Casale M, Oliveri P, Lanteri S (2009). "CAIMAN brothers: A Family of Powerful Classification and Class Modeling Techniques." *Chemometrics and Intelligent Laboratory Systems*, **96**(2), 239–245.

Frank E, Wang Y, Inglis S, Holmes G (1998). "Using Model Trees for Classification." *Machine Learning*.

Frank E, Witten I (1998). "Generating Accurate Rule Sets Without Global Optimization." *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 144–151.

Free Software Foundation (June 2007). *GNU General Public License*.

Freund Y (1995). "Boosting a Weak Learning Algorithm by Majority." *Information and Computation*, **121**, 256–285.

Freund Y, Schapire R (1996). "Experiments with a New Boosting Algorithm." *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156.

Friedman J (1989). "Regularized Discriminant Analysis." *Journal of the American Statistical Association*, **84**(405), 165–175.

Friedman J (1991). "Multivariate Adaptive Regression Splines." *The Annals of Statistics*, **19**(1), 1–141.

Friedman J (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, **29**(5), 1189–1232.

Friedman J (2002). "Stochastic Gradient Boosting." *Computational Statistics and Data Analysis*, **38**(4), 367–378.

Friedman J, Hastie T, Tibshirani R (2000). "Additive Logistic Regression: A Statistical View of Boosting." *Annals of Statistics*, **38**, 337–374.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22.

Geisser S (1993). *Predictive Inference: An Introduction*. Chapman and Hall.

Geladi P, Kowalski B (1986). "Partial Least-Squares Regression: A Tutorial." *Analytica Chimica Acta*, **185**, 1–17.

Geladi P, Manley M, Lestander T (2003). "Scatter Plotting in Multivariate Data Analysis." *Journal of Chemometrics*, **17**(8–9), 503–511.

Gentleman R (2008). *R Programming for Bioinformatics*. CRC Press.

Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber M, Iacus S, Irizarry R, Leisch F, Li C, Mächler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004). "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**(10), R80.

Giuliano K, DeBiasio R, Dunlay R, Gough A, Volosky J, Zock J, Pavlakis G, Taylor D (1997). "High–Content Screening: A New Approach to Easing Key Bottlenecks in the Drug Discovery Process." *Journal of Biomolecular Screening*, **2**(4), 249–259.

Goldberg D (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison–Wesley, Boston.

Golub G, Heath M, Wahba G (1979). "Generalized Cross–Validation as a Method for Choosing a Good Ridge Parameter." *Technometrics*, **21**(2), 215–223.

Good P (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer.

Gowen A, Downey G, Esquerre C, O'Donnell C (2010). "Preventing Over–Fitting in PLS Calibration Models of Near-Infrared (NIR) Spectroscopy Data Using Regression Coefficients." *Journal of Chemometrics*, **25**, 375–381.

Graybill F (1976). *Theory and Application of the Linear Model*. Wadsworth & Brooks, Pacific Grove, CA.

Guo Y, Hastie T, Tibshirani R (2007). "Regularized Linear Discriminant Analysis and its Application in Microarrays." *Biostatistics*, **8**(1), 86–100.

Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, Sriram S (2006). "Modeling Customer Lifetime Value." *Journal of Service Research*, **9**(2), 139–155.

Guyon I, Elisseeff A (2003). "An Introduction to Variable and Feature Selection." *The Journal of Machine Learning Research*, **3**, 1157–1182.

Guyon I, Weston J, Barnhill S, Vapnik V (2002). "Gene Selection for Cancer Classification Using Support Vector Machines." *Machine Learning*, **46**(1), 389–422.

Hall M, Smith L (1997). "Feature Subset Selection: A Correlation Based Filter Approach." *International Conference on Neural Information Processing and Intelligent Information Systems*, pp. 855–858.

Hall P, Hyndman R, Fan Y (2004). "Nonparametric Confidence Intervals for Receiver Operating Characteristic Curves." *Biometrika*, **91**, 743–750.

Hampel H, Frank R, Broich K, Teipel S, Katz R, Hardy J, Herholz K, Bokde A, Jessen F, Hoessler Y (2010). "Biomarkers for Alzheimer's Disease: Academic, Industry and Regulatory Perspectives." *Nature Reviews Drug Discovery*, **9**(7), 560–574.

Hand D, Till R (2001). "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning*, **45**(2), 171–186.

Hanley J, McNeil B (1982). "The Meaning and Use of the Area under a Receiver Operating (ROC) Curvel Characteristic." *Radiology*, **143**(1), 29–36.

Hardle W, Werwatz A, Müller M, Sperlich S, Hardle W, Werwatz A, Müller M, Sperlich S (2004). "Nonparametric Density Estimation." In "Nonparametric and Semiparametric Models," pp. 39–83. Springer Berlin Heidelberg.

Harrell F (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.

Hastie T, Pregibon D (1990). "Shrinking Trees." *Technical report*, AT&T Bell Laboratories Technical Report.

Hastie T, Tibshirani R (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hastie T, Tibshirani R (1996). "Discriminant Analysis by Gaussian Mixtures." *Journal of the Royal Statistical Society. Series B*, pp. 155–176.

Hastie T, Tibshirani R, Buja A (1994). "Flexible Discriminant Analysis by Optimal Scoring." *Journal of the American Statistical Association*, **89**(428), 1255–1270.

Hastie T, Tibshirani R, Friedman J (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2 edition.

Hawkins D (2004). "The Problem of Overfitting." *Journal of Chemical Information and Computer Sciences*, **44**(1), 1–12.

Hawkins D, Basak S, Mills D (2003). "Assessing Model Fit by Cross–Validation." *Journal of Chemical Information and Computer Sciences*, **43**(2), 579–586.

Henderson H, Velleman P (1981). "Building Multiple Regression Models Interactively." *Biometrics*, pp. 391–411.

Hesterberg T, Choi N, Meier L, Fraley C (2008). "Least Angle and L_1 Penalized Regression: A Review." *Statistics Surveys*, **2**, 61–93.

Heyman R, Slep A (2001). "The Hazards of Predicting Divorce Without Cross-validation." *Journal of Marriage and the Family*, **63**(2), 473.

Hill A, LaPan P, Li Y, Haney S (2007). "Impact of Image Segmentation on High–Content Screening Data Quality for SK–BR-3 Cells." *BMC Bioinformatics*, **8**(1), 340.

Ho T (1998). "The Random Subspace Method for Constructing Decision Forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 340–354.

Hoerl A (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, **12**(1), 55–67.

Holland J (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

Holland J (1992). *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.

Holmes G, Hall M, Frank E (1993). "Generating Rule Sets from Model Trees." In "Australian Joint Conference on Artificial Intelligence," .

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674.

Hothorn T, Leisch F, Zeileis A, Hornik K (2005). "The Design and Analysis of Benchmark Experiments." *Journal of Computational and Graphical Statistics*, **14**(3), 675–699.

Hsieh W, Tang B (1998). "Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography." *Bulletin of the American Meteorological Society*, **79**(9), 1855–1870.

Hsu C, Lin C (2002). "A Comparison of Methods for Multiclass Support Vector Machines." *IEEE Transactions on Neural Networks*, **13**(2), 415–425.

Huang C, Chang B, Cheng D, Chang C (2012). "Feature Selection and Parameter Optimization of a Fuzzy-Based Stock Selection Model Using Genetic Algorithms." *International Journal of Fuzzy Systems*, **14**(1), 65–75.

Huuskonen J (2000). "Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology." *Journal of Chemical Information and Computer Sciences*, **40**(3), 773–777.

Ihaka R, Gentleman R (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Jeatrakul P, Wong K, Fung C (2010). "Classification of Imbalanced Data By Combining the Complementary Neural Network and SMOTE Algorithm." *Neural Information Processing. Models and Applications*, pp. 152–159.

Jerez J, Molina I, Garcia-Laencina P, Alba R, Ribelles N, Martin M, Franco L (2010). "Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem." *Artificial Intelligence in Medicine*, **50**, 105–115.

John G, Kohavi R, Pfleger K (1994). "Irrelevant Features and the Subset Selection Problem." *Proceedings of the Eleventh International Conference on Machine Learning*, **129**, 121–129.

Johnson K, Rayens W (2007). "Modern Classification Methods for Drug Discovery." In A Dmitrienko, C Chuang-Stein, R D'Agostino (eds.), "Pharmaceutical Statistics Using SAS: A Practical Guide," pp. 7–43. Cary, NC: SAS Institute Inc.

Johnson R, Wichern D (2001). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Jolliffe I, Trendafilov N, Uddin M (2003). "A Modified Principal Component Technique Based on the lasso." *Journal of Computational and Graphical Statistics*, **12**(3), 531–547.

Kansy M, Senner F, Gubernator K (1998). "Physiochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes." *Journal of Medicinal Chemistry*, **41**, 1007–1010.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). "kernlab - An S4 Package for Kernel Methods in R." *Journal of Statistical Software*, **11**(9), 1–20.

Kearns M, Valiant L (1989). "Cryptographic Limitations on Learning Boolean Formulae and Finite Automata." In "Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing," .

Kim J, Basak J, Holtzman D (2009). "The Role of Apolipoprotein E in Alzheimer's Disease." *Neuron*, **63**(3), 287–303.

Kim JH (2009). "Estimating Classification Error Rate: Repeated Cross–Validation, Repeated Hold–Out and Bootstrap." *Computational Statistics & Data Analysis*, **53**(11), 3735–3745.

Kimball A (1957). "Errors of the Third Kind in Statistical Consulting." *Journal of the American Statistical Association*, **52**, 133–142.

Kira K, Rendell L (1992). "The Feature Selection Problem: Traditional Methods and a New Algorithm." *Proceedings of the National Conference on Artificial Intelligence*, pp. 129–129.

Kline DM, Berardi VL (2005). "Revisiting Squared–Error and Cross–Entropy Functions for Training Neural Network Classifiers." *Neural Computing and Applications*, **14**(4), 310–318.

Kohavi R (1995). "A Study of Cross–Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence*, **14**, 1137–1145.

Kohavi R (1996). "Scaling Up the Accuracy of Naive–Bayes Classifiers: A Decision–Tree Hybrid." In "Proceedings of the second international conference on knowledge discovery and data mining," volume 7.

Kohonen T (1995). *Self–Organizing Maps*. Springer.

Kononenko I (1994). "Estimating Attributes: Analysis and Extensions of Relief." In F Bergadano, L De Raedt (eds.), "Machine Learning: ECML–94," volume 784, pp. 171–182. Springer Berlin / Heidelberg.

Kuhn M (2008). "Building Predictive Models in R Using the `caret` Package." *Journal of Statistical Software*, **28**(5).

Kuhn M (2010). "The caret Package Homepage." URL http://caret.r-forge.r-project.org/.

Kuiper S (2008). "Introduction to Multiple Regression: How Much Is Your Car Worth?" *Journal of Statistics Education*, **16**(3).

Kvålseth T (1985). "Cautionary Note About R^2." *American Statistician*, **39**(4), 279–285.

Lachiche N, Flach P (2003). "Improving Accuracy and Cost of Two–Class and Multi–Class Probabilistic Classifiers using ROC Curves." In "Proceedings of the Twentieth International Conference on Machine Learning," volume 20, pp. 416–424.

Larose D (2006). *Data Mining Methods and Models*. Wiley.

Lavine B, Davidson C, Moores A (2002). "Innovative Genetic Algorithms for Chemoinformatics." *Chemometrics and Intelligent Laboratory Systems*, **60**(1), 161–171.

Leach A, Gillet V (2003). *An Introduction to Chemoinformatics*. Springer.

Leisch F (2002a). "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In W Härdle, B Rönz (eds.), "Compstat 2002 — Proceedings in Computational Statistics," pp. 575–580. Physica Verlag, Heidelberg.

Leisch F (2002b). "Sweave, Part I: Mixing R and LaTeX." *R News*, **2**(3), 28–31.

Levy S (2010). "The AI Revolution is On." *Wired*.

Li J, Fine JP (2008). "ROC Analysis with Multiple Classes and Multiple Tests: Methodology and Its Application in Microarray Studies." *Biostatistics*, **9**(3), 566–576.

Lindgren F, Geladi P, Wold S (1993). "The Kernel Algorithm for PLS." *Journal of Chemometrics*, **7**, 45–59.

Ling C, Li C (1998). "Data Mining for Direct Marketing: Problems and solutions." In "Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining," pp. 73–79.

Lipinski C, Lombardo F, Dominy B, Feeney P (1997). "Experimental and Computational Approaches To Estimate Solubility and Permeability In Drug Discovery and Development Settings." *Advanced Drug Delivery Reviews*, **23**, 3–25.

Liu B (2007). *Web Data Mining*. Springer Berlin / Heidelberg.

Liu Y, Rayens W (2007). "PLS and Dimension Reduction for Classification." *Computational Statistics*, pp. 189–208.

Lo V (2002). "The True Lift Model: A Novel Data Mining Approach To Response Modeling in Database Marketing." *ACM SIGKDD Explorations Newsletter*, **4**(2), 78–86.

Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C (2002). "Text Classification Using String Kernels." *The Journal of Machine Learning Research*, **2**, 419–444.

Loh WY (2002). "Regression Trees With Unbiased Variable Selection and Interaction Detection." *Statistica Sinica*, **12**, 361–386.

Loh WY (2010). "Tree–Structured Classifiers." *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 364–369.

Loh WY, Shih YS (1997). "Split Selection Methods for Classification Trees." *Statistica Sinica*, **7**, 815–840.

Mahé P, Ueda N, Akutsu T, Perret J, Vert J (2005). "Graph Kernels for Molecular Structure–Activity Relationship Analysis with Support Vector Machines." *Journal of Chemical Information and Modeling*, **45**(4), 939–951.

Mahé P, Vert J (2009). "Graph Kernels Based on Tree Patterns for Molecules." *Machine Learning*, **75**(1), 3–35.

Maindonald J, Braun J (2007). *Data Analysis and Graphics Using R*. Cambridge University Press, 2nd edition.

Mandal A, Johnson K, Wu C, Bornemeier D (2007). "Identifying Promising Compounds in Drug Discovery: Genetic Algorithms and Some New Statistical Techniques." *Journal of Chemical Information and Modeling*, **47**(3), 981–988.

Mandal A, Wu C, Johnson K (2006). "SELC: Sequential Elimination of Level Combinations by Means of Modified Genetic Algorithms." *Technometrics*, **48**(2), 273–283.

Martin J, Hirschberg D (1996). "Small Sample Statistics for Classification Error Rates I: Error Rate Measurements." *Department of Informatics and Computer Science Technical Report*.

Martin T, Harten P, Young D, Muratov E, Golbraikh A, Zhu H, Tropsha A (2012). "Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling?" *Journal of Chemical Information and Modeling*, **52**(10), 2570–2578.

Massy W (1965). "Principal Components Regression in Exploratory Statistical Research." *Journal of the American Statistical Association*, **60**, 234–246.

McCarren P, Springer C, Whitehead L (2011). "An Investigation into Pharmaceutically Relevant Mutagenicity Data and the Influence on Ames Predictive Potential." *Journal of Cheminformatics*, **3**(51).

McClish D (1989). "Analyzing a Portion of the ROC Curve." *Medical Decision Making*, **9**, 190–195.

Melssen W, Wehrens R, Buydens L (2006). "Supervised Kohonen Networks for Classification Problems." *Chemometrics and Intelligent Laboratory Systems*, **83**(2), 99–113.

Mente S, Lombardo F (2005). "A Recursive–Partitioning Model for Blood–Brain Barrier Permeation." *Journal of Computer–Aided Molecular Design*, **19**(7), 465–481.

Menze B, Kelm B, Splitthoff D, Koethe U, Hamprecht F (2011). "On Oblique Random Forests." *Machine Learning and Knowledge Discovery in Databases*, pp. 453–469.

Mevik B, Wehrens R (2007). "The pls Package: Principal Component and Partial Least Squares Regression in R." *Journal of Statistical Software*, **18**(2), 1–24.

Michailidis G, de Leeuw J (1998). "The Gifi System Of Descriptive Multivariate Analysis." *Statistical Science*, **13**, 307–336.

Milborrow S (2012). *Notes On the earth Package*. URL http://cran.r-project.org/package=earth.

Min S, Lee J, Han I (2006). "Hybrid Genetic Algorithms and Support Vector Machines for Bankruptcy Prediction." *Expert Systems with Applications*, **31**(3), 652–660.

Mitchell M (1998). *An Introduction to Genetic Algorithms*. MIT Press.

Molinaro A (2005). "Prediction Error Estimation: A Comparison of Resampling Methods." *Bioinformatics*, **21**(15), 3301–3307.

Molinaro A, Lostritto K, Van Der Laan M (2010). "partDSA: Deletion/Substitution/Addition Algorithm for Partitioning the Covariate Space in Prediction." *Bioinformatics*, **26**(10), 1357–1363.

Montgomery D, Runger G (1993). "Gauge Capability and Designed Experiments. Part I: Basic Methods." *Quality Engineering*, **6**(1), 115–135.

Muenchen R (2009). *R for SAS and SPSS Users*. Springer.

Myers R (1994). *Classical and Modern Regression with Applications*. PWS-KENT Publishing Company, Boston, MA, second edition.

Myers R, Montgomery D (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley, New York, NY.

Neal R (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag.

Nelder J, Mead R (1965). "A Simplex Method for Function Minimization." *The Computer Journal*, **7**(4), 308–313.

Netzeva T, Worth A, Aldenberg T, Benigni R, Cronin M, Gramatica P, Jaworska J, Kahn S, Klopman G, Marchant C (2005). "Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure–Activity Relationships." In "The Report and Recommendations of European Centre for the Validation of Alternative Methods Workshop 52," volume 33, pp. 1–19.

Niblett T (1987). "Constructing Decision Trees in Noisy Domains." In I Bratko, N Lavrač (eds.), "Progress in Machine Learning: Proceedings of EWSL–87," pp. 67–78. Sigma Press, Bled, Yugoslavia.

Olden J, Jackson D (2000). "Torturing Data for the Sake of Generality: How Valid Are Our Regression Models?" *Ecoscience*, **7**(4), 501–510.

Olsson D, Nelson L (1975). "The Nelder–Mead Simplex Procedure for Function Minimization." *Technometrics*, **17**(1), 45–51.

Osuna E, Freund R, Girosi F (1997). "Support Vector Machines: Training and Applications." *Technical report*, MIT Artificial Intelligence Laboratory.

Ozuysal M, Calonder M, Lepetit V, Fua P (2010). "Fast Keypoint Recognition Using Random Ferns." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(3), 448–461.

Park M, Hastie T (2008). "Penalized Logistic Regression for Detecting Gene Interactions." *Biostatistics*, **9**(1), 30.

Pepe MS, Longton G, Janes H (2009). "Estimation and Comparison of Receiver Operating Characteristic Curves." *Stata Journal*, **9**(1), 1–16.

Perrone M, Cooper L (1993). "When Networks Disagree: Ensemble Methods for Hybrid Neural Networks." In RJ Mammone (ed.), "Artificial Neural Networks for Speech and Vision," pp. 126–142. Chapman & Hall, London.

Piersma A, Genschow E, Verhoef A, Spanjersberg M, Brown N, Brady M, Burns A, Clemann N, Seiler A, Spielmann H (2004). "Validation of the Postimplantation Rat Whole-embryo Culture Test in the International ECVAM Validation Study on Three In Vitro Embryotoxicity Tests." *Alternatives to Laboratory Animals*, **32**, 275–307.

Platt J (2000). "Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods." In B Bartlett, B Schölkopf, D Schuurmans, A Smola (eds.), "Advances in Kernel Methods Support Vector Learning," pp. 61–74. Cambridge, MA: MIT Press.

Provost F, Domingos P (2003). "Tree Induction for Probability–Based Ranking." *Machine Learning*, **52**(3), 199–215.

Provost F, Fawcett T, Kohavi R (1998). "The Case Against Accuracy Estimation for Comparing Induction Algorithms." *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 445–453.

Quinlan R (1987). "Simplifying Decision Trees." *International Journal of Man–Machine Studies*, **27**(3), 221–234.

Quinlan R (1992). "Learning with Continuous Classes." *Proceedings of the 5th Australian Joint Conference On Artificial Intelligence*, pp. 343–348.

Quinlan R (1993a). "Combining Instance–Based and Model–Based Learning." *Proceedings of the Tenth International Conference on Machine Learning*, pp. 236–243.

Quinlan R (1993b). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Quinlan R (1996a). "Bagging, Boosting, and C4.5." In "In Proceedings of the Thirteenth National Conference on Artificial Intelligence," .

Quinlan R (1996b). "Improved use of continuous attributes in C4.5." *Journal of Artificial Intelligence Research*, **4**, 77–90.

Quinlan R, Rivest R (1989). "Inferring Decision Trees Using the Minimum Description Length Principle." *Information and computation*, **80**(3), 227–248.

Radcliffe N, Surry P (2011). "Real–World Uplift Modelling With Significance–Based Uplift Trees." *Technical report*, Stochastic Solutions.

Rännar S, Lindgren F, Geladi P, Wold S (1994). "A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects. Part 1: Theory and Algorithm." *Journal of Chemometrics*, **8**, 111–125.

R Development Core Team (2008). *R: Regulatory Compliance and Validation Issues A Guidance Document for the Use of R in Regulated Clinical Trial Environments*. R Foundation for Statistical Computing, Vienna, Austria.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reshef D, Reshef Y, Finucane H, Grossman S, McVean G, Turnbaugh P, Lander E, Mitzenmacher M, Sabeti P (2011). "Detecting Novel Associations in Large Data Sets." *Science*, **334**(6062), 1518–1524.

Richardson M, Dominowska E, Ragno R (2007). "Predicting Clicks: Estimating the Click–Through Rate for New Ads." In "Proceedings of the 16^{th} International Conference on the World Wide Web," pp. 521–530.

Ridgeway G (2007). "Generalized Boosted Models: A Guide to the gbm Package." URL http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

Ripley B (1995). "Statistical Ideas for Selecting Network Architectures." *Neural Networks: Artificial Intelligence and Industrial Applications*, pp. 183–190.

Ripley B (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." *BMC Bioinformatics*, **12**(1), 77.

Robnik-Sikonja M, Kononenko I (1997). "An Adaptation of Relief for Attribute Estimation in Regression." *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 296–304.

Rodriguez M (2011). "The Failure of Predictive Modeling and Why We Follow the Herd." *Technical report*, Concepcion, Martinez & Bellido.

Ruczinski I, Kooperberg C, Leblanc M (2003). "Logic Regression." *Journal of Computational and Graphical Statistics*, **12**(3), 475–511.

Rumelhart D, Hinton G, Williams R (1986). "Learning Internal Representations by Error Propagation." In "Parallel Distributed Processing: Explorations in the Microstructure of Cognition," The MIT Press.

Rzepakowski P, Jaroszewicz S (2012). "Uplift Modeling in Direct Marketing." *Journal of Telecommunications and Information Technology*, **2**, 43–50.

Saar-Tsechansky M, Provost F (2007a). "Decision–Centric Active Learning of Binary–Outcome Models." *Information Systems Research*, **18**(1), 4–22.

Saar-Tsechansky M, Provost F (2007b). "Handling Missing Values When Applying Classification Models." *Journal of Machine Learning Research*, **8**, 1625–1657.

Saeys Y, Inza I, Larranaga P (2007). "A Review of Feature Selection Techniques in Bioinformatics." *Bioinformatics*, **23**(19), 2507–2517.

Schapire R (1990). "The Strength of Weak Learnability." *Machine Learning*, **45**, 197–227.

Schapire YFR (1999). "Adaptive Game Playing Using Multiplicative Weights." *Games and Economic Behavior*, **29**, 79–103.

Schmidberger M, Morgan M, Eddelbuettel D, Yu H, Tierney L, Mansmann U (2009). "State–of–the–Art in Parallel Computing with R." *Journal of Statistical Software*, **31**(1).

Serneels S, Nolf ED, Espen PV (2006). "Spatial Sign Pre-processing: A Simple Way to Impart Moderate Robustness to Multivariate Estimators." *Journal of Chemical Information and Modeling*, **46**(3), 1402–1409.

Shachtman N (2011). "Pentagon's Prediction Software Didn't Spot Egypt Unrest." *Wired*.

Shannon C (1948). "A Mathematical Theory of Communication." *The Bell System Technical Journal*, **27**(3), 379–423.

Siegel E (2011). "Uplift Modeling: Predictive Analytics Can't Optimize Marketing Decisions Without It." *Technical report*, Prediction Impact Inc.

Simon R, Radmacher M, Dobbin K, McShane L (2003). "Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification." *Journal of the National Cancer Institute*, **95**(1), 14–18.

Smola A (1996). "Regression Estimation with Support Vector Learning Machines." *Master's thesis, Technische Universit at Munchen*.

Spector P (2008). *Data Manipulation with R*. Springer.

Steyerberg E (2010). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, 1st ed. softcover of orig. ed. 2009 edition.

Stone M, Brooks R (1990). "Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares, and Principal Component Regression." *Journal of the Royal Statistical Society, Series B*, **52**, 237–269.

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics*, **8**(1), 25.

Suykens J, Vandewalle J (1999). "Least Squares Support Vector Machine Classifiers." *Neural processing letters*, **9**(3), 293–300.

Tetko I, Tanchuk V, Kasheva T, Villa A (2001). "Estimation of Aqueous Solubility of Chemical Compounds Using E–State Indices." *Journal of Chemical Information and Computer Sciences*, **41**(6), 1488–1493.

Tibshirani R (1996). "Regression Shrinkage and Selection via the lasso." *Journal of the Royal Statistical Society Series B (Methodological)*, **58**(1), 267–288.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002). "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression." *Proceedings of the National Academy of Sciences*, **99**(10), 6567–6572.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2003). "Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays." *Statistical Science*, **18**(1), 104–117.

Ting K (2002). "An Instance–Weighting Method to Induce Cost–Sensitive Trees." *IEEE Transactions on Knowledge and Data Engineering*, **14**(3), 659–665.

Tipping M (2001). "Sparse Bayesian Learning and the Relevance Vector Machine." *Journal of Machine Learning Research*, **1**, 211–244.

Titterington M (2010). "Neural Networks." *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(1), 1–8.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R (2001). "Missing Value Estimation Methods for DNA Microarrays." *Bioinformatics*, **17**(6), 520–525.

Tumer K, Ghosh J (1996). "Analysis of Decision Boundaries in Linearly Combined Neural Classifiers." *Pattern Recognition*, **29**(2), 341–348.

US Commodity Futures Trading Commission and US Securities & Exchange Commission (2010). *Findings Regarding the Market Events of May 6, 2010*.

Valiant L (1984). "A Theory of the Learnable." *Communications of the ACM*, **27**, 1134–1142.

Van Der Putten P, Van Someren M (2004). "A Bias–Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000." *Machine Learning*, **57**(1), 177–195.

Van Hulse J, Khoshgoftaar T, Napolitano A (2007). "Experimental Perspectives On Learning From Imbalanced Data." In "Proceedings of the 24^{th} International Conference On Machine learning," pp. 935–942.

Vapnik V (2010). *The Nature of Statistical Learning Theory*. Springer.

Varma S, Simon R (2006). "Bias in Error Estimation When Using Cross–Validation for Model Selection." *BMC Bioinformatics*, **7**(1), 91.

Varmuza K, He P, Fang K (2003). "Boosting Applied to Classification of Mass Spectral Data." *Journal of Data Science*, **1**, 391–404.

Venables W, Ripley B (2002). *Modern Applied Statistics with S*. Springer.

Venables W, Smith D, the R Development Core Team (2003). *An Introduction to R*. R Foundation for Statistical Computing, Vienna, Austria, version 1.6.2 edition. ISBN 3-901167-55-2, URL http://www.R-project.org.

Venkatraman E (2000). "A Permutation Test to Compare Receiver Operating Characteristic Curves." *Biometrics*, **56**(4), 1134–1138.

Veropoulos K, Campbell C, Cristianini N (1999). "Controlling the Sensitivity of Support Vector Machines." *Proceedings of the International Joint Conference on Artificial Intelligence*, **1999**, 55–60.

Verzani J (2002). "simpleR – Using R for Introductory Statistics." URL http://www.math.csi.cuny.edu/Statistics/R/simpleR.

Wager TT, Hou X, Verhoest PR, Villalobos A (2010). "Moving Beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach To Enable Alignment of Druglike Properties." *ACS Chemical Neuroscience*, **1**(6), 435–449.

Wallace C (2005). *Statistical and Inductive Inference by Minimum Message Length*. Springer–Verlag.

Wang C, Venkatesh S (1984). "Optimal Stopping and Effective Machine Complexity in Learning." *Advances in NIPS*, pp. 303–310.

Wang Y, Witten I (1997). "Inducing Model Trees for Continuous Classes." *Proceedings of the Ninth European Conference on Machine Learning*, pp. 128–137.

Weiss G, Provost F (2001a). "The Effect of Class Distribution on Classifier Learning: An Empirical Study." *Department of Computer Science, Rutgers University*.

Weiss G, Provost F (2001b). "The Effect of Class Distribution On Classifier Learning: An Empirical Study." *Technical Report ML-TR-44*, Department of Computer Science, Rutgers University.

Welch B (1939). "Note on Discriminant Functions." *Biometrika*, **31**, 218–220.

Westfall P, Young S (1993). *Resampling–Based Multiple Testing: Examples and Methods for P–Value Adjustment*. Wiley.

Westphal C (2008). *Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies*. CRC Press.

Whittingham M, Stephens P, Bradbury R, Freckleton R (2006). "Why Do We Still Use Stepwise Modelling in Ecology and Behaviour?" *Journal of Animal Ecology*, **75**(5), 1182–1189.

Willett P (1999). "Dissimilarity–Based Algorithms for Selecting Structurally Diverse Sets of Compounds." *Journal of Computational Biology*, **6**(3), 447–457.

Williams G (2011). *Data Mining with Rattle and R : The Art of Excavating Data for Knowledge Discovery*. Springer.

Witten D, Tibshirani R (2009). "Covariance–Regularized Regression and Classification For High Dimensional Problems." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **71**(3), 615–636.

Witten D, Tibshirani R (2011). "Penalized Classification Using Fisher's Linear Discriminant." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **73**(5), 753–772.

Wold H (1966). "Estimation of Principal Components and Related Models by Iterative Least Squares." In P Krishnaiah (ed.), "Multivariate Analyses," pp. 391–420. Academic Press, New York.

Wold H (1982). "Soft Modeling: The Basic Design and Some Extensions." In K Joreskog, H Wold (eds.), "Systems Under Indirect Observation: Causality, Structure, Prediction," pt. 2, pp. 1–54. North–Holland, Amsterdam.

Wold S (1995). "PLS for Multivariate Linear Modeling." In H van de Waterbeemd (ed.), "Chemometric Methods in Molecular Design," pp. 195–218. VCH, Weinheim.

Wold S, Johansson M, Cocchi M (1993). "PLS–Partial Least-Squares Projections to Latent Structures." In H Kubinyi (ed.), "3D QSAR in Drug Design," volume 1, pp. 523–550. Kluwer Academic Publishers, The Netherlands.

Wold S, Martens H, Wold H (1983). "The Multivariate Calibration Problem in Chemistry Solved by the PLS Method." In "Proceedings from the Conference on Matrix Pencils," Springer–Verlag, Heidelberg.

Wolpert D (1996). "The Lack of *a priori* Distinctions Between Learning Algorithms." *Neural Computation*, **8**(7), 1341–1390.

Yeh I (1998). "Modeling of Strength of High-Performance Concrete Using Artificial Neural Networks." *Cement and Concrete research*, **28**(12), 1797–1808.

Yeh I (2006). "Analysis of Strength of Concrete Using Design of Experiments and Neural Networks." *Journal of Materials in Civil Engineering*, **18**, 597–604.

Youden W (1950). "Index for Rating Diagnostic Tests." *Cancer*, **3**(1), 32–35.

Zadrozny B, Elkan C (2001). "Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers." In "Proceedings of the 18th International Conference on Machine Learning," pp. 609–616. Morgan Kaufmann.

Zeileis A, Hothorn T, Hornik K (2008). "Model–Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514.

Zhu J, Hastie T (2005). "Kernel Logistic Regression and the Import Vector Machine." *Journal of Computational and Graphical Statistics*, **14**(1), 185–205.

Zou H, Hastie T (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B*, **67**(2), 301–320.

Zou H, Hastie T, Tibshirani R (2004). "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics*, **15**, 2006.

Indicies

Computing

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3,

General

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3,