

Table of Contents

INTRODUCTION	1
About This Book	1
Foolish Assumptions	2
Icons Used in This Book	3
Beyond the Book	4
Where to Go from Here	5
PART 1: INTRODUCING HOW MACHINES LEARN	7
CHAPTER 1: Getting the Real Story about AI	9
Moving beyond the Hype	10
Dreaming of Electric Sheep	11
Understanding the history of AI and machine learning	12
Exploring what machine learning can do for AI	13
Considering the goals of machine learning	13
Defining machine learning limits based on hardware	14
Overcoming AI Fantasies	15
Discovering the fad uses of AI and machine learning	16
Considering the true uses of AI and machine learning	16
Being useful; being mundane	18
Considering the Relationship between AI and Machine Learning	19
Considering AI and Machine Learning Specifications	20
Defining the Divide between Art and Engineering	20
CHAPTER 2: Learning in the Age of Big Data	23
Defining Big Data	24
Considering the Sources of Big Data	25
Building a new data source	26
Using existing data sources	27
Locating test data sources	28
Specifying the Role of Statistics in Machine Learning	29
Understanding the Role of Algorithms	30
Defining what algorithms do	30
Considering the five main techniques	30
Defining What Training Means	32
CHAPTER 3: Having a Glance at the Future	35
Creating Useful Technologies for the Future	36
Considering the role of machine learning in robots	36
Using machine learning in health care	37
Creating smart systems for various needs	37

Using machine learning in industrial settings	38
Understanding the role of updated processors and other hardware	39
Discovering the New Work Opportunities with Machine Learning	39
Working for a machine	40
Working with machines	41
Repairing machines	41
Creating new machine learning tasks	42
Devising new machine learning environments	42
Avoiding the Potential Pitfalls of Future Technologies	43

PART 2: PREPARING YOUR LEARNING TOOLS..... 45

CHAPTER 4: Installing an R Distribution 47

Choosing an R Distribution with Machine Learning in Mind	48
Installing R on Windows	49
Installing R on Linux	56
Installing R on Mac OS X	57
Downloading the Datasets and Example Code	59
Understanding the datasets used in this book	59
Defining the code repository	60

CHAPTER 5: Coding in R Using RStudio 63

Understanding the Basic Data Types	64
Working with Vectors	66
Organizing Data Using Lists	66
Working with Matrices	67
Creating a basic matrix	68
Changing the vector arrangement	69
Accessing individual elements	69
Naming the rows and columns	70
Interacting with Multiple Dimensions Using Arrays	71
Creating a basic array	71
Naming the rows and columns	72
Creating a Data Frame	74
Understanding factors	74
Creating a basic data frame	76
Interacting with data frames	77
Expanding a data frame	79
Performing Basic Statistical Tasks	80
Making decisions	80
Working with loops	82

Performing looped tasks without loops.....	84
Working with functions.....	85
Finding mean and median.....	85
Charting your data.....	87
CHAPTER 6: Installing a Python Distribution	89
Choosing a Python Distribution with Machine Learning in Mind.....	90
Getting Continuum Analytics Anaconda	91
Getting Enthought Canopy Express	92
Getting pythonxy.....	93
Getting WinPython	93
Installing Python on Linux	93
Installing Python on Mac OS X.....	94
Installing Python on Windows.....	96
Downloading the Datasets and Example Code.....	99
Using Jupyter Notebook	100
Defining the code repository.....	101
Understanding the datasets used in this book.....	106
CHAPTER 7: Coding in Python Using Anaconda.....	109
Working with Numbers and Logic.....	110
Performing variable assignments.....	112
Doing arithmetic	113
Comparing data using Boolean expressions.....	115
Creating and Using Strings.....	117
Interacting with Dates.....	118
Creating and Using Functions	119
Creating reusable functions.....	119
Calling functions	121
Working with global and local variables.....	123
Using Conditional and Loop Statements	124
Making decisions using the if statement	124
Choosing between multiple options using nested decisions	125
Performing repetitive tasks using for.....	126
Using the while statement.....	127
Storing Data Using Sets, Lists, and Tuples.....	128
Creating sets.....	128
Performing operations on sets	128
Creating lists.....	129
Creating and using tuples	131
Defining Useful Iterators	132
Indexing Data Using Dictionaries	134
Storing Code in Modules	134

CHAPTER 8:	Exploring Other Machine Learning Tools	137
	Meeting the Precursors SAS, Stata, and SPSS	138
	Learning in Academia with Weka	140
	Accessing Complex Algorithms Easily Using LIBSVM	141
	Running As Fast As Light with Vowpal Wabbit	142
	Visualizing with Knime and RapidMiner	143
	Dealing with Massive Data by Using Spark	144

PART 3: GETTING STARTED WITH THE MATH BASICS..... 145

CHAPTER 9:	Demystifying the Math Behind Machine Learning	147
	Working with Data	148
	Creating a matrix	150
	Understanding basic operations	152
	Performing matrix multiplication	152
	Glancing at advanced matrix operations	155
	Using vectorization effectively	155
	Exploring the World of Probabilities	158
	Operating on probabilities	159
	Conditioning chance by Bayes' theorem	160
	Describing the Use of Statistics	163
CHAPTER 10:	Descending the Right Curve	167
	Interpreting Learning As Optimization	168
	Supervised learning	168
	Unsupervised learning	169
	Reinforcement learning	169
	The learning process	170
	Exploring Cost Functions	173
	Descending the Error Curve	174
	Updating by Mini-Batch and Online	177
CHAPTER 11:	Validating Machine Learning	181
	Checking Out-of-Sample Errors	182
	Looking for generalization	183
	Getting to Know the Limits of Bias	184
	Keeping Model Complexity in Mind	186
	Keeping Solutions Balanced	188
	Depicting learning curves	189
	Training, Validating, and Testing	191
	Resorting to Cross-Validation	191
	Looking for Alternatives in Validation	193

Optimizing Cross-Validation Choices	194
Exploring the space of hyper-parameters	195
Avoiding Sample Bias and Leakage Traps	196
Watching out for snooping.	198
CHAPTER 12: Starting with Simple Learners.....	199
Discovering the Incredible Perceptron.....	200
Falling short of a miracle	200
Touching the nonseparability limit.....	202
Growing Greedy Classification Trees	204
Predicting outcomes by splitting data	204
Pruning overgrown trees	208
Taking a Probabilistic Turn.....	209
Understanding Naïve Bayes.....	209
Estimating response with Naïve Bayes.....	212
PART 4: LEARNING FROM SMART AND BIG DATA	217
CHAPTER 13: Preprocessing Data.....	219
Gathering and Cleaning Data	220
Repairing Missing Data.....	221
Identifying missing data	221
Choosing the right replacement strategy	222
Transforming Distributions	225
Creating Your Own Features	227
Understanding the need to create features	227
Creating features automatically	228
Compressing Data.....	230
Delimiting Anomalous Data.....	232
CHAPTER 14: Leveraging Similarity	237
Measuring Similarity between Vectors.....	238
Understanding similarity	238
Computing distances for learning.....	239
Using Distances to Locate Clusters.....	240
Checking assumptions and expectations.....	241
Inspecting the gears of the algorithm	243
Tuning the K-Means Algorithm	244
Experimenting K-means reliability	245
Experimenting with how centroids converge	247
Searching for Classification by K-Nearest Neighbors.....	251
Leveraging the Correct K Parameter	252
Understanding the k parameter	252
Experimenting with a flexible algorithm	253

CHAPTER 15:	Working with Linear Models the Easy Way	257
	Starting to Combine Variables	258
	Mixing Variables of Different Types	264
	Switching to Probabilities	267
	Specifying a binary response	267
	Handling multiple classes	270
	Guessing the Right Features	271
	Defining the outcome of features that don't work together	271
	Solving overfitting by using selection	272
	Learning One Example at a Time	274
	Using gradient descent	275
	Understanding how SGD is different	275
CHAPTER 16:	Hitting Complexity with Neural Networks	279
	Learning and Imitating from Nature	280
	Going forth with feed-forward	281
	Going even deeper down the rabbit hole	283
	Getting Back with Backpropagation	286
	Struggling with Overfitting	289
	Understanding the problem	289
	Opening the black box	290
	Introducing Deep Learning	293
CHAPTER 17:	Going a Step beyond Using Support Vector Machines	297
	Revisiting the Separation Problem: A New Approach	298
	Explaining the Algorithm	299
	Getting into the math of an SVM	301
	Avoiding the pitfalls of nonseparability	302
	Applying Nonlinearity	303
	Demonstrating the kernel trick by example	305
	Discovering the different kernels	306
	Illustrating Hyper-Parameters	308
	Classifying and Estimating with SVM	309
CHAPTER 18:	Resorting to Ensembles of Learners	315
	Leveraging Decision Trees	316
	Growing a forest of trees	317
	Understanding the importance measures	321
	Working with Almost Random Guesses	324
	Bagging predictors with Adaboost	324
	Boosting Smart Predictors	327
	Meeting again with gradient descent	328
	Averaging Different Predictors	329

PART 5: APPLYING LEARNING TO REAL PROBLEMS..... 331

CHAPTER 19: Classifying Images..... 333

Working with a Set of Images	334
Extracting Visual Features	338
Recognizing Faces Using Eigenfaces.....	340
Classifying Images.....	343

CHAPTER 20: Scoring Opinions and Sentiments..... 349

Introducing Natural Language Processing.....	349
Understanding How Machines Read	350
Processing and enhancing text	352
Scraping textual datasets from the web	357
Handling problems with raw text	360
Using Scoring and Classification	362
Performing classification tasks	362
Analyzing reviews from e-commerce	365

CHAPTER 21: Recommending Products and Movies..... 369

Realizing the Revolution.....	370
Downloading Rating Data.....	371
Trudging through the MovieLens dataset	371
Navigating through anonymous web data	373
Encountering the limits of rating data	374
Leveraging SVD	375
Considering the origins of SVD	376
Understanding the SVD connection	377
Seeing SVD in action	378

PART 6: THE PART OF TENS..... 383

CHAPTER 22: Ten Machine Learning Packages to Master..... 385

Cloudera Oryx	386
CUDA-Convnet.....	386
ConvNetJS.....	387
e1071.....	387
gbm	388
Gensim	388
glmnet.....	388
randomForest	389
SciPy	389
XGBoost	390

CHAPTER 23: Ten Ways to Improve Your Machine Learning Models	391
Studying Learning Curves	392
Using Cross-Validation Correctly	393
Choosing the Right Error or Score Metric	394
Searching for the Best Hyper-Parameters	395
Testing Multiple Models	395
Averaging Models	396
Stacking Models	396
Applying Feature Engineering	397
Selecting Features and Examples	397
Looking for More Data	398
INDEX	399