

# BRIEF CONTENTS

Acknowledgments . . . . .	xxi
Introduction . . . . .	xxiii

## **PART I: FOUNDATIONAL IDEAS . . . . . 1**

Chapter 1: An Overview of Machine Learning . . . . .	3
Chapter 2: Essential Statistics . . . . .	15
Chapter 3: Measuring Performance . . . . .	47
Chapter 4: Bayes' Rule . . . . .	83
Chapter 5: Curves and Surfaces . . . . .	117
Chapter 6: Information Theory . . . . .	133

## **PART II: BASIC MACHINE LEARNING . . . . . 153**

Chapter 7: Classification . . . . .	155
Chapter 8: Training and Testing . . . . .	181
Chapter 9: Overfitting and Underfitting . . . . .	195
Chapter 10: Data Preparation . . . . .	221
Chapter 11: Classifiers . . . . .	263
Chapter 12: Ensembles . . . . .	297

## **PART III: DEEP LEARNING BASICS . . . . . 311**

Chapter 13: Neural Networks . . . . .	313
Chapter 14: Backpropagation . . . . .	351
Chapter 15: Optimizers . . . . .	387

<b>PART IV: BEYOND THE BASICS</b> . . . . .	<b>427</b>
Chapter 16: Convolutional Neural Networks. . . . .	429
Chapter 17: Convnets in Practice. . . . .	473
Chapter 18: Autoencoders . . . . .	495
Chapter 19: Recurrent Neural Networks . . . . .	539
Chapter 20: Attention and Transformers . . . . .	565
Chapter 21: Reinforcement Learning . . . . .	601
Chapter 22: Generative Adversarial Networks . . . . .	649
Chapter 23: Creative Applications. . . . .	675
References. . . . .	693
Image Credits . . . . .	717
Index . . . . .	721

# CONTENTS IN DETAIL

## ACKNOWLEDGMENTS

xxi

## INTRODUCTION

xxiii

Who This Book Is For . . . . .	xxiv
This Book Has No Math and No Code . . . . .	xxv
There Is Code, If You Want It . . . . .	xxv
The Figures Are Available, Too! . . . . .	xxv
Errata . . . . .	xxvi
About This Book . . . . .	xxvi
Part I: Foundational Ideas . . . . .	xxvi
Part II: Basic Machine Learning . . . . .	xxvii
Part III: Deep Learning Basics . . . . .	xxvii
Part IV: Deep Beyond the Basics . . . . .	xxvii
Final Words . . . . .	xxviii

## PART I: FOUNDATIONAL IDEAS

1

### 1 AN OVERVIEW OF MACHINE LEARNING

3

Expert Systems . . . . .	4
Supervised Learning . . . . .	6
Unsupervised Learning . . . . .	8
Reinforcement Learning . . . . .	9
Deep Learning . . . . .	10
Summary . . . . .	13

### 2 ESSENTIAL STATISTICS

15

Describing Randomness . . . . .	16
Random Variables and Probability Distributions . . . . .	17
Some Common Distributions . . . . .	21
Continuous Distributions . . . . .	21
Discrete Distributions . . . . .	26
Collections of Random Values . . . . .	28
Expected Value . . . . .	28
Dependence . . . . .	29
Independent and Identically Distributed Variables . . . . .	29
Sampling and Replacement . . . . .	29
Selection with Replacement . . . . .	30
Selection Without Replacement . . . . .	30
Bootstrapping . . . . .	31
Covariance and Correlation . . . . .	35
Covariance . . . . .	36
Correlation . . . . .	37

Statistics Don't Tell Us Everything . . . . .	40
High-Dimensional Spaces . . . . .	42
Summary . . . . .	44

### 3

## MEASURING PERFORMANCE

47

Different Types of Probability . . . . .	48
Dart Throwing . . . . .	48
Simple Probability . . . . .	50
Conditional Probability . . . . .	50
Joint Probability . . . . .	53
Marginal Probability . . . . .	55
Measuring Correctness . . . . .	56
Classifying Samples . . . . .	57
The Confusion Matrix . . . . .	60
Characterizing Incorrect Predictions . . . . .	61
Measuring Correct and Incorrect. . . . .	63
Accuracy . . . . .	64
Precision . . . . .	65
Recall . . . . .	66
Precision-Recall Tradeoff . . . . .	67
Misleading Measures . . . . .	69
f1 Score. . . . .	71
About These Terms . . . . .	72
Other Measures . . . . .	72
Constructing a Confusion Matrix Correctly . . . . .	74
Summary . . . . .	81

### 4

## BAYES' RULE

83

Frequentist and Bayesian Probability . . . . .	84
The Frequentist Approach. . . . .	84
The Bayesian Approach. . . . .	85
Frequentists vs. Bayesians . . . . .	85
Frequentist Coin Flipping. . . . .	86
Bayesian Coin Flipping. . . . .	87
A Motivating Example . . . . .	87
Picturing the Coin Probabilities . . . . .	88
Expressing Coin Flips as Probabilities . . . . .	90
Bayes' Rule . . . . .	94
Discussion of Bayes' Rule . . . . .	95
Bayes' Rule and Confusion Matrices. . . . .	97
Repeating Bayes' Rule . . . . .	101
The Posterior-Prior Loop . . . . .	102
The Bayes Loop in Action. . . . .	103
Multiple Hypotheses . . . . .	109
Summary . . . . .	115

## 5 CURVES AND SURFACES

117

The Nature of Functions . . . . .	118
The Derivative . . . . .	119
Maximums and Minimums . . . . .	119
Tangent Lines . . . . .	122
Finding Minimums and Maximums with Derivatives . . . . .	125
The Gradient . . . . .	126
Water, Gravity, and the Gradient . . . . .	127
Finding Maximums and Minimums with Gradients . . . . .	128
Saddle Points . . . . .	130
Summary . . . . .	131

## 6 INFORMATION THEORY

133

Surprise and Context . . . . .	134
Understanding Surprise . . . . .	134
Unpacking Context . . . . .	135
Measuring Information . . . . .	136
Adaptive Codes . . . . .	137
Speaking Morse . . . . .	138
Customizing Morse Code . . . . .	141
Entropy . . . . .	143
Cross Entropy . . . . .	145
Two Adaptive Codes . . . . .	146
Using the Codes . . . . .	148
Cross Entropy in Practice . . . . .	150
Kullback–Leibler Divergence . . . . .	151
Summary . . . . .	152

## PART II: BASIC MACHINE LEARNING

153

### 7 CLASSIFICATION

155

Two-Dimensional Binary Classification . . . . .	156
2D Multiclass Classification . . . . .	160
Multiclass Classification . . . . .	161
One-Versus-Rest . . . . .	161
One-Versus-One . . . . .	163
Clustering . . . . .	166
The Curse of Dimensionality . . . . .	168
Dimensionality and Density . . . . .	169
High-Dimensional Weirdness . . . . .	175
Summary . . . . .	179

## 8

### TRAINING AND TESTING

181

Training . . . . .	182
Testing the Performance . . . . .	183
Test Data . . . . .	186
Validation Data . . . . .	187
Cross-Validation . . . . .	190
k-Fold Cross-Validation . . . . .	192
Summary . . . . .	194

## 9

### OVERFITTING AND UNDERFITTING

195

Finding a Good Fit . . . . .	196
Overfitting . . . . .	196
Underfitting . . . . .	197
Detecting and Addressing Overfitting . . . . .	197
Early Stopping . . . . .	202
Regularization . . . . .	203
Bias and Variance . . . . .	204
Matching the Underlying Data . . . . .	205
High Bias, Low Variance . . . . .	207
Low Bias, High Variance . . . . .	209
Comparing Curves . . . . .	210
Fitting a Line with Bayes' Rule . . . . .	212
Summary . . . . .	219

## 10

### DATA PREPARATION

221

Basic Data Cleaning . . . . .	222
The Importance of Consistency . . . . .	223
Types of Data . . . . .	225
One-Hot Encoding . . . . .	226
Normalizing and Standardizing . . . . .	227
Normalization . . . . .	228
Standardization . . . . .	229
Remembering the Transformation . . . . .	230
Types of Transformations . . . . .	231
Slice Processing . . . . .	232
Samplewise Processing . . . . .	232
Featurewise Processing . . . . .	233
Elementwise Processing . . . . .	234
Inverse Transformations . . . . .	234
Information Leakage in Cross-Validation . . . . .	239
Shrinking the Dataset . . . . .	242
Feature Selection . . . . .	243
Dimensionality Reduction . . . . .	243
Principal Component Analysis . . . . .	244
PCA for Simple Images . . . . .	250
PCA for Real Images . . . . .	255
Summary . . . . .	260

## 11 CLASSIFIERS

263

Types of Classifiers . . . . .	264
k-Nearest Neighbors . . . . .	264
Decision Trees . . . . .	269
Introduction to Trees . . . . .	269
Using Decision Trees . . . . .	271
Overfitting Trees . . . . .	275
Splitting Nodes . . . . .	280
Support Vector Machines . . . . .	282
The Basic Algorithm . . . . .	282
The SVM Kernel Trick . . . . .	287
Naive Bayes . . . . .	290
Comparing Classifiers . . . . .	295
Summary . . . . .	296

## 12 ENSEMBLES

297

Voting . . . . .	298
Ensembles of Decision Trees . . . . .	299
Bagging . . . . .	299
Random Forests . . . . .	301
Extra Trees . . . . .	302
Boosting . . . . .	302
Summary . . . . .	309

## PART III: DEEP LEARNING BASICS

311

### 13 NEURAL NETWORKS

313

Real Neurons . . . . .	314
Artificial Neurons . . . . .	315
The Perceptron . . . . .	315
Modern Artificial Neurons . . . . .	317
Drawing the Neurons . . . . .	319
Feed-Forward Networks . . . . .	322
Neural Network Graphs . . . . .	323
Initializing the Weights . . . . .	325
Deep Networks . . . . .	326
Fully Connected Layers . . . . .	328
Tensors . . . . .	328
Preventing Network Collapse . . . . .	329
Activation Functions . . . . .	331
Straight-Line Functions . . . . .	331
Step Functions . . . . .	333
Piecewise Linear Functions . . . . .	336
Smooth Functions . . . . .	339

Activation Function Gallery . . . . .	344
Comparing Activation Functions . . . . .	344
Softmax . . . . .	345
Summary . . . . .	348

## **14**

### **BACKPROPAGATION** **351**

A High-Level Overview of Training . . . . .	352
Punishing Error . . . . .	352
A Slow Way to Learn . . . . .	354
Gradient Descent . . . . .	355
Getting Started . . . . .	356
Backprop on a Tiny Neural Network . . . . .	358
Finding Deltas for the Output Neurons . . . . .	360
Using Deltas to Change Weights . . . . .	366
Other Neuron Deltas . . . . .	368
Backprop on a Larger Network . . . . .	372
The Learning Rate . . . . .	376
Building a Binary Classifier . . . . .	378
Picking a Learning Rate . . . . .	379
An Even Smaller Learning Rate . . . . .	383
Summary . . . . .	386

## **15**

### **OPTIMIZERS** **387**

Error as a 2D Curve . . . . .	388
Adjusting the Learning Rate . . . . .	389
Constant-Sized Updates . . . . .	391
Changing the Learning Rate over Time . . . . .	396
Decay Schedules . . . . .	398
Updating Strategies . . . . .	400
Batch Gradient Descent . . . . .	401
Stochastic Gradient Descent . . . . .	403
Mini-Batch Gradient Descent . . . . .	405
Gradient Descent Variations . . . . .	407
Momentum . . . . .	408
Nesterov Momentum . . . . .	414
Adagrad . . . . .	417
Adadelta and RMSprop . . . . .	418
Adam . . . . .	420
Choosing an Optimizer . . . . .	421
Regularization . . . . .	422
Dropout . . . . .	422
Batchnorm . . . . .	424
Summary . . . . .	425



<b>16</b>	<b>CONVOLUTIONAL NEURAL NETWORKS</b>	<b>429</b>
Introducing Convolution . . . . .		430
Detecting Yellow . . . . .		431
Weight Sharing . . . . .		433
Larger Filters . . . . .		434
Filters and Features . . . . .		437
Padding . . . . .		440
Multidimensional Convolution . . . . .		443
Multiple Filters . . . . .		444
Convolution Layers . . . . .		446
1D Convolution . . . . .		446
1×1 Convolutions . . . . .		447
Changing Output Size . . . . .		449
Pooling . . . . .		449
Striding . . . . .		453
Transposed Convolution . . . . .		457
Hierarchies of Filters . . . . .		461
Simplifying Assumptions . . . . .		461
Finding Face Masks . . . . .		462
Finding Eyes, Noses, and Mouths . . . . .		465
Applying Our Filters . . . . .		467
Summary . . . . .		472
<b>17</b>	<b>CONVNETS IN PRACTICE</b>	<b>473</b>
Categorizing Handwritten Digits . . . . .		473
VGG16 . . . . .		478
Visualizing Filters, Part 1 . . . . .		481
Visualizing Filters, Part 2 . . . . .		487
Adversaries . . . . .		491
Summary . . . . .		493
<b>18</b>	<b>AUTOENCODERS</b>	<b>495</b>
Introduction to Encoding . . . . .		496
Lossless and Lossy Encoding . . . . .		496
Blending Representations . . . . .		498
The Simplest Autoencoder . . . . .		500
A Better Autoencoder . . . . .		505
Exploring the Autoencoder . . . . .		508
A Closer Look at the Latent Variables . . . . .		508
The Parameter Space . . . . .		508
Blending Latent Variables . . . . .		513
Predicting from Novel Input . . . . .		515
Convolutional Autoencoders . . . . .		516
Blending Latent Variables . . . . .		517
Predicting from Novel Input . . . . .		519

Denoising . . . . .	519
Variational Autoencoders . . . . .	521
Distribution of Latent Variables . . . . .	522
Variational Autoencoder Structure . . . . .	523
Exploring the VAE . . . . .	530
Working with the MNIST Samples . . . . .	530
Working with Two Latent Variables . . . . .	533
Producing New Input . . . . .	535
Summary . . . . .	538

## **19 RECURRENT NEURAL NETWORKS 539**

Working with Language . . . . .	540
Common Natural Language Processing Tasks . . . . .	540
Transforming Text into Numbers . . . . .	541
Fine-Tuning and Downstream Networks . . . . .	542
Fully Connected Prediction . . . . .	542
Testing Our Network . . . . .	543
Why Our Network Failed . . . . .	546
Recurrent Neural Networks . . . . .	548
Introducing State . . . . .	548
Rolling Up Our Diagram . . . . .	549
Recurrent Cells in Action . . . . .	552
Training a Recurrent Neural Network . . . . .	552
Long Short-Term Memory and Gated Recurrent Networks . . . . .	553
Using Recurrent Neural Networks . . . . .	554
Working with Sunspot Data . . . . .	554
Generating Text . . . . .	555
Different Architectures . . . . .	557
Seq2Seq . . . . .	561
Summary . . . . .	564

## **20 ATTENTION AND TRANSFORMERS 565**

Embedding . . . . .	566
Embedding Words . . . . .	569
ELMo . . . . .	571
Attention . . . . .	574
A Motivating Analogy . . . . .	574
Self-Attention . . . . .	576
Q/KV Attention . . . . .	579
Multi-Head Attention . . . . .	580
Layer Icons . . . . .	581
Transformers . . . . .	581
Skip Connections . . . . .	582
Norm-Add . . . . .	583
Positional Encoding . . . . .	584
Assembling a Transformer . . . . .	586
Transformers in Action . . . . .	589

BERT and GPT-2 . . . . .	590
BERT . . . . .	590
GPT-2 . . . . .	593
Generators Discussion . . . . .	596
Data Poisoning . . . . .	598
Summary . . . . .	599

## **21 REINFORCEMENT LEARNING 601**

Basic Ideas . . . . .	602
Learning a New Game . . . . .	603
The Structure of Reinforcement Learning . . . . .	605
Step 1: The Agent Selects an Action . . . . .	605
Step 2: The Environment Responds . . . . .	606
Step 3: The Agent Updates Itself . . . . .	607
Back to the Big Picture . . . . .	608
Understanding Rewards . . . . .	608
Flippers . . . . .	614
L-Learning . . . . .	616
The Basics . . . . .	616
The L-Learning Algorithm . . . . .	619
Testing Our Algorithm . . . . .	621
Handling Unpredictability . . . . .	624
Q-Learning . . . . .	626
Q-Values and Updates . . . . .	627
Q-Learning Policy . . . . .	630
Putting It All Together . . . . .	632
The Elephant in the Room . . . . .	633
Q-learning in Action . . . . .	634
SARSA . . . . .	638
The Algorithm . . . . .	639
SARSA in Action . . . . .	642
Comparing Q-Learning and SARSA . . . . .	644
The Big Picture . . . . .	646
Summary . . . . .	648

## **22 GENERATIVE ADVERSARIAL NETWORKS 649**

Forging Money . . . . .	650
Learning from Experience . . . . .	652
Forging with Neural Networks . . . . .	653
A Learning Round . . . . .	655
Why Adversarial? . . . . .	656
Implementing GANs . . . . .	657
The Discriminator . . . . .	657
The Generator . . . . .	658
Training the GAN . . . . .	658
GANs in Action . . . . .	660
Building a Discriminator and Generator . . . . .	662
Training Our Network . . . . .	664
Testing Our Network . . . . .	665

DCGANs .....	666
Challenges .....	669
Using Big Samples .....	670
Modal Collapse .....	671
Training with Generated Data .....	671
Summary .....	673

## 23

### CREATIVE APPLICATIONS

Deep Dreaming .....	675
Stimulating Filters .....	676
Running Deep Dreaming .....	678
Neural Style Transfer .....	680
Representing Style .....	680
Representing Content .....	683
Style and Content Together .....	683
Running Style Transfer .....	685
Generating More of This Book .....	688
Summary .....	690
Final Thoughts .....	690

### REFERENCES

693

### IMAGE CREDITS

717

### INDEX

721