Essentials of Pattern Recognition

This textbook introduces fundamental concepts, major models, and popular applications of pattern recognition for a one-semester undergraduate course. To ensure student understanding, the text focuses on a relatively small number of core concepts with an abundance of illustrations and examples. Concepts are reinforced with hands-on exercises to nurture the student's skill in problem solving. New concepts and algorithms are framed by real-world context, and established as part of the big picture introduced in an early chapter. A problem-solving strategy is employed in several chapters to equip students with an approach for new problems in pattern recognition. This text also points out common errors that a new player in pattern recognition may encounter, and fosters the ability of readers to find useful resources and independently solve a new pattern-recognition task through various working examples. Students with an undergraduate understanding of mathematical analysis, linear algebra, and probability will be well prepared to master the concepts and mathematical analysis presented here.

Jianxin Wu is a professor in the Department of Computer Science and Technology and the School of Artificial Intelligence at Nanjing University, China. He received his BS and MS degrees in computer science from Nanjing University and his PhD degree in computer science from the Georgia Institute of Technology. Professor Wu has served as an area chair for the conference on Computer Vision and Pattern Recognition (CVPR), the International Conference on Computer Vision (ICCV), and the AAAI Conference on Artificial Intelligence. He is also an associate editor for the *Pattern Recognition* journal. His research interests are computer vision and machine learning.

Essentials of Pattern Recognition

An Accessible Approach

JIANXIN WU
Nanjing University, China

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108483469
DOI: 10.1017/9781108650212

First published 2021

Printed in the United Kingdom by TJ Books Ltd, Padstow Cornwall, 2021

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-48346-9 Hardback

Additional resources for this publication are at www.cambridge.org/patternrecognition

CAMBRIDGE

Cambridge University Press
978-1-108-48346-9 — Essentials of Pattern Recognition
Jianxin Wu
Frontmatter
More Information

Contents

	List of Figures	*page* ix
	List of Tables	xi
	Preface	xiii
	Notation	xvi
	Part I Introduction and Overview	
1	**Introduction**	3
	1.1 An Example: Autonomous Driving	4
	1.2 Pattern Recognition and Machine Learning	6
	Exercises	12
2	**Mathematical Background**	15
	2.1 Linear Algebra	15
	2.2 Probability	25
	2.3 Optimization and Matrix Calculus	34
	2.4 Complexity of Algorithms	39
	2.5 Miscellaneous Notes and Additional Resources	40
	Exercises	41
3	**Overview of a Pattern Recognition System**	44
	3.1 Face Recognition	44
	3.2 A Simple Nearest Neighbor Classifier	45
	3.3 The Ugly Details	49
	3.4 Making Assumptions and Simplifications	52
	3.5 A Framework	59
	3.6 Miscellaneous Notes and Additional Resources	59
	Exercises	61
4	**Evaluation**	63
	4.1 Accuracy and Error in the Simple Case	63
	4.2 Minimizing the Cost/Loss	70
	4.3 Evaluation in Imbalanced Problems	73

4.4 Can We Reach 100% Accuracy? 79
4.5 Confidence in the Evaluation Results 85
4.6 Miscellaneous Notes and Additional Resources 92
Exercises 93

Part II Domain-Independent Feature Extraction

5 Principal Component Analysis 101
5.1 Motivation 101
5.2 PCA to Zero-Dimensional Subspace 104
5.3 PCA to One-Dimensional Subspace 106
5.4 PCA for More Dimensions 110
5.5 The Complete PCA Algorithm 110
5.6 Variance Analysis 111
5.7 When to Use or Not to Use PCA? 115
5.8 The Whitening Transform 118
5.9 Eigen-Decomposition vs. SVD 118
5.10 Miscellaneous Notes and Additional Resources 119
Exercises 119

6 Fisher's Linear Discriminant 123
6.1 FLD for Binary Classification 125
6.2 FLD for More Classes 132
6.3 Miscellaneous Notes and Additional Resources 135
Exercises 136

Part III Classifiers and Tools

7 Support Vector Machines 143
7.1 The Key SVM Idea 143
7.2 Visualizing and Calculating the Margin 147
7.3 Maximizing the Margin 150
7.4 The Optimization and the Solution 152
7.5 Extensions for Linearly Inseparable and Multiclass Problems 157
7.6 Kernel SVMs 161
7.7 Miscellaneous Notes and Additional Resources 167
Exercises 167

8 Probabilistic Methods 173
8.1 The Probabilistic Way of Thinking 173
8.2 Choices 175
8.3 Parametric Estimation 178
8.4 Nonparametric Estimation 184

	8.5 Making Decisions	191
	8.6 Miscellaneous Notes and Additional Resources	192
	Exercises	192
9	**Distance Metrics and Data Transformations**	196
	9.1 Distance Metrics and Similarity Measures	196
	9.2 Data Transformation and Normalization	207
	9.3 Miscellaneous Notes and Additional Resources	213
	Exercises	213
10	**Information Theory and Decision Trees**	219
	10.1 Prefix Code and Huffman Tree	219
	10.2 Basics of Information Theory	221
	10.3 Information Theory for Continuous Distributions	226
	10.4 Information Theory in ML and PR	231
	10.5 Decision Trees	234
	10.6 Miscellaneous Notes and Additional Resources	239
	Exercises	239
	Part IV Handling Diverse Data Formats	
11	**Sparse and Misaligned Data**	245
	11.1 Sparse Machine Learning	245
	11.2 Dynamic Time Warping	254
	11.3 Miscellaneous Notes and Additional Resources	262
	Exercises	262
12	**Hidden Markov Model**	266
	12.1 Sequential Data and the Markov Property	266
	12.2 Three Basic Problems in HMM Learning	274
	12.3 α, β, and the Evaluation Problem	275
	12.4 γ, δ, ψ, and the Decoding Problem	280
	12.5 ξ and Learning HMM Parameters	283
	12.6 Miscellaneous Notes and Additional Resources	286
	Exercises	287
	Part V Advanced Topics	
13	**The Normal Distribution**	293
	13.1 Definition	293
	13.2 Notation and Parameterization	296
	13.3 Linear Operation and Summation	297
	13.4 Geometry and the Mahalanobis Distance	299

	13.5 Conditioning	300
	13.6 Product of Gaussians	302
	13.7 Application I: Parameter Estimation	303
	13.8 Application II: Kalman Filter	305
	13.9 Useful Math in This Chapter	307
	Exercises	312
14	**The Basic Idea behind Expectation-Maximization**	316
	14.1 GMM: A Worked Example	316
	14.2 An Informal Description of the EM Algorithm	321
	14.3 The Expectation-Maximization Algorithm	321
	14.4 EM for GMM	328
	14.5 Miscellaneous Notes and Additional Resources	330
	Exercises	331
15	**Convolutional Neural Networks**	333
	15.1 Preliminaries	334
	15.2 CNN Overview	336
	15.3 Layer Input, Output, and Notation	341
	15.4 The ReLU Layer	342
	15.5 The Convolution Layer	344
	15.6 The Pooling Layer	356
	15.7 A Case Study: The VGG16 Net	359
	15.8 Hands-On CNN Experiences	361
	15.9 Miscellaneous Notes and Additional Resources	362
	Exercises	362
	Bibliography	365
	Index	379

The plate section can be found between pages 208 and 209

Figures

1.1	A typical pattern recognition pipeline	*page* 7
1.2	A small grayscale image	8
1.3	A typical pattern recognition pipeline with feedback loop	11
2.1	Illustration of vector projection	18
2.2	Probability density function of example normal distributions	33
2.3	Illustration of a simple convex function	37
4.1	Fitting data with polynomials with different degrees of freedom	66
4.2	Illustration of the overfitting phenomenon	69
4.3	One example of the receiver operating characteristics (ROC) curve	76
4.4	One example of the precision–recall (PR) curve	78
4.5	Illustration of bias and variance in a simple polynomial regression task	84
4.6	The standard normal probability density function and the one- and two-sigma ranges	87
4.7	Probability density function of Student's t-distribution	90
5.1	Illustration of various types of relationships between dimensions	102
5.2	Variance of projected values	113
5.3	PCA and the whitening transform applied to Gaussian data	116
5.4	Eigenvalues shown in decreasing order	117
6.1	FLD vs. PCA	124
6.2	Histograms of projected values of the dataset in Figure 6.1	126
7.1	Illustration of complex classification problems	145
7.2	Illustration of the large margin idea	146
7.3	Illustration of projection, margin, and normal vector	148
7.4	Illustration of support vectors	156
7.5	Illustration of nonlinear classifiers	162
8.1	An illustration of Bayesian parameter estimation	183
8.2	Histograms with different numbers of bins	186
9.1	Illustration of two-dimensional points that satisfy $\|\boldsymbol{x}\|_p = 1$	200
9.2	The Manhattan (city block) distance	201
9.3	Illustration of the similarity of two distributions	203
9.4	Illustration of different power mean values	205
9.5	Use the power mean kernel to approximate the histogram intersection kernel	206
9.6	The logistic sigmoid function	211

10.1 An example Huffman tree 220
10.2 Relationships between entropy, conditional entropy, and mutual information 225
10.3 The XOR problem and its decision tree model 236
10.4 Information-gain-based internal node split 237
11.1 The soft thresholding solution 248
11.2 Alignment for `university` vs. `unverstiy` 255
11.3 Visualizing a match between two sequences as a path 257
11.4 The dynamic programming strategy for DTW 259
11.5 The (partial) expansion tree for recursive DTW computations 260
12.1 A simple RNN hidden unit, its unfolded version, and its application in language translation 267
12.2 A discrete-time Markov chain example and its graphical model illustration 270
12.3 The hidden Markov model 272
12.4 Sample code to compute α, β, and (unnormalized) γ variables 279
12.5 Illustration of how to compute $\xi_t(i, j)$ 285
12.6 Various graphical model structures 288
12.7 Example of d-separation 290
13.1 Bivariate normal p.d.f. 299
13.2 Equal probability contour of a bivariate normal distribution 300
14.1 A simple GMM illustration 317
14.2 GMM as a graphical model 318
15.1 Illustration of the gradient descent method, in which η is the learning rate 339
15.2 The ReLU function 343
15.3 Illustration of the convolution operation 344
15.4 An image and the effect of different convolution kernels 347
15.5 Illustration of how to compute $\frac{\partial z}{\partial X}$ 355

Tables

4.1	True and learned parameters of a polynomial regression	*page* 67
4.2	Possible combinations for the groundtruth and the predicted label	74
4.3	Calculation of AUC-PR and AP	95
5.1	MATLAB/GNU Octave code to generate the data points in Figure 5.1	102
5.2	PCA outcome for data in Figures 5.1b, 5.1c, and 5.1d	118
9.1	Ridge regression under different λ values	217
11.1	An example of the integral image	265
12.1	Summary of the variables in HMM learning	285
15.1	Variables, their sizes, and meanings	351
15.2	The VGG-Verydeep-16 architecture and receptive field	359

Preface

Pattern recognition, a research area that extracts useful patterns (or regularities) from data and applies these patterns to subsequent decision processes, has always been an important topic in computer science and related subject areas. Applications of deep learning, the current focus of attention in artificial intelligence, are mainly pattern recognition tasks. Although pattern recognition has direct applications in our society, the shortage of well-trained pattern recognition researchers and practitioners is also obvious.

As an introductory textbook, the purpose of this book is to introduce background and fundamental concepts, major models, and popular applications of pattern recognition. By learning the theories and techniques, followed by hands-on exercises, I hope a beginner will nurture the ability for independent problem solving in the pattern recognition field.

Several classic textbooks have been published in this field. Do we need yet another new one (such as this book)? My answer to this question is yes. These widely adopted pattern recognition textbooks were mostly published a decade ago, but nowadays quite a number of characteristics differ significantly from where the pattern recognition area was ten years ago. Deep learning is a typical example of such novel characteristics. The final chapter of this book introduces convolutional neural networks, the most important deep learning model. Recent achievements and views from the pattern recognition research frontier are also reflected throughout this book.

The major goal, and hopefully the most important feature of this book, however, is *to ensure that all readers understand its contents—even a reader who is not strong (or is even slightly weak) in mathematical and other background knowledge related to pattern recognition.* To achieve this goal, I have used many illustrations and examples, emphasized the cause and effect of various methods (e.g., their motivations, applications, and applicable conditions), and have not omitted any steps in the mathematical derivations. I also provide all necessary background knowledge and encourage the reader to obtain hands-on experience when appropriate. I also wish this book will *serve as an excellent reference book for practitioners* in pattern recognition and machine learning (including deep learning).

Chapter 14 is a good example to illustrate these features. Expectation-maximization (EM) is an algorithm that is important in both pattern recognition and machine learning. However, in the classic textbook (Bishop 1995a), EM occupies only seven pages and the core mathematical derivation only two pages! This succinct treatment may be suitable for experienced or talented readers, but not necessarily for the general audience who are interested in learning pattern recognition.

In Chapter 14 of this book, I introduce the EM algorithm's necessity and main idea through an example (the Gaussian mixture model, GMM), which paves the way for a formal description of EM. The EM algorithm, although very short in its mathematical form, is not easy to follow for beginners. I continue to use GMM as a worked example and prepare the derivation and meaning of every step in full detail. Finally, the EM updating equations for the GMM become obvious and neat. In the exercise problems, I ask the reader to derive the EM algorithm independently without resorting to the chapter. In another exercise problem, I ask the reader to independently derive Baum–Welch, another classic EM derivation—with the help of well-designed hints and steps. For this same EM algorithm, I use 17 pages, and I believe this chapter will help readers not only to learn the EM algorithm smoothly, but also to understand its key ideas and its merits and drawbacks.

Obviously, this book can elaborate on only a carefully selected small subset of core contents. However, other important topics are also briefly mentioned in chapters and exercise problems (e.g., locally linear embedding and the exponential family), and I provide pointers to resources at the end of most chapters if a reader wants to dive deeper into pattern recognition. The core contents of this book may also help a reader to form a foundation for understanding deep learning.

This book also emphasizes hands-on experience. Some details, although not relevant to mathematical derivations, are vital in practical systems. These details are emphasized when appropriate in the book. The design of the exercise problems took me one year. In order to fully understand this book, it is essential that a reader completes these problems. Some problems ask a reader to install software packages, read manuals, and solve problems by writing code.

Finally, beyond teaching knowledge, I want to nurture two kinds of ability in this book. First, when presented with a new task, I want to encourage readers to independently solve it by following these steps: analyzing the problem, obtaining an idea to solve it, formalizing the idea, simplifying the formulation, and then solving it. Second, when encountering a problem that may be easily solved with the help of existing resources, I hope readers can actively search and find such resources (e.g., software packages, manuals, products) such that the problem can be solved promptly rather than reinventing the wheel.

It is always a difficult mission to write a textbook. The conception of this book began in July 2013, when I had just returned to my *alma mater*, Nanjing University, and planned to start a new course, Pattern Recognition. Throughout the six-and-a-half-year writing process, I have been grateful to many people for their kind support. A partial list of persons I wish to acknowledge is shown here in an approximate chronological order.

- Professor James M. Rehg at the Georgia Institute of Technology, my PhD supervisor. Jim's suggestions improved some critical parts of this book too.

- My colleagues at the LAMDA (Learning And Mining from DatA) institute, the Department of Computer Science and Technology (and School of Artificial Intelligence) at Nanjing University. The LAMDA institute provides an excellent research environment. Discussions and collaborations with LAMDA director Professor Zhi-Hua Zhou, my fellow teachers, and LAMDA postgraduate students have been very helpful. LAMDA secretaries also saved me a lot of time by taking care of many administrative procedures.
- Students enrolled in my Pattern Recognition courses. After I used a draft of a few chapters as course notes, the enthusiastic and positive feedback I received from them encouraged me to continue the writing process. Interactions in and after the lectures also greatly shaped the presentation of this book. In addition, numerous typos and errors have been pointed out by them. Although a full list of names is not provided here due to limitations on space , they will be acknowledged in this book's accompanying home page.
- Graduate students under my supervision. The collaborative efforts between myself and my supervisees ensured that research efforts in my small group carried on productively. Thus, I had more time to write this book. They were often also the first batch of readers of this book.
- The managers and editors at Cambridge University Press (CUP). Mr. David Liu from CUP visited my office and discussed a potential publication proposal with me when this book's draft was almost finished. His friendly reminders pushed me to devote more time and to finish the book writing and publication process as early as possible. Ms. Lisa Pinto, lead of the higher education branch's development team at CUP, supervised and greatly helped the editing process.
- My dear family members. Before the birth of my son, this book was mostly written at home, either before I went to the office or after I returned. However, since his birth, especially once he learned to walk, I could use only my time in the office, on an airplane, or on a train trip for this book. I wish to thank my wife, son, parents, and parents-in-law for their love and support!

Any feedback or comments are most welcome. Please send them to the following email address: pr.book.wujx@gmail.com.

Notation

$\mathbb{R}$	set of real values
$\mathbb{R}_+$	set of nonnegative real values
$\mathbb{Z}$	set of integers
$\triangleq$	defined as
$(\cdot)^T$	transpose of a matrix
$\mathbf{1}, \mathbf{0}$	vector of all 1s or all 0s, respectively
$\\|\cdot\\|$	norm of a matrix or a vector
$\boldsymbol{x} \perp \boldsymbol{y}$	two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal
I_n (I)	identity matrix of size $n \times n$
$\det(X)$ or $\lvert X\rvert$	determinant of X when X is a square matrix
$\lvert D\rvert$	size (cardinality) of D when D is a set
X^{-1}	inverse of a square matrix X
X^+	Moore–Penrose pseudoinverse of matrix X
$\operatorname{tr}(X)$	trace of a square matrix X
$\operatorname{rank}(X)$	rank of a matrix X
$\operatorname{diag}(a_1, a_2, \dots, a_n)$	diagonal matrix with diagonal entries being a_i
$\operatorname{diag}(X)$	vector formed from diagonal entries in square matrix X
$X \succ 0$ $(X \succcurlyeq 0)$	square matrix X is positive (semi)definite
$\Pr(\cdot)$	probability of an event
$\mathbb{E}_X[f(X)]$	expectation of $f(X)$ with respect to X
$\operatorname{Var}(X)$ $(\operatorname{Cov}(X))$	variance (covariance matrix) of X
$\rho_{X,Y}$	Pearson's correlation coefficient
$N(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$[\![\cdot]\!]$	indicator function
$\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$	ceiling and floor functions, respectively
$\operatorname{sign}(x)$	the sign of $x \in \mathbb{R}$, can be 0, 1 or -1
$\propto$	proportional to
x_+	hinge loss, $x_+ = \max(0, x)$
$\mathcal{O}(\cdot)$	big-O notation
$x_{1:t}$	abbreviation for the sequence $x_1, x_2, \dots, x_t$

