

# Applied Machine Learning

David Forsyth

# Applied Machine Learning

 Springer

David Forsyth  
Computer Science Department  
University of Illinois Urbana Champaign  
Urbana, IL, USA

ISBN 978-3-030-18113-0      ISBN 978-3-030-18114-7 (eBook)  
<https://doi.org/10.1007/978-3-030-18114-7>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG. The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Machine learning methods are now an important tool for scientists, researchers, engineers, and students in a wide range of areas. Many years ago, one could publish papers introducing (say) classifiers to one of the many fields that hadn't heard of them. Now, you need to know what a classifier is to get started in most fields. This book is written for people who want to adopt and use the main tools of machine learning but aren't necessarily going to want to be machine learning researchers—as of writing, this seems like almost everyone. There is no new fact about machine learning here, but the selection of topics is my own. I think it's different from what one sees in other books.

The book is revised and corrected from class notes written for a course I've taught on numerous occasions to very broad audiences. Most students were at the final year undergraduate/first year graduate student level in a US university. About half of each class consisted of students who weren't computer science students but still needed a background in learning methods. The course stressed applying a wide range of methods to real datasets, and the book does so, too.

The key principle in choosing what to write about was to cover the ideas in machine learning that I thought everyone who was going to use learning tools should have seen, whatever their chosen specialty or career. Although it's never a good thing to be ignorant of anything, an author must choose. Most people will find a broad shallow grasp of this field more useful than a deep and narrow grasp, so this book is broad, and coverage of many areas is shallow. I think that's fine, because my purpose is to ensure that all have seen enough to know that, say, firing up a classification package will make many problems go away. So I've covered enough to get you started and to get you to realize that it's worth knowing more.

The notes I wrote have been useful to more experienced students as well. In my experience, many learned some or all of this material without realizing how useful it was and then forgot it. If this happened to you, I hope the book is a stimulus to your memory. You really should have a grasp of all of this material. You might need to know more, but you certainly shouldn't know less.

## This Book

I wrote this book to be taught, or read, by starting at the beginning and proceeding to the end. In a 15-week semester, I cover a lot and usually set 12 assignments, always programming assignments. Different instructors or readers may have different needs, and so I sketch some pointers to what can be omitted below.

## What You Need to Know Before You Start

This book assumes you have a moderate background in probability and statistics before you start. I wrote a companion book, *Probability and Statistics for Computer Science*, which covers this background. There is a little overlap, because not everyone will read both books cover to cover (a mistake—you should!). But I've

kept the overlap small (about 40 pp.) and confined to material that is better repeated anyway. Here's what you should know from that book (or some other, if you insist):

- Various descriptive statistics (mean, standard deviation, variance) and visualization methods for 1D datasets
- Scatter plots, correlation, and prediction for 2D datasets
- A little discrete probability
- A very little continuous probability (rough grasp of probability density functions and how to interpret them)
- Random variables and expectations
- A little about samples and populations
- Maximum likelihood
- Simple Bayesian inference
- A selection of facts about an assortment of useful probability distributions, or where to look them up

**General Background:** Your linear algebra should be reasonably fluent at a practical level. Fairly soon, we will see matrices, vectors, orthonormal matrices, eigenvalues, eigenvectors, and the singular value decomposition. All of these ideas will be used without too much comment.

**Programming Background:** You should be able to pick up a practical grasp of a programming environment without too much fuss. I use either R or MATLAB for this sort of thing, depending on how reliable I think a particular package is. You should, too. In some places, it's a good idea to use Python.

**Survival Skills:** Most simple questions about programming can be answered by searching. I usually use a web search to supply details of syntax, particular packages, etc. which I forget easily. You should, too. When someone asks me, say, "how do I write a loop in R?" in office hours, I very often answer by searching for R loop (or whatever) and then pointing out that I'm not actually needed. The questioner is often embarrassed at this point. You could save everyone trouble and embarrassment by cutting out the middleman in this transaction.

### Datasets and Broken Links

I think using real datasets is an important feature of this book. But real life is messy, and datasets I refer to here may have been moved by the time you read this. Generally, a little poking around using a web search engine will find datasets or items that have been moved. I will also try to provide pointers to missing or moved datasets on a webpage which will also depend from my home page. It's easy to find me on the Internet by searching for my name, and ignoring the soap opera star (that's really not me).

### Citations

Generally, I have followed the natural style of a textbook and tried not to cite papers in the text. I'm not up to providing a complete bibliography of modern machine learning and didn't want to provide an incomplete one. However, I have mentioned papers in some places and have done so when it seemed very important for users

to be aware of the paper or when I was using a dataset whose compilers asked for a citation. I will try to correct errors or omissions in citation on a webpage, which will depend from my home page.

### What Has Been Omitted

A list of everything omitted would be impractically too long. There are three topics I most regret omitting: kernel methods, reinforcement learning, and neural sequence models like LSTM. I omitted each because I thought that, while each is an important part of the toolbox of a practitioner, there are other topics with more claim on space. I may well write additional chapters on these topics after I recover from finishing this book. When they're in a reasonable shape, I'll put them on a webpage that depends from my home page.

There is very little learning theory here. While learning theory is very important (and I put in a brief chapter to sketch what's there and give readers a flavor), it doesn't directly change what practitioners *do*. And there's quite a lot of machinery with weak theoretical underpinnings that is extremely useful.

Urbana, IL, USA

David Forsyth

# Acknowledgments

I acknowledge a wide range of intellectual debts, starting at kindergarten. Important figures in the very long list of my creditors include Gerald Alanthwaite, Mike Brady, Tom Fair, Margaret Fleck, Jitendra Malik, Joe Mundy, Jean Ponce, Mike Rodd, Charlie Rothwell, and Andrew Zisserman.

I have benefited from looking at a variety of sources, though this work really is my own. I particularly enjoyed the following books:

- *The Nature of Statistical Learning Theory*, V. Vapnik; Springer, 1999
- *Machine Learning: A Probabilistic Perspective*, K. P. Murphy; MIT Press, 2012
- *Pattern Recognition and Machine Learning*, C. M. Bishop; Springer, 2011
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, T. Hastie, R. Tibshirani, and J. Friedman; Springer, 2016
- *An Introduction to Statistical Learning: With Applications in R*, G. James, D. Witten, T. Hastie, and R. Tibshirani; Springer, 2013
- *Deep Learning*, I. Goodfellow, Y. Bengio, and A. Courville; MIT Press, 2016
- *Probabilistic Graphical Models: Principles and Techniques*, D. Koller and N. Friedman; MIT Press, 2009
- *Artificial Intelligence: A Modern Approach*, Third Edition, S. J. Russell and P. Norvig; Pearson, 2015
- *Data Analysis and Graphics Using R: An Example-Based Approach*, J. Maindonald and W. J. Braun; Cambridge University Press, 2e, 2003

A wonderful feature of modern scientific life is the willingness of people to share data on the Internet. I have roamed the Internet widely looking for datasets and have tried to credit the makers and sharers of data accurately and fully when I use the dataset. If, by some oversight, I have left you out, please tell me and I will try and fix this. I have been particularly enthusiastic about using data from the following repositories:

- *The UC Irvine Machine Learning Repository*, at <http://archive.ics.uci.edu/ml/>
- *Dr. John Rasp's Statistics Website*, at <http://www2.stetson.edu/~jrasp/>
- *OzDasl: The Australasian Data and Story Library*, at <http://www.statsci.org/data/>
- *The Center for Genome Dynamics, at the Jackson Laboratory*, at <http://cgd.jax.org/> (which contains staggering amounts of information about mice) and the datasets listed and described in Sects. 17.2 and 18.1

I looked at Wikipedia regularly when preparing this manuscript, and I've pointed readers to neat stories there when they're relevant. I don't think one could learn the material in this book by reading Wikipedia, but it's been tremendously helpful in restoring ideas that I have mislaid, mangled, or simply forgotten.

When I did the first version of this course, Alyosha Efros let me look at notes for a learning course he taught, and these affected my choices of topic. Ben Recht gave me advice on several choices of topic. I co-taught this class with Trevor Walker

for one semester, and his comments were extremely valuable. Eric Huber has made numerous suggestions in his role as course lead for an offering that included an online component. TA's for various versions of this and related classes have also helped improve the notes. Thanks to: Jyoti Aneja, Lavisha Aggarwal, Xiaoyang Bai, Christopher Benson, Shruti Bhargava, Anand Bhattad, Daniel Calzada, Binglin Chen, Taiyu Dong, Tanmay Gangwani, Sili Hui, Ayush Jain, Krishna Kothapalli, Maghav Kumar, Ji Li, Qixuan Li, Jiajun Lu, Shreya Rajpal, Jason Rock, Daeyun Shin, Mariya Vasileva, and Anirud Yadav. Typo's were spotted by (at least!): Johnny Chang, Yan Geng Niv Hadas, Vivian Hu, Eric Huber, Michael McCarrin, Thai Duy Cuong Nguyen, Jian Peng, and Victor Sui.

Several people commented very helpfully on the deep network part of the book, including Mani Golparvar Fard, Tanmay Gupta, Arun Mallya, Amin Sadeghi, Sepehr Sameni, and Alex Schwing.

I have benefited hugely from reviews organized by the publisher. Reviewers made many extremely helpful suggestions, which I have tried to adopt, by cutting chapters, moving chapters around, and general re-engineering of topics. Reviewers were anonymous to me at the time of review, but their names were later revealed so I can thank them by name. Thanks to:

Xiaoming Huo, Georgia Institute of Technology  
 Georgios Lazarou, University of South Alabama  
 Ilias Tagkopoulos, University of California, Davis  
 Matthew Turk, University of California, Santa Barbara  
 George Tzanetakis, University of Victoria  
 Qin Wang, University of Alabama  
 Guanghui Wang, University of Kansas  
 Jie Yang, University of Illinois at Chicago  
 Lisa Zhang, University of Toronto, Mississauga

A long list of people have tried to help me make this book better, and I'm very grateful for their efforts. But what remains is my fault, not theirs. Sorry.



# Contents

<b>I</b>	<b>Classification</b>	<b>1</b>
<b>1</b>	<b>Learning to Classify</b>	<b>3</b>
1.1	Classification: The Big Ideas . . . . .	4
1.1.1	The Error Rate and Other Summaries of Performance . . . . .	4
1.1.2	More Detailed Evaluation . . . . .	5
1.1.3	Overfitting and Cross-Validation . . . . .	6
1.2	Classifying with Nearest Neighbors . . . . .	7
1.2.1	Practical Considerations for Nearest Neighbors . . . . .	8
1.3	Naive Bayes . . . . .	10
1.3.1	Cross-Validation to Choose a Model . . . . .	13
1.3.2	Missing Data . . . . .	15
1.4	You Should . . . . .	16
1.4.1	Remember These Terms . . . . .	16
1.4.2	Remember These Facts . . . . .	16
1.4.3	Remember These Procedures . . . . .	17
1.4.4	Be Able to . . . . .	17
<b>2</b>	<b>SVMs and Random Forests</b>	<b>21</b>
2.1	The Support Vector Machine . . . . .	21
2.1.1	The Hinge Loss . . . . .	22
2.1.2	Regularization . . . . .	24
2.1.3	Finding a Classifier with Stochastic Gradient Descent . . . . .	25
2.1.4	Searching for $\lambda$ . . . . .	27
2.1.5	Summary: Training with Stochastic Gradient Descent . . . . .	29
2.1.6	Example: Adult Income with an SVM . . . . .	30
2.1.7	Multiclass Classification with SVMs . . . . .	33
2.2	Classifying with Random Forests . . . . .	34
2.2.1	Building a Decision Tree . . . . .	35
2.2.2	Choosing a Split with Information Gain . . . . .	38
2.2.3	Forests . . . . .	41
2.2.4	Building and Evaluating a Decision Forest . . . . .	41
2.2.5	Classifying Data Items with a Decision Forest . . . . .	42
2.3	You Should . . . . .	44
2.3.1	Remember These Terms . . . . .	44
2.3.2	Remember These Facts . . . . .	44
2.3.3	Use These Procedures . . . . .	45
2.3.4	Be Able to . . . . .	45

<b>3</b>	<b>A Little Learning Theory</b>	<b>49</b>
3.1	Held-Out Loss Predicts Test Loss . . . . .	49
3.1.1	Sample Means and Expectations . . . . .	50
3.1.2	Using Chebyshev’s Inequality . . . . .	52
3.1.3	A Generalization Bound . . . . .	52
3.2	Test and Training Error for a Classifier from a Finite Family . . . . .	53
3.2.1	Hoeffding’s Inequality . . . . .	54
3.2.2	Test from Training for a Finite Family of Predictors . . . . .	55
3.2.3	Number of Examples Required . . . . .	56
3.3	An Infinite Collection of Predictors . . . . .	57
3.3.1	Predictors and Binary Functions . . . . .	57
3.3.2	Symmetrization . . . . .	61
3.3.3	Bounding the Generalization Error . . . . .	62
3.4	You Should . . . . .	64
3.4.1	Remember These Terms . . . . .	64
3.4.2	Remember These Facts . . . . .	64
3.4.3	Be Able to . . . . .	65
<b>II</b>	<b>High Dimensional Data</b>	<b>67</b>
<b>4</b>	<b>High Dimensional Data</b>	<b>69</b>
4.1	Summaries and Simple Plots . . . . .	69
4.1.1	The Mean . . . . .	70
4.1.2	Stem Plots and Scatterplot Matrices . . . . .	70
4.1.3	Covariance . . . . .	73
4.1.4	The Covariance Matrix . . . . .	74
4.2	The Curse of Dimension . . . . .	77
4.2.1	The Curse: Data Isn’t Where You Think It Is . . . . .	77
4.2.2	Minor Banes of Dimension . . . . .	78
4.3	Using Mean and Covariance to Understand High Dimensional Data . . . . .	79
4.3.1	Mean and Covariance Under Affine Transformations . . . . .	80
4.3.2	Eigenvectors and Diagonalization . . . . .	81
4.3.3	Diagonalizing Covariance by Rotating Blobs . . . . .	82
4.4	The Multivariate Normal Distribution . . . . .	83
4.4.1	Affine Transformations and Gaussians . . . . .	84
4.4.2	Plotting a 2D Gaussian: Covariance Ellipses . . . . .	85
4.4.3	Descriptive Statistics and Expectations . . . . .	86
4.4.4	More from the Curse of Dimension . . . . .	87
4.5	You Should . . . . .	88
4.5.1	Remember These Terms . . . . .	88
4.5.2	Remember These Facts . . . . .	88
4.5.3	Remember These Procedures . . . . .	89
<b>5</b>	<b>Principal Component Analysis</b>	<b>93</b>
5.1	Representing Data on Principal Components . . . . .	93
5.1.1	Approximating Blobs . . . . .	93
5.1.2	Example: Transforming the Height–Weight Blob . . . . .	94

5.1.3	Representing Data on Principal Components . . . . .	96
5.1.4	The Error in a Low Dimensional Representation . . . . .	98
5.1.5	Extracting a Few Principal Components with NIPALS . . . . .	99
5.1.6	Principal Components and Missing Values . . . . .	101
5.1.7	PCA as Smoothing . . . . .	103
5.2	Example: Representing Colors with Principal Components . . . . .	105
5.3	Example: Representing Faces with Principal Components . . . . .	109
5.4	You Should . . . . .	111
5.4.1	Remember These Terms . . . . .	111
5.4.2	Remember These Facts . . . . .	111
5.4.3	Remember These Procedures . . . . .	111
5.4.4	Be Able to . . . . .	111
<b>6</b>	<b>Low Rank Approximations</b>	<b>117</b>
6.1	The Singular Value Decomposition . . . . .	117
6.1.1	SVD and PCA . . . . .	119
6.1.2	SVD and Low Rank Approximations . . . . .	120
6.1.3	Smoothing with the SVD . . . . .	120
6.2	Multidimensional Scaling . . . . .	122
6.2.1	Choosing Low D Points Using High D Distances . . . . .	122
6.2.2	Using a Low Rank Approximation to Factor . . . . .	123
6.2.3	Example: Mapping with Multidimensional Scaling . . . . .	124
6.3	Example: Text Models and Latent Semantic Analysis . . . . .	126
6.3.1	The Cosine Distance . . . . .	127
6.3.2	Smoothing Word Counts . . . . .	128
6.3.3	Example: Mapping NIPS Documents . . . . .	130
6.3.4	Obtaining the Meaning of Words . . . . .	130
6.3.5	Example: Mapping NIPS Words . . . . .	133
6.3.6	TF-IDF . . . . .	134
6.4	You Should . . . . .	136
6.4.1	Remember These Terms . . . . .	136
6.4.2	Remember These Facts . . . . .	136
6.4.3	Remember These Procedures . . . . .	136
6.4.4	Be Able to . . . . .	136
<b>7</b>	<b>Canonical Correlation Analysis</b>	<b>139</b>
7.1	Canonical Correlation Analysis . . . . .	139
7.2	Example: CCA of Words and Pictures . . . . .	142
7.3	Example: CCA of Albedo and Shading . . . . .	144
7.3.1	Are Correlations Significant? . . . . .	148
7.4	You Should . . . . .	150
7.4.1	Remember These Terms . . . . .	150
7.4.2	Remember These Facts . . . . .	150
7.4.3	Remember These Procedures . . . . .	150
7.4.4	Be Able to . . . . .	150

### III Clustering 153

#### 8 Clustering 155

8.1	Agglomerative and Divisive Clustering . . . . .	155
8.1.1	Clustering and Distance . . . . .	157
8.2	The $k$ -Means Algorithm and Variants . . . . .	159
8.2.1	How to Choose $k$ . . . . .	163
8.2.2	Soft Assignment . . . . .	164
8.2.3	Efficient Clustering and Hierarchical $k$ -Means . . . . .	166
8.2.4	$k$ -Medoids . . . . .	167
8.2.5	Example: Groceries in Portugal . . . . .	167
8.2.6	General Comments on $k$ -Means . . . . .	170
8.3	Describing Repetition with Vector Quantization . . . . .	171
8.3.1	Vector Quantization . . . . .	172
8.3.2	Example: Activity from Accelerometer Data . . . . .	175
8.4	You Should . . . . .	178
8.4.1	Remember These Terms . . . . .	178
8.4.2	Remember These Facts . . . . .	178
8.4.3	Remember These Procedures . . . . .	178

#### 9 Clustering Using Probability Models 183

9.1	Mixture Models and Clustering . . . . .	183
9.1.1	A Finite Mixture of Blobs . . . . .	184
9.1.2	Topics and Topic Models . . . . .	185
9.2	The EM Algorithm . . . . .	188
9.2.1	Example: Mixture of Normals: The E-step . . . . .	189
9.2.2	Example: Mixture of Normals: The M-step . . . . .	191
9.2.3	Example: Topic Model: The E-step . . . . .	192
9.2.4	Example: Topic Model: The M-step . . . . .	193
9.2.5	EM in Practice . . . . .	193
9.3	You Should . . . . .	198
9.3.1	Remember These Terms . . . . .	198
9.3.2	Remember These Facts . . . . .	198
9.3.3	Remember These Procedures . . . . .	198
9.3.4	Be Able to . . . . .	198

### IV Regression 203

#### 10 Regression 205

10.1	Overview . . . . .	205
10.1.1	Regression to Spot Trends . . . . .	206
10.2	Linear Regression and Least Squares . . . . .	208
10.2.1	Linear Regression . . . . .	209
10.2.2	Choosing $\beta$ . . . . .	210
10.2.3	Residuals . . . . .	212
10.2.4	$R$ -squared . . . . .	212

10.2.5	Transforming Variables . . . . .	214
10.2.6	Can You Trust Your Regression? . . . . .	217
10.3	Visualizing Regressions to Find Problems . . . . .	218
10.3.1	Problem Data Points Have Significant Impact . . . . .	219
10.3.2	The Hat Matrix and Leverage . . . . .	222
10.3.3	Cook's Distance . . . . .	223
10.3.4	Standardized Residuals . . . . .	224
10.4	Many Explanatory Variables . . . . .	225
10.4.1	Functions of One Explanatory Variable . . . . .	227
10.4.2	Regularizing Linear Regressions . . . . .	227
10.4.3	Example: Weight Against Body Measurements . . . . .	232
10.5	You Should . . . . .	236
10.5.1	Remember These Terms . . . . .	236
10.5.2	Remember These Facts . . . . .	236
10.5.3	Remember These Procedures . . . . .	236
10.5.4	Be Able to . . . . .	237
<b>11</b>	<b>Regression: Choosing and Managing Models</b>	<b>245</b>
11.1	Model Selection: Which Model Is Best? . . . . .	245
11.1.1	Bias and Variance . . . . .	246
11.1.2	Choosing a Model Using Penalties: AIC and BIC . . . . .	248
11.1.3	Choosing a Model Using Cross-Validation . . . . .	250
11.1.4	Greedy Search with Stagewise Regression . . . . .	251
11.1.5	What Variables Are Important? . . . . .	252
11.2	Robust Regression . . . . .	253
11.2.1	M-Estimators and Iteratively Reweighted Least Squares . . . . .	254
11.2.2	Scale for M-Estimators . . . . .	257
11.3	Generalized Linear Models . . . . .	258
11.3.1	Logistic Regression . . . . .	258
11.3.2	Multiclass Logistic Regression . . . . .	260
11.3.3	Regressing Count Data . . . . .	261
11.3.4	Deviance . . . . .	262
11.4	L1 Regularization and Sparse Models . . . . .	262
11.4.1	Dropping Variables with L1 Regularization . . . . .	263
11.4.2	Wide Datasets . . . . .	267
11.4.3	Using Sparsity Penalties with Other Models . . . . .	270
11.5	You Should . . . . .	271
11.5.1	Remember These Terms . . . . .	271
11.5.2	Remember These Facts . . . . .	271
11.5.3	Remember These Procedures . . . . .	272
<b>12</b>	<b>Boosting</b>	<b>275</b>
12.1	Greedy and Stagewise Methods for Regression . . . . .	276
12.1.1	Example: Greedy Stagewise Linear Regression . . . . .	276
12.1.2	Regression Trees . . . . .	279
12.1.3	Greedy Stagewise Regression with Trees . . . . .	279

12.2	Boosting a Classifier . . . . .	284
12.2.1	The Loss . . . . .	284
12.2.2	Recipe: Stagewise Reduction of Loss . . . . .	286
12.2.3	Example: Boosting Decision Stumps . . . . .	288
12.2.4	Gradient Boost with Decision Stumps . . . . .	289
12.2.5	Gradient Boost with Other Predictors . . . . .	290
12.2.6	Example: Is a Prescriber an Opiate Prescriber? . . . . .	291
12.2.7	Pruning the Boosted Predictor with the Lasso . . . . .	293
12.2.8	Gradient Boosting Software . . . . .	294
12.3	You Should . . . . .	299
12.3.1	Remember These Definitions . . . . .	299
12.3.2	Remember These Terms . . . . .	299
12.3.3	Remember These Facts . . . . .	299
12.3.4	Remember These Procedures . . . . .	300
12.3.5	Be Able to . . . . .	300
<b>V</b>	<b>Graphical Models</b>	<b>303</b>
<b>13</b>	<b>Hidden Markov Models</b>	<b>305</b>
13.1	Markov Chains . . . . .	305
13.1.1	Transition Probability Matrices . . . . .	309
13.1.2	Stationary Distributions . . . . .	311
13.1.3	Example: Markov Chain Models of Text . . . . .	313
13.2	Hidden Markov Models and Dynamic Programming . . . . .	316
13.2.1	Hidden Markov Models . . . . .	316
13.2.2	Picturing Inference with a Trellis . . . . .	317
13.2.3	Dynamic Programming for HMMs: Formalities . . . . .	320
13.2.4	Example: Simple Communication Errors . . . . .	321
13.3	Learning an HMM . . . . .	323
13.3.1	When the States Have Meaning . . . . .	324
13.3.2	Learning an HMM with EM . . . . .	324
13.4	You Should . . . . .	329
13.4.1	Remember These Terms . . . . .	329
13.4.2	Remember These Facts . . . . .	330
13.4.3	Be Able to . . . . .	330
<b>14</b>	<b>Learning Sequence Models Discriminatively</b>	<b>333</b>
14.1	Graphical Models . . . . .	333
14.1.1	Inference and Graphs . . . . .	334
14.1.2	Graphical Models . . . . .	336
14.1.3	Learning in Graphical Models . . . . .	337
14.2	Conditional Random Field Models for Sequences . . . . .	338
14.2.1	MEMMs and Label Bias . . . . .	339
14.2.2	Conditional Random Field Models . . . . .	341
14.2.3	Learning a CRF Takes Care . . . . .	342

14.3	Discriminative Learning of CRFs . . . . .	343
14.3.1	Representing the Model . . . . .	343
14.3.2	Example: Modelling a Sequence of Digits . . . . .	344
14.3.3	Setting Up the Learning Problem . . . . .	345
14.3.4	Evaluating the Gradient . . . . .	346
14.4	You Should . . . . .	348
14.4.1	Remember These Terms . . . . .	348
14.4.2	Remember These Procedures . . . . .	348
14.4.3	Be Able to . . . . .	348
<b>15</b>	<b>Mean Field Inference</b>	<b>351</b>
15.1	Useful but Intractable Models . . . . .	351
15.1.1	Denoising Binary Images with Boltzmann Machines . . . . .	352
15.1.2	A Discrete Markov Random Field . . . . .	353
15.1.3	Denoising and Segmenting with Discrete MRFs . . . . .	354
15.1.4	MAP Inference in Discrete MRFs Can Be Hard . . . . .	357
15.2	Variational Inference . . . . .	358
15.2.1	The KL Divergence . . . . .	359
15.2.2	The Variational Free Energy . . . . .	360
15.3	Example: Variational Inference for Boltzmann Machines . . . . .	361
15.4	You Should . . . . .	364
15.4.1	Remember These Terms . . . . .	364
15.4.2	Remember These Facts . . . . .	364
15.4.3	Be Able to . . . . .	364
<b>VI</b>	<b>Deep Networks</b>	<b>365</b>
<b>16</b>	<b>Simple Neural Networks</b>	<b>367</b>
16.1	Units and Classification . . . . .	367
16.1.1	Building a Classifier out of Units: The Cost Function . . . . .	368
16.1.2	Building a Classifier out of Units: Strategy . . . . .	369
16.1.3	Building a Classifier out of Units: Training . . . . .	370
16.2	Example: Classifying Credit Card Accounts . . . . .	372
16.3	Layers and Networks . . . . .	377
16.3.1	Stacking Layers . . . . .	377
16.3.2	Jacobians and the Gradient . . . . .	379
16.3.3	Setting up Multiple Layers . . . . .	380
16.3.4	Gradients and Backpropagation . . . . .	381
16.4	Training Multilayer Networks . . . . .	383
16.4.1	Software Environments . . . . .	385
16.4.2	Dropout and Redundant Units . . . . .	386
16.4.3	Example: Credit Card Accounts Revisited . . . . .	387
16.4.4	Advanced Tricks: Gradient Scaling . . . . .	390
16.5	You Should . . . . .	394
16.5.1	Remember These Terms . . . . .	394
16.5.2	Remember These Facts . . . . .	394
16.5.3	Remember These Procedures . . . . .	394
16.5.4	Be Able to . . . . .	395

<b>17 Simple Image Classifiers</b>	<b>399</b>
17.1 Image Classification	399
17.1.1 Pattern Detection by Convolution	401
17.1.2 Convolutional Layers upon Convolutional Layers	407
17.2 Two Practical Image Classifiers	408
17.2.1 Example: Classifying MNIST	410
17.2.2 Example: Classifying CIFAR-10	412
17.2.3 Quirks: Adversarial Examples	418
17.3 You Should	420
17.3.1 Remember These Definitions	420
17.3.2 Remember These Terms	420
17.3.3 Remember These Facts	420
17.3.4 Remember These Procedures	420
17.3.5 Be Able to	420
<b>18 Classifying Images and Detecting Objects</b>	<b>423</b>
18.1 Image Classification	423
18.1.1 Datasets for Classifying Images of Objects	424
18.1.2 Datasets for Classifying Images of Scenes	426
18.1.3 Augmentation and Ensembles	427
18.1.4 AlexNet	428
18.1.5 VGGNet	430
18.1.6 Batch Normalization	432
18.1.7 Computation Graphs	433
18.1.8 Inception Networks	434
18.1.9 Residual Networks	436
18.2 Object Detection	438
18.2.1 How Object Detectors Work	438
18.2.2 Selective Search	440
18.2.3 R-CNN, Fast R-CNN and Faster R-CNN	441
18.2.4 YOLO	443
18.2.5 Evaluating Detectors	445
18.3 Further Reading	447
18.4 You Should	449
18.4.1 Remember These Terms	449
18.4.2 Remember These Facts	449
18.4.3 Be Able to	450
<b>19 Small Codes for Big Signals</b>	<b>455</b>
19.1 Better Low Dimensional Maps	455
19.1.1 Sammon Mapping	456
19.1.2 T-SNE	457
19.2 Maps That Make Low-D Representations	460
19.2.1 Encoders, Decoders, and Autoencoders	461
19.2.2 Making Data Blocks Bigger	462
19.2.3 The Denoising Autoencoder	465



19.3	Generating Images from Examples . . . . .	469
19.3.1	Variational Autoencoders . . . . .	470
19.3.2	Adversarial Losses: Fooling a Classifier . . . . .	471
19.3.3	Matching Distributions with Test Functions . . . . .	473
19.3.4	Matching Distributions by Looking at Distances . . . . .	474
19.4	You Should . . . . .	475
19.4.1	Remember These Terms . . . . .	475
19.4.2	Remember These Facts . . . . .	476
19.4.3	Be Able to . . . . .	476
	<b>Index</b>	<b>479</b>
	<b>Index: Useful Facts</b>	<b>485</b>
	<b>Index: Procedures</b>	<b>487</b>
	<b>Index: Worked Examples</b>	<b>489</b>
	<b>Index: Remember This</b>	<b>491</b>

# About the Author

**David Forsyth** grew up in Cape Town. He received his B.Sc. (Elec. Eng.) and M.Sc. (Elec. Eng.) from the University of the Witwatersrand, Johannesburg, in 1984 and in 1986, respectively and D.Phil. from Balliol College, Oxford, in 1989. He spent 3 years on the Faculty at the University of Iowa and 10 years on the Faculty at the University of California at Berkeley and then moved to the University of Illinois. He served as program co-chair for IEEE Computer Vision and Pattern Recognition in 2000, 2011, 2018, and 2021; general co-chair for CVPR 2006 and ICCV 2019, and program co-chair for the European Conference on Computer Vision 2008 and is a regular member of the program committee of all major international conferences on computer vision. He has also served six terms on the SIGGRAPH program committee. In 2006, he received an IEEE Technical Achievement Award and in 2009 and 2014 he was named an IEEE Fellow and an ACM Fellow, respectively; he served as editor in chief of IEEE TPAMI from 2014 to 2017; he is lead coauthor of *Computer Vision: A Modern Approach*, a textbook of computer vision that ran to two editions and four languages; and is sole author of *Probability and Statistics for Computer Science*, which provides the background for this book. Among a variety of odd hobbies, he is a compulsive diver, certified up to normoxic trimix level.