
Probability and Statistics for Computer Science

David Forsyth

Probability and Statistics for Computer Science

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

ISBN 978-3-319-64409-7 ISBN 978-3-319-64410-3 (eBook)
<https://doi.org/10.1007/978-3-319-64410-3>

Library of Congress Control Number: 2017950289

© Springer International Publishing AG 2018

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.

Printed on acid-free paper

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

To my family

Preface

An understanding of probability and statistics is an essential tool for a modern computer scientist. If your tastes run to theory, then you need to know a lot of probability (e.g., to understand randomized algorithms, to understand the probabilistic method in graph theory, to understand a lot of work on approximation, and so on) and at least enough statistics to bluff successfully on occasion. If your tastes run to the practical, you will find yourself constantly raiding the larder of statistical techniques (particularly classification, clustering, and regression). For example, much of modern artificial intelligence is built on clever pirating of statistical ideas. As another example, thinking about statistical inference for gigantic datasets has had a tremendous influence on how people build modern computer systems.

Computer science undergraduates traditionally are required to take either a course in probability, typically taught by the math department, or a course in statistics, typically taught by the statistics department. A curriculum committee in my department decided that the curricula of these courses could do with some revision. So I taught a trial version of a course, for which I wrote notes; these notes became this book. There is no new fact about probability or statistics here, but the selection of topics is my own; I think it's quite different from what one sees in other books.

The key principle in choosing what to write about was to cover the ideas in probability and statistics that I thought every computer science undergraduate student should have seen, whatever their chosen specialty or career. This means the 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 graduate 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.

Reading and Teaching This Book

I wrote this book to be taught, or read, by starting at the beginning and proceeding to the end. Different instructors or readers may have different needs, and so I sketch some pointers to what can be omitted below.

Describing Datasets

This part covers:

- Various descriptive statistics (mean, standard deviation, variance) and visualization methods for 1D datasets
- Scatter plots, correlation, and prediction for 2D datasets

Most people will have seen some, but not all, of this material. In my experience, it takes some time for people to really internalize just how useful it is to make pictures of datasets. I've tried to emphasize this point strongly by investigating a variety of datasets in worked examples. When I teach this material, I move through these chapters slowly and carefully.

Probability

This part covers:

- Discrete probability, developed fairly formally
- Conditional probability, with a particular emphasis on examples, because people find this topic counterintuitive
- Random variables and expectations
- Just a little continuous probability (probability density functions and how to interpret them)
- Markov's inequality, Chebyshev's inequality, and the weak law of large numbers
- A selection of facts about an assortment of useful probability distributions
- The normal approximation to a binomial distribution with large N

I've been quite careful developing discrete probability fairly formally. Most people find conditional probability counterintuitive (or, at least, behave as if they do—you can still start a fight with the Monty Hall problem), and so I've used a number of (sometimes startling) examples to emphasize how useful it is to tread carefully here. In my experience, worked examples help learning, but I found that too many worked examples in any one section could become distracting, so there's an entire section of extra worked examples. You can't omit anything here, except perhaps the extra worked examples.

The chapter on random variables largely contains routine material, but there I've covered Markov's inequality, Chebyshev's inequality, and the weak law of large numbers. In my experience, computer science undergraduates find simulation absolutely natural (why do sums when you can write a program?) and enjoy the weak law as a license to do what they would do anyway. You could omit the inequalities and just describe the weak law, though most students run into the inequalities in later theory courses; the experience is usually happier if they've seen them once before.

The chapter on useful probability distributions again largely contains routine material. When I teach this course, I skim through the chapter fairly fast and rely on students reading the chapter. However, there is a detailed discussion of a normal approximation to a binomial distribution with large N . In my experience, no one enjoys the derivation, but you should know the approximation is available, and roughly how it works. I lecture this topic in some detail, mainly by giving examples.

Inference

This part covers:

- Samples and populations
- Confidence intervals for sampled estimates of population means
- Statistical significance, including t-tests, F-tests, and χ^2 -tests
- Very simple experimental design, including one-way and two-way experiments
- ANOVA for experiments
- Maximum likelihood inference
- Simple Bayesian inference
- A very brief discussion of filtering

The material on samples covers only sampling with replacement; if you need something more complicated, this will get you started. Confidence intervals are not much liked by students, I think because the true definition is quite delicate; but getting a grasp of the general idea is useful. You really shouldn't omit these topics.

You shouldn't omit statistical significance either, though you might feel the impulse. I have never dealt with anyone who found their first encounter with statistical significance pleasurable (such a person might exist, the population being very large). But the idea is so useful and so valuable that you just have to take your medicine. Statistical significance is often seen and sometimes taught as a powerful but fundamentally mysterious apotropaic ritual. I try very hard not to do this.

I have often omitted teaching simple experimental design and ANOVA, but in retrospect this was a mistake. The ideas are straightforward and useful. There's a bit of hypocrisy involved in teaching experimental design using other people's datasets. The (correct) alternative is to force students to plan and execute experiments; there just isn't enough time in a usual course to fit this in.

Finally, you shouldn't omit maximum likelihood inference or Bayesian inference. Many people don't need to know about filtering, though.

Tools

This part covers:

- Principal component analysis
- Simple multidimensional scaling with principal coordinate analysis;
- Basic ideas in classification;
- Nearest neighbors classification;
- Naive Bayes classification;
- Classifying with a linear SVM trained with stochastic gradient descent;
- Classifying with a random forest;
- The curse of dimension;
- Agglomerative and divisive clustering;
- K-means clustering;
- Vector quantization;
- A superficial mention of the multivariate normal distribution;
- Linear regression;
- A variety of tricks to analyze and improve regressions;
- Nearest neighbors regression;
- Simple Markov chains;
- Hidden Markov models.

Most students in my institution take this course at the same time they take a linear algebra course. When I teach the course, I try and time things so they hit PCA shortly after hitting eigenvalues and eigenvectors. You shouldn't omit PCA. I lecture principal coordinate analysis very superficially, just describing what it does and why it's useful.

I've been told, often quite forcefully, you can't teach classification to undergraduates. I think you have to, and in my experience, they like it a lot. Students really respond to being taught something that is extremely useful and really easy to do. Please, please, don't omit any of this stuff.

The clustering material is quite simple and easy to teach. In my experience, the topic is a little baffling without an application. I always set a programming exercise where one must build a classifier using features derived from vector quantization. This is a great way of identifying situations where people think they understand something, but don't really. Most students find the exercise challenging, because they must use several concepts together. But most students overcome the challenges and are pleased to see the pieces intermeshing well. The discussion of the multivariate normal distribution is not much more than a mention. I don't think you could omit anything in this chapter.

The regression material is also quite simple and is also easy to teach. The main obstacle here is that students feel something more complicated must necessarily work better (and they're not the only ones). I also don't think you could omit anything in this chapter.

In my experience, computer science students find simple Markov chains natural (though they might find the notation annoying) and will suggest simulating a chain before the instructor does. The examples of using Markov chains to produce natural language (particularly Garkov and wine reviews) are wonderful fun and you really should show them in lectures. You could omit the discussion of ranking the Web. About half of each class I've dealt with has found hidden Markov models easy and natural, and the other half has been wishing the end of the semester was closer. You could omit this topic if you sense likely resistance, and have those who might find it interesting read it.

Mathematical Bits and Pieces

This is a chapter of collected mathematical facts some readers might find useful, together with some slightly deeper information on decision tree construction. Not necessary to lecture this.

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:

- *Elementary Probability*, D. Stirzaker; Cambridge University Press, 2e, 2003.
- *What is a p -value anyway? 34 Stories to Help You Actually Understand Statistics*, A. J. Vickers; Pearson, 2009.
- *Elementary Probability for Applications*, R. Durrett; Cambridge University Press, 2009.
- *Statistics*, D. Freedman, R. Pisani and R. Purves; W. W. Norton & Company, 4e, 2007.
- *Data Analysis and Graphics Using R: An Example-Based Approach*, J. Maindonald and W. J. Braun; Cambridge University Press, 2e, 2003.
- *The Nature of Statistical Learning Theory*, V. Vapnik; Springer, 1999.

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).

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.

Typos spotted by Han Chen (numerous!), Henry Lin (numerous!), Eric Huber, Brian Lunt, Yusuf Sobh, and Scott Walters. Some names might be missing due to poor record-keeping on my part; I apologize. Jian Peng and Paris Smaragdis taught courses from versions of these notes and improved them by detailed comments, suggestions, and typo lists. TAs for this course have helped improve the notes. Thanks to Minje Kim, Henry Lin, Zicheng Liao, Karthik Ramaswamy, Saurabh Singh, Michael Sittig, Nikita Spirin, and Daphne Tsatsoulis. TAs for related classes have also helped improve the notes. Thanks to Tanmay Gangwani, Sili Hui, Ayush Jain, Maghav Kumar, Jiajun Lu, Jason Rock, Daeyun Shin, Mariya Vasileva, and Anirud Yadav.

I have benefited hugely from reviews organized by the publisher. Reviewers made many extremely helpful suggestions, which I have tried to adopt; among many other things, the current material on inference is the product of a complete

overhaul recommended by a reviewer. Reviewers were anonymous to me at time of review, but their names were later revealed so I can thank them by name. Thanks to:

University of Texas, Arlington
University of California, Davis
St. Louis University
University of Tulsa
University of Rhode Island
Morgan State University
University of Texas, Dallas

Dr. Ashis Biswas
Dr. Dipak Ghosal
James Mixco
Sabrina Ripp
Catherine Robinson
Dr. Eric Sakk
Dr. William Semper

Remaining typos, errors, howlers, infelicities, cliché, slang, jargon, cant, platitude, attitude, inaccuracy, fatuousness, etc., are all my fault: Sorry.

Contents

Part I Describing Datasets

1	First Tools for Looking at Data	3
1.1	Datasets	3
1.2	What's Happening? Plotting Data	4
1.2.1	Bar Charts	5
1.2.2	Histograms	6
1.2.3	How to Make Histograms	6
1.2.4	Conditional Histograms	7
1.3	Summarizing 1D Data	7
1.3.1	The Mean	7
1.3.2	Standard Deviation	9
1.3.3	Computing Mean and Standard Deviation Online	12
1.3.4	Variance	12
1.3.5	The Median	13
1.3.6	Interquartile Range	14
1.3.7	Using Summaries Sensibly	15
1.4	Plots and Summaries	16
1.4.1	Some Properties of Histograms	16
1.4.2	Standard Coordinates and Normal Data	18
1.4.3	Box Plots	20
1.5	Whose is Bigger? Investigating Australian Pizzas	20
1.6	You Should	24
1.6.1	Remember These Definitions	24
1.6.2	Remember These Terms	25
1.6.3	Remember These Facts	25
1.6.4	Be Able to	25
2	Looking at Relationships	29
2.1	Plotting 2D Data	29
2.1.1	Categorical Data, Counts, and Charts	29
2.1.2	Series	31
2.1.3	Scatter Plots for Spatial Data	33
2.1.4	Exposing Relationships with Scatter Plots	34
2.2	Correlation	36
2.2.1	The Correlation Coefficient	39
2.2.2	Using Correlation to Predict	42
2.2.3	Confusion Caused by Correlation	44
2.3	Sterile Males in Wild Horse Herds	45
2.4	You Should	47
2.4.1	Remember These Definitions	47
2.4.2	Remember These Terms	47

2.4.3	Remember These Facts	47
2.4.4	Use These Procedures	47
2.4.5	Be Able to	47

Part II Probability

3	Basic Ideas in Probability	53
3.1	Experiments, Outcomes and Probability	53
3.1.1	Outcomes and Probability	53
3.2	Events	55
3.2.1	Computing Event Probabilities by Counting Outcomes	56
3.2.2	The Probability of Events	58
3.2.3	Computing Probabilities by Reasoning About Sets	60
3.3	Independence	61
3.3.1	Example: Airline Overbooking	64
3.4	Conditional Probability	66
3.4.1	Evaluating Conditional Probabilities	67
3.4.2	Detecting Rare Events Is Hard	70
3.4.3	Conditional Probability and Various Forms of Independence	71
3.4.4	Warning Example: The Prosecutor's Fallacy	72
3.4.5	Warning Example: The Monty Hall Problem	73
3.5	Extra Worked Examples	75
3.5.1	Outcomes and Probability	75
3.5.2	Events	76
3.5.3	Independence	77
3.5.4	Conditional Probability	78
3.6	You Should	80
3.6.1	Remember These Definitions	80
3.6.2	Remember These Terms	80
3.6.3	Remember and Use These Facts	80
3.6.4	Remember These Points	80
3.6.5	Be Able to	81
4	Random Variables and Expectations	87
4.1	Random Variables	87
4.1.1	Joint and Conditional Probability for Random Variables	89
4.1.2	Just a Little Continuous Probability	91
4.2	Expectations and Expected Values	93
4.2.1	Expected Values	93
4.2.2	Mean, Variance and Covariance	95
4.2.3	Expectations and Statistics	98
4.3	The Weak Law of Large Numbers	99
4.3.1	IID Samples	99
4.3.2	Two Inequalities	100
4.3.3	Proving the Inequalities	100
4.3.4	The Weak Law of Large Numbers	102
4.4	Using the Weak Law of Large Numbers	103
4.4.1	Should You Accept a Bet?	103
4.4.2	Odds, Expectations and Bookmaking: A Cultural Diversion	104
4.4.3	Ending a Game Early	105
4.4.4	Making a Decision with Decision Trees and Expectations	105
4.4.5	Utility	106

4.5	You Should	108
4.5.1	Remember These Definitions	108
4.5.2	Remember These Terms	108
4.5.3	Use and Remember These Facts	109
4.5.4	Remember These Points	109
4.5.5	Be Able to	109
5	Useful Probability Distributions	115
5.1	Discrete Distributions	115
5.1.1	The Discrete Uniform Distribution	115
5.1.2	Bernoulli Random Variables	116
5.1.3	The Geometric Distribution	116
5.1.4	The Binomial Probability Distribution	116
5.1.5	Multinomial Probabilities	118
5.1.6	The Poisson Distribution	118
5.2	Continuous Distributions	120
5.2.1	The Continuous Uniform Distribution	120
5.2.2	The Beta Distribution	120
5.2.3	The Gamma Distribution	121
5.2.4	The Exponential Distribution	122
5.3	The Normal Distribution	123
5.3.1	The Standard Normal Distribution	123
5.3.2	The Normal Distribution	124
5.3.3	Properties of the Normal Distribution	124
5.4	Approximating Binomials with Large N	126
5.4.1	Large N	127
5.4.2	Getting Normal	128
5.4.3	Using a Normal Approximation to the Binomial Distribution	129
5.5	You Should	130
5.5.1	Remember These Definitions	130
5.5.2	Remember These Terms	130
5.5.3	Remember These Facts	131
5.5.4	Remember These Points	131

Part III Inference

6	Samples and Populations	141
6.1	The Sample Mean	141
6.1.1	The Sample Mean Is an Estimate of the Population Mean	141
6.1.2	The Variance of the Sample Mean	142
6.1.3	When The Urn Model Works	144
6.1.4	Distributions Are Like Populations	145
6.2	Confidence Intervals	146
6.2.1	Constructing Confidence Intervals	146
6.2.2	Estimating the Variance of the Sample Mean	146
6.2.3	The Probability Distribution of the Sample Mean	148
6.2.4	Confidence Intervals for Population Means	149
6.2.5	Standard Error Estimates from Simulation	152
6.3	You Should	154
6.3.1	Remember These Definitions	154
6.3.2	Remember These Terms	154
6.3.3	Remember These Facts	154

6.3.4	Use These Procedures	154
6.3.5	Be Able to	154
7	The Significance of Evidence	159
7.1	Significance	160
7.1.1	Evaluating Significance	160
7.1.2	P-Values	161
7.2	Comparing the Mean of Two Populations	165
7.2.1	Assuming Known Population Standard Deviations	165
7.2.2	Assuming Same, Unknown Population Standard Deviation	167
7.2.3	Assuming Different, Unknown Population Standard Deviation	168
7.3	Other Useful Tests of Significance	169
7.3.1	F-Tests and Standard Deviations	169
7.3.2	χ^2 Tests of Model Fit	171
7.4	P-Value Hacking and Other Dangerous Behavior	174
7.5	You Should	174
7.5.1	Remember These Definitions	174
7.5.2	Remember These Terms	175
7.5.3	Remember These Facts	175
7.5.4	Use These Procedures	175
7.5.5	Be Able to	175
8	Experiments	179
8.1	A Simple Experiment: The Effect of a Treatment	179
8.1.1	Randomized Balanced Experiments	180
8.1.2	Decomposing Error in Predictions	180
8.1.3	Estimating the Noise Variance	181
8.1.4	The ANOVA Table	182
8.1.5	Unbalanced Experiments	183
8.1.6	Significant Differences	185
8.2	Two Factor Experiments	186
8.2.1	Decomposing the Error	188
8.2.2	Interaction Between Effects	189
8.2.3	The Effects of a Treatment	190
8.2.4	Setting Up An ANOVA Table	191
8.3	You Should	194
8.3.1	Remember These Definitions	194
8.3.2	Remember These Terms	194
8.3.3	Remember These Facts	194
8.3.4	Use These Procedures	194
8.3.5	Be Able to	194
9	Inferring Probability Models from Data	197
9.1	Estimating Model Parameters with Maximum Likelihood	197
9.1.1	The Maximum Likelihood Principle	198
9.1.2	Binomial, Geometric and Multinomial Distributions	199
9.1.3	Poisson and Normal Distributions	201
9.1.4	Confidence Intervals for Model Parameters	204
9.1.5	Cautions About Maximum Likelihood	206
9.2	Incorporating Priors with Bayesian Inference	206
9.2.1	Conjugacy	209
9.2.2	MAP Inference	210
9.2.3	Cautions About Bayesian Inference	211

9.3	Bayesian Inference for Normal Distributions	211
9.3.1	Example: Measuring Depth of a Borehole	212
9.3.2	Normal Prior and Normal Likelihood Yield Normal Posterior	212
9.3.3	Filtering	214
9.4	You Should	215
9.4.1	Remember These Definitions	215
9.4.2	Remember These Terms	216
9.4.3	Remember These Facts	216
9.4.4	Use These Procedures	217
9.4.5	Be Able to	217

Part IV Tools

10	Extracting Important Relationships in High Dimensions	225
10.1	Summaries and Simple Plots	225
10.1.1	The Mean	226
10.1.2	Stem Plots and Scatterplot Matrices	226
10.1.3	Covariance	227
10.1.4	The Covariance Matrix	228
10.2	Using Mean and Covariance to Understand High Dimensional Data	231
10.2.1	Mean and Covariance Under Affine Transformations	231
10.2.2	Eigenvectors and Diagonalization	232
10.2.3	Diagonalizing Covariance by Rotating Blobs	233
10.2.4	Approximating Blobs	235
10.2.5	Example: Transforming the Height-Weight Blob	235
10.3	Principal Components Analysis	236
10.3.1	The Low Dimensional Representation	236
10.3.2	The Error Caused by Reducing Dimension	238
10.3.3	Example: Representing Colors with Principal Components	241
10.3.4	Example: Representing Faces with Principal Components	242
10.4	Multi-Dimensional Scaling	242
10.4.1	Choosing Low D Points Using High D Distances	243
10.4.2	Factoring a Dot-Product Matrix	245
10.4.3	Example: Mapping with Multidimensional Scaling	246
10.5	Example: Understanding Height and Weight	247
10.6	You Should	250
10.6.1	Remember These Definitions	250
10.6.2	Remember These Terms	250
10.6.3	Remember These Facts	250
10.6.4	Use These Procedures	250
10.6.5	Be Able to	250
11	Learning to Classify	253
11.1	Classification: The Big Ideas	253
11.1.1	The Error Rate, and Other Summaries of Performance	254
11.1.2	More Detailed Evaluation	254
11.1.3	Overfitting and Cross-Validation	255
11.2	Classifying with Nearest Neighbors	256
11.2.1	Practical Considerations for Nearest Neighbors	256
11.3	Classifying with Naive Bayes	257
11.3.1	Cross-Validation to Choose a Model	259

11.4	The Support Vector Machine	260
11.4.1	The Hinge Loss	261
11.4.2	Regularization	262
11.4.3	Finding a Classifier with Stochastic Gradient Descent	262
11.4.4	Searching for λ	264
11.4.5	Example: Training an SVM with Stochastic Gradient Descent	266
11.4.6	Multi-Class Classification with SVMs	268
11.5	Classifying with Random Forests	268
11.5.1	Building a Decision Tree: General Algorithm	270
11.5.2	Building a Decision Tree: Choosing a Split	270
11.5.3	Forests	272
11.6	You Should	274
11.6.1	Remember These Definitions	274
11.6.2	Remember These Terms	274
11.6.3	Remember These Facts	275
11.6.4	Use These Procedures	275
11.6.5	Be Able to	276
12	Clustering: Models of High Dimensional Data	281
12.1	The Curse of Dimension	281
12.1.1	Minor Banes of Dimension	281
12.1.2	The Curse: Data Isn't Where You Think It Is	282
12.2	Clustering Data	283
12.2.1	Agglomerative and Divisive Clustering	283
12.2.2	Clustering and Distance	285
12.3	The K-Means Algorithm and Variants	287
12.3.1	How to Choose K	288
12.3.2	Soft Assignment	290
12.3.3	Efficient Clustering and Hierarchical K Means	291
12.3.4	K-Medoids	292
12.3.5	Example: Groceries in Portugal	292
12.3.6	General Comments on K-Means	293
12.4	Describing Repetition with Vector Quantization	294
12.4.1	Vector Quantization	296
12.4.2	Example: Activity from Accelerometer Data	298
12.5	The Multivariate Normal Distribution	300
12.5.1	Affine Transformations and Gaussians	301
12.5.2	Plotting a 2D Gaussian: Covariance Ellipses	301
12.6	You Should	302
12.6.1	Remember These Definitions	302
12.6.2	Remember These Terms	302
12.6.3	Remember These Facts	303
12.6.4	Use These Procedures	303
13	Regression	305
13.1	Regression to Make Predictions	305
13.2	Regression to Spot Trends	306
13.3	Linear Regression and Least Squares	308
13.3.1	Linear Regression	308
13.3.2	Choosing β	309
13.3.3	Solving the Least Squares Problem	309
13.3.4	Residuals	310
13.3.5	R-Squared	310

13.4	Producing Good Linear Regressions	313
13.4.1	Transforming Variables	313
13.4.2	Problem Data Points Have Significant Impact	314
13.4.3	Functions of One Explanatory Variable	317
13.4.4	Regularizing Linear Regressions	318
13.5	Exploiting Your Neighbors for Regression	321
13.5.1	Using Your Neighbors to Predict More than a Number	323
13.6	You Should	323
13.6.1	Remember These Definitions	323
13.6.2	Remember These Terms	324
13.6.3	Remember These Facts	324
13.6.4	Remember These Procedures	324
14	Markov Chains and Hidden Markov Models	331
14.1	Markov Chains	331
14.1.1	Transition Probability Matrices	333
14.1.2	Stationary Distributions	335
14.1.3	Example: Markov Chain Models of Text	336
14.2	Estimating Properties of Markov Chains	338
14.2.1	Simulation	338
14.2.2	Simulation Results as Random Variables	339
14.2.3	Simulating Markov Chains	341
14.3	Example: Ranking the Web by Simulating a Markov Chain	342
14.4	Hidden Markov Models and Dynamic Programming	344
14.4.1	Hidden Markov Models	344
14.4.2	Picturing Inference with a Trellis	344
14.4.3	Dynamic Programming for HMM's: Formalities	346
14.4.4	Example: Simple Communication Errors	348
14.5	You Should	349
14.5.1	Remember These Definitions	349
14.5.2	Remember These Terms	349
14.5.3	Remember These Facts	350
14.5.4	Be Able to	350
Part V Mathematical Bits and Pieces		
15	Resources and Extras	355
15.1	Useful Material About Matrices	355
15.1.1	The Singular Value Decomposition	356
15.1.2	Approximating A Symmetric Matrix	356
15.2	Some Special Functions	358
15.3	Splitting a Node in a Decision Tree	359
15.3.1	Accounting for Information with Entropy	359
15.3.2	Choosing a Split with Information Gain	360
Index	363

About the Author

David Forsyth grew up in Cape Town. He received a B.Sc. (Elec. Eng.) from the University of the Witwatersrand, Johannesburg, in 1984, an M.Sc. (Elec. Eng.) from that university in 1986, and a 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 cochair for IEEE Computer Vision and Pattern Recognition in 2000, 2011, and 2018; general cochair for CVPR 2006 and ICCV 2019; and program cochair 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 served six terms on the SIGGRAPH program committee. In 2006, he received an IEEE technical achievement award, in 2009 he was named an IEEE Fellow, and in 2014 he was named an ACM Fellow. 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. Among a variety of odd hobbies, he is a compulsive diver, certified up to normoxic trimix level.

Notation and Conventions

A dataset is a collection of d -tuples (a d -tuple is an ordered list of d elements). Tuples differ from vectors, because we can always add and subtract vectors, but we cannot necessarily add or subtract tuples. There are always N items in any dataset. There are always d elements in each tuple in a dataset. The number of elements will be the same for every tuple in any given tuple. Sometimes we may not know the value of some elements in some tuples.

We use the same notation for a tuple and for a vector. Most of our data will be vectors. We write a vector in bold, so \mathbf{x} could represent a vector or a tuple (the context will make it obvious which is intended).

The entire dataset is $\{x\}$. When we need to refer to the i th data item, we write \mathbf{x}_i . Assume we have N data items, and we wish to make a new dataset out of them; we write the dataset made out of these items as $\{x_i\}$ (the i is to suggest you are taking a set of items and making a dataset out of them). If we need to refer to the j th component of a vector \mathbf{x}_i , we will write $x_i^{(j)}$ (notice this isn't in bold, because it is a component, not a vector, and the j is in parentheses because it isn't a power). Vectors are always column vectors.

When I write $\{kx\}$, I mean the dataset created by taking each element of the dataset $\{x\}$ and multiplying by k ; and when I write $\{x + c\}$, I mean the dataset created by taking each element of the dataset $\{x\}$ and adding c .

Terms

- $\text{mean}(\{x\})$ is the mean of the dataset $\{x\}$ (Definition 1.1, page 7).
- $\text{std}(\{x\})$ is the standard deviation of the dataset $\{x\}$ (Definition 1.2, page 10).
- $\text{var}(\{x\})$ is the standard deviation of the dataset $\{x\}$ (Definition 1.3, page 13).
- $\text{median}(\{x\})$ is the standard deviation of the dataset $\{x\}$ (Definition 1.4, page 13).
- $\text{percentile}(\{x\}, k)$ is the $k\%$ percentile of the dataset $\{x\}$ (Definition 1.5, page 14).
- $\text{iqr}\{x\}$ is the interquartile range of the dataset $\{x\}$ (Definition 1.7, page 15).
- $\{\hat{x}\}$ is the dataset $\{x\}$, transformed to standard coordinates (Definition 1.8, page 18).
- Standard normal data is defined in Definition 18 (page 19).
- Normal data is defined in Definition 1.10 (page 19).
- $\text{corr}(\{(x, y)\})$ is the correlation between two components x and y of a dataset (Definition 2.1, page 39).
- \emptyset is the empty set.
- Ω is the set of all possible outcomes of an experiment.
- Sets are written as \mathcal{A} .
- \mathcal{A}^c is the complement of the set \mathcal{A} (i.e., $\Omega - \mathcal{A}$).
- \mathcal{E} is an event (page 341).
- $P(\{\mathcal{E}\})$ is the probability of event \mathcal{E} (page 341).
- $P(\{\mathcal{E}\}|\{\mathcal{F}\})$ is the probability of event \mathcal{E} , conditioned on event \mathcal{F} (page 341).
- $p(x)$ is the probability that random variable X will take the value x , also written as $P(\{X = x\})$ (page 341).
- $p(x, y)$ is the probability that random variable X will take the value x and random variable Y will take the value y , also written as $P(\{X = x\} \cap \{Y = y\})$ (page 341).
- $\arg\max_x f(x)$ means the value of x that maximizes $f(x)$.
- $\arg\min_x f(x)$ means the value of x that minimizes $f(x)$.
- $\max_i(f(x_i))$ means the largest value that f takes on different elements of the dataset $\{x_i\}$.
- $\hat{\theta}$ is an estimated value of a parameter θ .

Background Information

Cards: A standard deck of playing cards contains 52 cards. These cards are divided into four suits. The suits are spades and clubs (which are black) and hearts and diamonds (which are red). Each suit contains 13 cards: ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, jack (sometimes called knave), queen, and king. It is common to call jack, queen, and king *court cards*.

Dice: If you look hard enough, you can obtain dice with many different numbers of sides (though I've never seen a three-sided die). We adopt the convention that the sides of an N -sided die are labeled with numbers $1 \dots N$ and that no number is used twice. Most dice are like this.

Fairness: Each face of a fair coin or die has the same probability of landing upmost in a flip or roll.

Roulette: A roulette wheel has a collection of slots. There are 36 slots numbered with digits $1 \dots 36$, and then one, two, or even three slots numbered with zero. There are no other slots. Odd-numbered slots are colored red, and even-numbered slots are colored black. Zeros are green. A ball is thrown at the wheel when it is spinning, and it bounces around and eventually falls into a slot. If the wheel is properly balanced, the ball has the same probability of falling into each slot. The number of the slot the ball falls into is said to "come up."