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

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

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.