



# Machine Learning

A Probabilistic Perspective

**Kevin P. Murphy**

# Machine Learning

## A Probabilistic Perspective

Kevin P. Murphy



The MIT Press  
Cambridge, Massachusetts  
London, England

© 2012 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email [special\\_sales@mitpress.mit.edu](mailto:special_sales@mitpress.mit.edu)

This book was set in the L<sup>A</sup>T<sub>E</sub>X programming language by the author. Printed and bound in the United States of America.

#### Library of Congress Cataloging-in-Publication Information

Murphy, Kevin P.  
Machine learning : a probabilistic perspective / Kevin P. Murphy.  
p. cm. — (Adaptive computation and machine learning series)  
Includes bibliographical references and index.  
ISBN 978-0-262-01802-9 (hardcover : alk. paper)  
1. Machine learning. 2. Probabilities. I. Title.  
Q325.5.M87 2012  
006.3'1—dc23  
2012004558

10 9 8 7 6 5 4 3

## Machine Learning: A Probabilistic Perspective

## **Adaptive Computation and Machine Learning**

Thomas Dietterich, Editor

Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns, Associate Editors

*Bioinformatics: The Machine Learning Approach*, Pierre Baldi and Søren Brunak

*Reinforcement Learning: An Introduction*, Richard S. Sutton and Andrew G. Barto

*Graphical Models for Machine Learning and Digital Communication*, Brendan J. Frey

*Learning in Graphical Models*, Michael I. Jordan

*Causation, Prediction, and Search*, second edition, Peter Spirtes, Clark Glymour, and Richard Scheines

*Principles of Data Mining*, David Hand, Heikki Mannila, and Padhraic Smyth

*Bioinformatics: The Machine Learning Approach*, second edition, Pierre Baldi and Søren Brunak

*Learning Kernel Classifiers: Theory and Algorithms*, Ralf Herbrich

*Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Bernhard Schölkopf and Alexander J. Smola

*Introduction to Machine Learning*, Ethem Alpaydin

*Gaussian Processes for Machine Learning*, Carl Edward Rasmussen and Christopher K.I. Williams

*Semi-Supervised Learning*, Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, Eds.

*The Minimum Description Length Principle*, Peter D. GrÅijnwald

*Introduction to Statistical Relational Learning*, Lise Getoor and Ben Taskar, Eds.

*Probabilistic Graphical Models: Principles and Techniques*, Daphne Koller and Nir Friedman

*Introduction to Machine Learning*, second edition, Ethem Alpaydin

*Boosting: Foundations and Algorithms*, Robert E. Schapire and Yoav Freund

*Machine Learning: A Probabilistic Perspective*, Kevin P. Murphy

*Foundations of Machine Learning*, Mehryar Mohri, Afshin Rostami, and Ameet Talwalkar

This book is dedicated to Alessandro, Michael and Stefano,  
and to the memory of Gerard Joseph Murphy.

## Preface

### Introduction

With the ever increasing amounts of data in electronic form, the need for automated methods for data analysis continues to grow. The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest. Machine learning is thus closely related to the fields of statistics and data mining, but differs slightly in terms of its emphasis and terminology. This book provides a detailed introduction to the field, and includes worked examples drawn from application domains such as molecular biology, text processing, computer vision, and robotics.

### Target audience

This book is suitable for upper-level undergraduate students and beginning graduate students in computer science, statistics, electrical engineering, econometrics, or any one else who has the appropriate mathematical background. Specifically, the reader is assumed to already be familiar with basic multivariate calculus, probability, linear algebra, and computer programming. Prior exposure to statistics is helpful but not necessary.

### A probabilistic approach

This book adopts the view that the best way to make machines that can learn from data is to use the tools of probability theory, which has been the mainstay of statistics and engineering for centuries. Probability theory can be applied to any problem involving uncertainty. In machine learning, uncertainty comes in many forms: what is the best prediction (or decision) given some data? what is the best model given some data? what measurement should I perform next? etc.

The systematic application of probabilistic reasoning to all inferential problems, including inferring parameters of statistical models, is sometimes called a Bayesian approach. However, this term tends to elicit very strong reactions (either positive or negative, depending on who you ask), so we prefer the more neutral term “probabilistic approach”. Besides, we will often use techniques such as maximum likelihood estimation, which are not Bayesian methods, but certainly fall within the probabilistic paradigm.

Rather than describing a cookbook of different heuristic methods, this book stresses a principled model-based approach to machine learning. For any given model, a variety of algorithms



can often be applied. Conversely, any given algorithm can often be applied to a variety of models. This kind of modularity, where we distinguish model from algorithm, is good pedagogy and good engineering.

We will often use the language of graphical models to specify our models in a concise and intuitive way. In addition to aiding comprehension, the graph structure aids in developing efficient algorithms, as we will see. However, this book is not primarily about graphical models; it is about probabilistic modeling in general.

## A practical approach

Nearly all of the methods described in this book have been implemented in a MATLAB software package called PMTK, which stands for probabilistic modeling toolkit. This is freely available from [pmtk3.googlecode.com](http://pmtk3.googlecode.com) (the digit 3 refers to the third edition of the toolkit, which is the one used in this version of the book). There are also a variety of supporting files, written by other people, available at [pmtksupport.googlecode.com](http://pmtksupport.googlecode.com). These will be downloaded automatically, if you follow the setup instructions described on the PMTK website.

MATLAB is a high-level, interactive scripting language ideally suited to numerical computation and data visualization, and can be purchased from [www.mathworks.com](http://www.mathworks.com). Some of the code requires the Statistics toolbox, which needs to be purchased separately. There is also a free version of Matlab called Octave, available at <http://www.gnu.org/software/octave/>, which supports most of the functionality of MATLAB. Some (but not all) of the code in this book also works in Octave. See the PMTK website for details.

PMTK was used to generate many of the figures in this book; the source code for these figures is included on the PMTK website, allowing the reader to easily see the effects of changing the data or algorithm or parameter settings. The book refers to files by name, e.g., `naiveBayesFit`. In order to find the corresponding file, you can use two methods: within Matlab you can type `which naiveBayesFit` and it will return the full path to the file; or, if you do not have Matlab but want to read the source code anyway, you can use your favorite search engine, which should return the corresponding file from the [pmtk3.googlecode.com](http://pmtk3.googlecode.com) website.

Details on *how to use* PMTK can be found on the website, which will be updated over time. Details on the *underlying theory* behind these methods can be found in this book.

## Acknowledgments

A book this large is obviously a team effort. I would especially like to thank the following people: my wife Margaret, for keeping the home fires burning as I toiled away in my office for the last six years; Matt Dunham, who created many of the figures in this book, and who wrote much of the code in PMTK; Baback Moghaddam (RIP), who gave extremely detailed feedback on every page of an earlier draft of the book; Chris Williams, who also gave very detailed feedback; Cody Severinski and Wei-Lwun Lu, who assisted with figures; generations of UBC students, who gave helpful comments on earlier drafts; Daphne Koller, Nir Friedman, and Chris Manning, for letting me use their latex style files; Stanford University, Google Research and Skyline College for hosting me during part of my sabbatical; and various Canadian funding agencies (NSERC, CRC and CIFAR) who have supported me financially over the years.

In addition, I would like to thank the following people for giving me helpful feedback on



parts of the book, and/or for sharing figures, code, exercises or even (in some cases) text: David Blei, Sebastien Bratieres, Hannes Bretschneider, Greg Corrado, Arnaud Doucet, Mario Figueiredo, Nando de Freitas, Mark Girolami, Gabriel Goh, Tom Griffiths, Katherine Heller, Geoff Hinton, Aapo Hyvarinen, Tommi Jaakkola, Mike Jordan, Charles Kemp, Emtiyaz Khan, Bonnie Kirkpatrick, Daphne Koller, Zico Kolter, Honglak Lee, Julien Mairal, Andrew McPherson, Tom Minka, Ian Nabney, Robert Piche, Arthur Pope, Carl Rasmussen, Ryan Rifkin, Ruslan Salakhutdinov, Mark Schmidt, Daniel Selsam, David Sontag, Erik Sudderth, Josh Tenenbaum, Martin Wainwright, Yair Weiss, Kai Yu.

Kevin Patrick Murphy  
Palo Alto, California  
June 2012 (Minor updates February 2013)

# Contents

## Preface xxvii

## 1 Introduction 1

- 1.1 Machine learning: what and why? 1
  - 1.1.1 Types of machine learning 2
- 1.2 Supervised learning 3
  - 1.2.1 Classification 3
  - 1.2.2 Regression 8
- 1.3 Unsupervised learning 9
  - 1.3.1 Discovering clusters 10
  - 1.3.2 Discovering latent factors 11
  - 1.3.3 Discovering graph structure 13
  - 1.3.4 Matrix completion 14
- 1.4 Some basic concepts in machine learning 16
  - 1.4.1 Parametric vs non-parametric models 16
  - 1.4.2 A simple non-parametric classifier:  $K$ -nearest neighbors 16
  - 1.4.3 The curse of dimensionality 18
  - 1.4.4 Parametric models for classification and regression 19
  - 1.4.5 Linear regression 19
  - 1.4.6 Logistic regression 21
  - 1.4.7 Overfitting 22
  - 1.4.8 Model selection 22
  - 1.4.9 No free lunch theorem 24

## 2 Probability 27

- 2.1 Introduction 27
- 2.2 A brief review of probability theory 28
  - 2.2.1 Discrete random variables 28
  - 2.2.2 Fundamental rules 29
  - 2.2.3 Bayes' rule 29
  - 2.2.4 Independence and conditional independence 31
  - 2.2.5 Continuous random variables 32

2.2.6	Quantiles	33
2.2.7	Mean and variance	33
2.3	Some common discrete distributions	34
2.3.1	The binomial and Bernoulli distributions	34
2.3.2	The multinomial and multinoulli distributions	35
2.3.3	The Poisson distribution	37
2.3.4	The empirical distribution	37
2.4	Some common continuous distributions	38
2.4.1	Gaussian (normal) distribution	38
2.4.2	Degenerate pdf	39
2.4.3	The Student $t$ distribution	39
2.4.4	The Laplace distribution	41
2.4.5	The gamma distribution	41
2.4.6	The beta distribution	42
2.4.7	Pareto distribution	43
2.5	Joint probability distributions	44
2.5.1	Covariance and correlation	44
2.5.2	The multivariate Gaussian	46
2.5.3	Multivariate Student $t$ distribution	46
2.5.4	Dirichlet distribution	47
2.6	Transformations of random variables	49
2.6.1	Linear transformations	49
2.6.2	General transformations	50
2.6.3	Central limit theorem	51
2.7	Monte Carlo approximation	52
2.7.1	Example: change of variables, the MC way	53
2.7.2	Example: estimating $\pi$ by Monte Carlo integration	54
2.7.3	Accuracy of Monte Carlo approximation	54
2.8	Information theory	56
2.8.1	Entropy	56
2.8.2	KL divergence	57
2.8.3	Mutual information	59
<b>3</b>	<b>Generative models for discrete data</b>	<b>65</b>
3.1	Introduction	65
3.2	Bayesian concept learning	65
3.2.1	Likelihood	67
3.2.2	Prior	67
3.2.3	Posterior	68
3.2.4	Posterior predictive distribution	71
3.2.5	A more complex prior	72
3.3	The beta-binomial model	72
3.3.1	Likelihood	73
3.3.2	Prior	74
3.3.3	Posterior	75

3.3.4	Posterior predictive distribution	77
3.4	The Dirichlet-multinomial model	78
3.4.1	Likelihood	79
3.4.2	Prior	79
3.4.3	Posterior	79
3.4.4	Posterior predictive	81
3.5	Naive Bayes classifiers	82
3.5.1	Model fitting	83
3.5.2	Using the model for prediction	85
3.5.3	The log-sum-exp trick	86
3.5.4	Feature selection using mutual information	86
3.5.5	Classifying documents using bag of words	87
<b>4</b>	<b>Gaussian models</b>	<b>97</b>
4.1	Introduction	97
4.1.1	Notation	97
4.1.2	Basics	97
4.1.3	MLE for an MVN	99
4.1.4	Maximum entropy derivation of the Gaussian *	101
4.2	Gaussian discriminant analysis	101
4.2.1	Quadratic discriminant analysis (QDA)	102
4.2.2	Linear discriminant analysis (LDA)	103
4.2.3	Two-class LDA	104
4.2.4	MLE for discriminant analysis	106
4.2.5	Strategies for preventing overfitting	106
4.2.6	Regularized LDA *	107
4.2.7	Diagonal LDA	108
4.2.8	Nearest shrunken centroids classifier *	109
4.3	Inference in jointly Gaussian distributions	110
4.3.1	Statement of the result	111
4.3.2	Examples	111
4.3.3	Information form	115
4.3.4	Proof of the result *	116
4.4	Linear Gaussian systems	119
4.4.1	Statement of the result	120
4.4.2	Examples	120
4.4.3	Proof of the result *	125
4.5	Digression: The Wishart distribution *	126
4.5.1	Inverse Wishart distribution	127
4.5.2	Visualizing the Wishart distribution *	127
4.6	Inferring the parameters of an MVN	127
4.6.1	Posterior distribution of $\mu$	128
4.6.2	Posterior distribution of $\Sigma$ *	129
4.6.3	Posterior distribution of $\mu$ and $\Sigma$ *	132
4.6.4	Sensor fusion with unknown precisions *	138

<b>5</b>	<b>Bayesian statistics</b>	<b>149</b>
5.1	Introduction	149
5.2	Summarizing posterior distributions	149
5.2.1	MAP estimation	149
5.2.2	Credible intervals	152
5.2.3	Inference for a difference in proportions	154
5.3	Bayesian model selection	155
5.3.1	Bayesian Occam's razor	156
5.3.2	Computing the marginal likelihood (evidence)	158
5.3.3	Bayes factors	163
5.3.4	Jeffreys-Lindley paradox *	164
5.4	Priors	165
5.4.1	Uninformative priors	165
5.4.2	Jeffreys priors *	166
5.4.3	Robust priors	168
5.4.4	Mixtures of conjugate priors	169
5.5	Hierarchical Bayes	171
5.5.1	Example: modeling related cancer rates	171
5.6	Empirical Bayes	172
5.6.1	Example: beta-binomial model	173
5.6.2	Example: Gaussian-Gaussian model	174
5.7	Bayesian decision theory	176
5.7.1	Bayes estimators for common loss functions	177
5.7.2	The false positive vs false negative tradeoff	180
5.7.3	Other topics *	184
<b>6</b>	<b>Frequentist statistics</b>	<b>191</b>
6.1	Introduction	191
6.2	Sampling distribution of an estimator	191
6.2.1	Bootstrap	192
6.2.2	Large sample theory for the MLE *	193
6.3	Frequentist decision theory	195
6.3.1	Bayes risk	195
6.3.2	Minimax risk	196
6.3.3	Admissible estimators	197
6.4	Desirable properties of estimators	200
6.4.1	Consistent estimators	200
6.4.2	Unbiased estimators	201
6.4.3	Minimum variance estimators	201
6.4.4	The bias-variance tradeoff	202
6.5	Empirical risk minimization	205
6.5.1	Regularized risk minimization	206
6.5.2	Structural risk minimization	206
6.5.3	Estimating the risk using cross validation	207
6.5.4	Upper bounding the risk using statistical learning theory *	209

6.5.5	Surrogate loss functions	211
6.6	Pathologies of frequentist statistics *	212
6.6.1	Counter-intuitive behavior of confidence intervals	212
6.6.2	p-values considered harmful	213
6.6.3	The likelihood principle	215
6.6.4	Why isn't everyone a Bayesian?	215
<b>7</b>	<b>Linear regression</b>	<b>217</b>
7.1	Introduction	217
7.2	Model specification	217
7.3	Maximum likelihood estimation (least squares)	217
7.3.1	Derivation of the MLE	219
7.3.2	Geometric interpretation	220
7.3.3	Convexity	221
7.4	Robust linear regression *	223
7.5	Ridge regression	225
7.5.1	Basic idea	225
7.5.2	Numerically stable computation *	227
7.5.3	Connection with PCA *	228
7.5.4	Regularization effects of big data	230
7.6	Bayesian linear regression	231
7.6.1	Computing the posterior	232
7.6.2	Computing the posterior predictive	233
7.6.3	Bayesian inference when $\sigma^2$ is unknown *	234
7.6.4	EB for linear regression (evidence procedure)	238
<b>8</b>	<b>Logistic regression</b>	<b>245</b>
8.1	Introduction	245
8.2	Model specification	245
8.3	Model fitting	245
8.3.1	MLE	246
8.3.2	Steepest descent	247
8.3.3	Newton's method	249
8.3.4	Iteratively reweighted least squares (IRLS)	250
8.3.5	Quasi-Newton (variable metric) methods	251
8.3.6	$\ell_2$ regularization	252
8.3.7	Multi-class logistic regression	252
8.4	Bayesian logistic regression	254
8.4.1	Laplace approximation	255
8.4.2	Derivation of the Bayesian information criterion (BIC)	255
8.4.3	Gaussian approximation for logistic regression	256
8.4.4	Approximating the posterior predictive	256
8.4.5	Residual analysis (outlier detection) *	260
8.5	Online learning and stochastic optimization	261
8.5.1	Online learning and regret minimization	262

8.5.2	Stochastic optimization and risk minimization	262
8.5.3	The LMS algorithm	265
8.5.4	The perceptron algorithm	266
8.5.5	A Bayesian view	267
8.6	Generative vs discriminative classifiers	267
8.6.1	Pros and cons of each approach	268
8.6.2	Dealing with missing data	269
8.6.3	Fisher's linear discriminant analysis (FLDA) *	271
<b>9</b>	<b><i>Generalized linear models and the exponential family</i></b>	<b>281</b>
9.1	Introduction	281
9.2	The exponential family	281
9.2.1	Definition	282
9.2.2	Examples	282
9.2.3	Log partition function	284
9.2.4	MLE for the exponential family	286
9.2.5	Bayes for the exponential family *	287
9.2.6	Maximum entropy derivation of the exponential family *	289
9.3	Generalized linear models (GLMs)	290
9.3.1	Basics	290
9.3.2	ML and MAP estimation	292
9.3.3	Bayesian inference	293
9.4	Probit regression	293
9.4.1	ML/MAP estimation using gradient-based optimization	294
9.4.2	Latent variable interpretation	294
9.4.3	Ordinal probit regression *	295
9.4.4	Multinomial probit models *	295
9.5	Multi-task learning	296
9.5.1	Hierarchical Bayes for multi-task learning	296
9.5.2	Application to personalized email spam filtering	296
9.5.3	Application to domain adaptation	297
9.5.4	Other kinds of prior	297
9.6	Generalized linear mixed models *	298
9.6.1	Example: semi-parametric GLMMs for medical data	298
9.6.2	Computational issues	300
9.7	Learning to rank *	300
9.7.1	The pointwise approach	301
9.7.2	The pairwise approach	301
9.7.3	The listwise approach	302
9.7.4	Loss functions for ranking	303
<b>10</b>	<b><i>Directed graphical models (Bayes nets)</i></b>	<b>307</b>
10.1	Introduction	307
10.1.1	Chain rule	307
10.1.2	Conditional independence	308



10.1.3	Graphical models	308
10.1.4	Graph terminology	309
10.1.5	Directed graphical models	310
10.2	Examples	311
10.2.1	Naive Bayes classifiers	311
10.2.2	Markov and hidden Markov models	312
10.2.3	Medical diagnosis	313
10.2.4	Genetic linkage analysis *	315
10.2.5	Directed Gaussian graphical models *	318
10.3	Inference	319
10.4	Learning	320
10.4.1	Plate notation	320
10.4.2	Learning from complete data	322
10.4.3	Learning with missing and/or latent variables	323
10.5	Conditional independence properties of DGMs	324
10.5.1	d-separation and the Bayes Ball algorithm (global Markov properties)	324
10.5.2	Other Markov properties of DGMs	327
10.5.3	Markov blanket and full conditionals	327
10.6	Influence (decision) diagrams *	328
<b>II</b>	<b>Mixture models and the EM algorithm</b>	<b>337</b>
11.1	Latent variable models	337
11.2	Mixture models	337
11.2.1	Mixtures of Gaussians	339
11.2.2	Mixture of multinoullis	340
11.2.3	Using mixture models for clustering	340
11.2.4	Mixtures of experts	342
11.3	Parameter estimation for mixture models	345
11.3.1	Unidentifiability	346
11.3.2	Computing a MAP estimate is non-convex	347
11.4	The EM algorithm	348
11.4.1	Basic idea	349
11.4.2	EM for GMMs	350
11.4.3	EM for mixture of experts	357
11.4.4	EM for DGMs with hidden variables	358
11.4.5	EM for the Student distribution *	359
11.4.6	EM for probit regression *	362
11.4.7	Theoretical basis for EM *	363
11.4.8	Online EM	365
11.4.9	Other EM variants *	367
11.5	Model selection for latent variable models	370
11.5.1	Model selection for probabilistic models	370
11.5.2	Model selection for non-probabilistic methods	370
11.6	Fitting models with missing data	372

11.6.1	EM for the MLE of an MVN with missing data	373
<b>12</b>	<b>Latent linear models</b>	<b>381</b>
12.1	Factor analysis	381
12.1.1	FA is a low rank parameterization of an MVN	381
12.1.2	Inference of the latent factors	382
12.1.3	Unidentifiability	383
12.1.4	Mixtures of factor analysers	385
12.1.5	EM for factor analysis models	386
12.1.6	Fitting FA models with missing data	387
12.2	Principal components analysis (PCA)	387
12.2.1	Classical PCA: statement of the theorem	387
12.2.2	Proof *	389
12.2.3	Singular value decomposition (SVD)	392
12.2.4	Probabilistic PCA	395
12.2.5	EM algorithm for PCA	396
12.3	Choosing the number of latent dimensions	398
12.3.1	Model selection for FA/PPCA	398
12.3.2	Model selection for PCA	399
12.4	PCA for categorical data	402
12.5	PCA for paired and multi-view data	404
12.5.1	Supervised PCA (latent factor regression)	404
12.5.2	Partial least squares	406
12.5.3	Canonical correlation analysis	407
12.6	Independent Component Analysis (ICA)	407
12.6.1	Maximum likelihood estimation	410
12.6.2	The FastICA algorithm	411
12.6.3	Using EM	414
12.6.4	Other estimation principles *	415
<b>13</b>	<b>Sparse linear models</b>	<b>421</b>
13.1	Introduction	421
13.2	Bayesian variable selection	422
13.2.1	The spike and slab model	424
13.2.2	From the Bernoulli-Gaussian model to $\ell_0$ regularization	425
13.2.3	Algorithms	426
13.3	$\ell_1$ regularization: basics	429
13.3.1	Why does $\ell_1$ regularization yield sparse solutions?	430
13.3.2	Optimality conditions for lasso	431
13.3.3	Comparison of least squares, lasso, ridge and subset selection	435
13.3.4	Regularization path	436
13.3.5	Model selection	439
13.3.6	Bayesian inference for linear models with Laplace priors	440
13.4	$\ell_1$ regularization: algorithms	441
13.4.1	Coordinate descent	441