

Machine Learning: A Probabilistic Perspective

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

This book was set in the \LaTeX 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 2 1

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

Contents

1	Introduction	1
1.1	Machine learning: what and why?	1
1.1.1	Types of machine learning	2
1.2	Supervised learning	2
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	12
1.3.4	Matrix completion	13
1.4	Some basic concepts in machine learning	15
1.4.1	Parametric vs non-parametric models	15
1.4.2	A simple non-parametric classifier: K -nearest neighbors	16
1.4.3	The curse of dimensionality	17
1.4.4	Parametric models for classification and regression	18
1.4.5	Linear regression	19
1.4.6	Logistic regression	20
1.4.7	Overfitting	22
1.4.8	Model selection	22
1.4.9	No free lunch theorem	24
2	Probability	25
2.1	Introduction	25
2.2	A brief review of probability theory	26
2.2.1	Discrete random variables	26
2.2.2	Fundamental rules	26
2.2.3	Bayes rule	27
2.2.4	Independence and conditional independence	28
2.2.5	Continuous random variables	30
2.2.6	Quantiles	31

2.2.7	Mean and variance	31
2.3	Some common discrete distributions	32
2.3.1	The binomial and Bernoulli distributions	32
2.3.2	The multinomial and multinoulli distributions	33
2.3.3	The Poisson distribution	35
2.3.4	The empirical distribution	35
2.4	Some common continuous distributions	36
2.4.1	Gaussian (normal) distribution	36
2.4.2	Degenerate pdf	37
2.4.3	The Student t distribution	37
2.4.4	The Laplace distribution	39
2.4.5	The gamma distribution	39
2.4.6	The beta distribution	40
2.4.7	Pareto distribution	41
2.5	Joint probability distributions	42
2.5.1	Covariance and correlation	42
2.5.2	The multivariate Gaussian	44
2.5.3	Multivariate Student t distribution	44
2.5.4	Dirichlet distribution	45
2.6	Transformations of random variables	47
2.6.1	Linear transformations	47
2.6.2	General transformations	48
2.6.3	Central limit theorem	49
2.7	Monte Carlo approximation	50
2.7.1	Example: change of variables, the MC way	51
2.7.2	Example: estimating π by Monte Carlo integration	52
2.7.3	Accuracy of Monte Carlo approximation	52
2.8	Information theory	54
2.8.1	Entropy	54
2.8.2	KL divergence	55
2.8.3	Mutual information	57
3	<i>Generative models for discrete data</i>	63
3.1	Introduction	63
3.2	Bayesian concept learning	63
3.2.1	Likelihood	65
3.2.2	Prior	65
3.2.3	Posterior	66
3.2.4	Posterior predictive distribution	69
3.2.5	A more complex prior	70
3.3	The Beta-Binomial model	70
3.3.1	Likelihood	71
3.3.2	Prior	72
3.3.3	Posterior	73
3.3.4	Posterior predictive distribution	75

3.4	The Dirichlet-multinomial model	76
3.4.1	Likelihood	77
3.4.2	Prior	77
3.4.3	Posterior	77
3.4.4	Posterior predictive	79
3.5	Naive Bayes classifiers	80
3.5.1	Model fitting	81
3.5.2	Using the model for prediction	83
3.5.3	The log-sum-exp trick	84
3.5.4	Feature selection using mutual information	84
3.5.5	Classifying documents using bag of words	85
4	<i>Gaussian models</i>	95
4.1	Introduction	95
4.1.1	Notation	95
4.1.2	Basics	95
4.1.3	MLE for an MVN	97
4.1.4	Maximum entropy derivation of the Gaussian *	99
4.2	Gaussian Discriminant analysis	99
4.2.1	Quadratic discriminant analysis (QDA)	100
4.2.2	Linear discriminant analysis (LDA)	101
4.2.3	Two-class LDA	102
4.2.4	MLE for discriminant analysis	104
4.2.5	Strategies for preventing overfitting	104
4.2.6	Regularized LDA *	105
4.2.7	Diagonal LDA	106
4.2.8	Nearest shrunken centroids classifier *	107
4.3	Inference in jointly Gaussian distributions	108
4.3.1	Statement of the result	109
4.3.2	Examples	109
4.3.3	Information form	113
4.3.4	Proof of the result *	114
4.4	Linear Gaussian systems	117
4.4.1	Statement of the result	117
4.4.2	Examples	118
4.4.3	Proof of the result *	122
4.5	Digression: The Wishart distribution *	123
4.5.1	Inverse Wishart distribution	124
4.5.2	Visualizing the Wishart distribution *	125
4.6	Inferring the parameters of an MVN	125
4.6.1	Posterior distribution of μ	126
4.6.2	Posterior distribution of Σ *	126
4.6.3	Posterior distribution of μ and Σ *	130
4.6.4	Sensor fusion with unknown precisions *	136

5	<i>Bayesian statistics</i>	147
5.1	Introduction	147
5.2	Summarizing posterior distributions	147
5.2.1	MAP estimation	147
5.2.2	Credible intervals	150
5.2.3	Inference for a difference in proportions	152
5.3	Bayesian model selection	154
5.3.1	Bayesian Occam's razor	154
5.3.2	Computing the marginal likelihood (evidence)	156
5.3.3	Bayes factors	161
5.3.4	Jeffreys-Lindley paradox *	162
5.4	Priors	163
5.4.1	Uninformative priors	163
5.4.2	Jeffreys priors *	164
5.4.3	Robust priors	166
5.4.4	Mixtures of conjugate priors	166
5.5	Hierarchical Bayes	169
5.5.1	Example: modeling related cancer rates	169
5.6	Empirical Bayes	170
5.6.1	Example: Beta-Binomial model	171
5.6.2	Example: Gaussian-Gaussian model	171
5.7	Bayesian decision theory	174
5.7.1	Bayes estimators for common loss functions	175
5.7.2	The false positive vs false negative tradeoff	178
5.7.3	Other topics *	182
6	<i>Frequentist statistics</i>	189
6.1	Introduction	189
6.2	Sampling distribution of an estimator	189
6.2.1	Bootstrap	190
6.2.2	Large sample theory for the MLE *	191
6.3	Frequentist decision theory	192
6.3.1	Bayes risk	193
6.3.2	Minimax risk	194
6.3.3	Admissible estimators	195
6.4	Desirable properties of estimators	198
6.4.1	Consistent estimators	198
6.4.2	Unbiased estimators	198
6.4.3	Minimum variance estimators	199
6.4.4	The bias-variance tradeoff	200
6.5	Empirical risk minimization	202
6.5.1	Regularized risk minimization	203
6.5.2	Structural risk minimization	204
6.5.3	Estimating the risk using cross validation	204
6.5.4	Upper bounding the risk using statistical learning theory *	207

	6.5.5	Surrogate loss functions	208
6.6		Pathologies of frequentist statistics *	209
	6.6.1	Counter-intuitive behavior of confidence intervals	210
	6.6.2	p-values considered harmful	211
	6.6.3	The likelihood principle	212
	6.6.4	Why isn't everyone a Bayesian?	213
7		Linear regression	215
	7.1	Introduction	215
	7.2	Model specification	215
	7.3	Maximum likelihood estimation (least squares)	215
	7.3.1	Derivation of the MLE	217
	7.3.2	Geometric interpretation	218
	7.3.3	Convexity	219
	7.4	Robust linear regression *	221
	7.5	Ridge regression	223
	7.5.1	Basic idea	223
	7.5.2	Numerically stable computation *	225
	7.5.3	Connection with PCA *	226
	7.5.4	Regularization effects of big data	228
	7.6	Bayesian linear regression	229
	7.6.1	Computing the posterior	230
	7.6.2	Computing the posterior predictive	231
	7.6.3	Bayesian inference when σ^2 is unknown *	232
	7.6.4	EB for linear regression (evidence procedure)	236
8		Logistic regression	243
	8.1	Introduction	243
	8.2	Model specification	243
	8.3	Model fitting	243
	8.3.1	MLE	244
	8.3.2	Steepest descent	245
	8.3.3	Newton's method	247
	8.3.4	Iteratively reweighted least squares (IRLS)	248
	8.3.5	Quasi-Newton (variable metric) methods	249
	8.3.6	ℓ_2 regularization	250
	8.3.7	Multi-class logistic regression	250
	8.4	Bayesian logistic regression	252
	8.4.1	Gaussian/ Laplace approximation in general	252
	8.4.2	Derivation of the BIC	253
	8.4.3	Gaussian approximation for logistic regression	254
	8.4.4	Approximating the posterior predictive	255
	8.4.5	Residual analysis (outlier detection) *	258
	8.5	Online learning and stochastic optimization	259
	8.5.1	Online learning and regret minimization	259

8.5.2	Stochastic optimization and risk minimization	260
8.5.3	The LMS algorithm	263
8.5.4	The perceptron algorithm	263
8.5.5	A Bayesian view	264
8.6	Generative vs discriminative classifiers	265
8.6.1	Pros and cons of each approach	265
8.6.2	Dealing with missing data	266
8.6.3	Fisher's linear discriminant analysis (FLDA) *	269
9	<i>Generalized linear models and the exponential family</i>	277
9.1	Introduction	277
9.2	The exponential family	277
9.2.1	Definition	278
9.2.2	Examples	278
9.2.3	Log partition function	280
9.2.4	MLE for the exponential family	282
9.2.5	Bayes for the exponential family *	283
9.2.6	Maximum entropy derivation of the exponential family *	285
9.3	Generalized linear models (GLMs)	286
9.3.1	Basics	286
9.3.2	ML and MAP estimation	288
9.3.3	Bayesian inference	289
9.4	Probit regression	289
9.4.1	ML/ MAP estimation using gradient-based optimization	290
9.4.2	Latent variable interpretation	290
9.4.3	Ordinal probit regression *	291
9.4.4	Multinomial probit models *	291
9.5	Multi-task learning and mixed effect GLMs *	293
9.5.1	Basic model	293
9.5.2	Example: semi-parametric GLMMs for medical data	294
9.5.3	Example: discrete choice modeling	294
9.5.4	Other kinds of prior	295
9.5.5	Computational issues	295
9.6	Learning to rank *	295
9.6.1	The pointwise approach	296
9.6.2	The pairwise approach	297
9.6.3	The listwise approach	297
9.6.4	Loss functions for ranking	298
10	<i>Directed graphical models (Bayes nets)</i>	301
10.1	Introduction	301
10.1.1	Chain rule	301
10.1.2	Conditional independence	302
10.1.3	Graphical models	302
10.1.4	Graph terminology	303

10.1.5	Directed graphical models	304
10.2	Examples	305
10.2.1	Naive Bayes classifiers	305
10.2.2	Markov and hidden Markov models	306
10.2.3	Medical diagnosis	307
10.2.4	Genetic linkage analysis *	309
10.2.5	Directed Gaussian graphical models *	312
10.3	Inference	313
10.4	Learning	314
10.4.1	Plate notation	314
10.4.2	Learning from complete data	316
10.4.3	Learning with missing and/or latent variables	317
10.5	Conditional independence properties of DGMs	318
10.5.1	d-separation and the Bayes Ball algorithm (global Markov properties)	318
10.5.2	Other Markov properties of DGMs	321
10.5.3	Markov blanket and full conditionals	321
10.6	Influence (decision) diagrams *	322
II	Mixture models and the EM algorithm	331
11.1	Latent variable models	331
11.2	Mixture models	331
11.2.1	Mixtures of Gaussians	333
11.2.2	Mixture of multinoullis	334
11.2.3	Using mixture models for clustering	334
11.2.4	Mixtures of experts	336
11.3	Parameter estimation for mixture models	339
11.3.1	Unidentifiability	340
11.3.2	Computing a MAP estimate is non-convex	341
11.4	The EM algorithm	342
11.4.1	Basic idea	343
11.4.2	EM for GMMs	344
11.4.3	EM for mixture of experts	351
11.4.4	EM for DGMs with hidden variables	352
11.4.5	EM for the Student distribution *	353
11.4.6	EM for probit regression *	356
11.4.7	Theoretical basis for EM *	357
11.4.8	Online EM	359
11.4.9	Other EM variants *	361
11.5	Model selection for latent variable models	363
11.5.1	Model selection for probabilistic models	364
11.5.2	Model selection for non-probabilistic methods	364
11.6	Fitting models with missing data	366
11.6.1	EM for the MLE of an MVN with missing data	367

12	<i>Latent linear models</i>	375
12.1	Factor analysis	375
12.1.1	FA is a low rank parameterization of an MVN	375
12.1.2	Inference of the latent factors	376
12.1.3	Unidentifiability	377
12.1.4	Mixtures of factor analysers	379
12.1.5	EM for factor analysis models	380
12.1.6	Fitting FA models with missing data	381
12.2	Principal components analysis (PCA)	381
12.2.1	Classical PCA: statement of the theorem	381
12.2.2	Proof *	383
12.2.3	Singular value decomposition (SVD)	386
12.2.4	Probabilistic PCA	389
12.2.5	EM algorithm for PCA	390
12.3	Choosing the number of latent dimensions	392
12.3.1	Model selection for FA/ PPCA	392
12.3.2	Model selection for PCA	393
12.4	PCA for categorical data	396
12.5	PCA for paired and multi-view data	398
12.5.1	Supervised PCA (latent factor regression)	399
12.5.2	Partial least squares	400
12.5.3	Canonical correlation analysis	401
12.6	Independent Component Analysis (ICA)	401
12.6.1	Maximum likelihood estimation	404
12.6.2	The FastICA algorithm	405
12.6.3	Using EM	408
12.6.4	Other estimation principles *	409
13	<i>Sparse linear models</i>	415
13.1	Introduction	415
13.2	Bayesian variable selection	416
13.2.1	The spike and slab model	418
13.2.2	From the Bernoulli-Gaussian model to ℓ_0 regularization	419
13.2.3	Algorithms	420
13.3	ℓ_1 regularization: basics	423
13.3.1	Why does ℓ_1 regularization yield sparse solutions?	424
13.3.2	Optimality conditions for lasso	425
13.3.3	Comparison of least squares, lasso, ridge and subset selection	429
13.3.4	Regularization path	430
13.3.5	Model selection	433
13.3.6	Bayesian inference for linear models with Laplace priors	434
13.4	ℓ_1 regularization: algorithms	435
13.4.1	Coordinate descent	435
13.4.2	LARS and other homotopy methods	435
13.4.3	Proximal and gradient projection methods	436

13.4.4	EM for lasso	441	
13.5	ℓ_1 regularization: extensions	443	
13.5.1	Group Lasso	443	
13.5.2	Fused lasso	448	
13.5.3	Elastic net (ridge and lasso combined)	449	
13.6	Non-convex regularizers	451	
13.6.1	Bridge regression	452	
13.6.2	Hierarchical adaptive lasso	452	
13.6.3	Other hierarchical priors	456	
13.7	Automatic relevance determination (ARD)/ sparse Bayesian learning (SBL)	457	
13.7.1	ARD for linear regression	457	
13.7.2	Whence sparsity?	459	
13.7.3	Connection to MAP estimation	459	
13.7.4	Algorithms for ARD *	460	
13.7.5	ARD for logistic regression	462	
13.8	Sparse coding *	462	
13.8.1	Learning a sparse coding dictionary	463	
13.8.2	Results of dictionary learning from image patches	464	
13.8.3	Compressed sensing	466	
13.8.4	Image inpainting and denoising	466	
14	Kernels	473	
14.1	Introduction	473	
14.2	Kernel functions	473	
14.2.1	RBF kernels	474	
14.2.2	Kernels for comparing documents	474	
14.2.3	Mercer (positive definite) kernels	475	
14.2.4	Linear kernels	476	
14.2.5	Matern kernels	476	
14.2.6	String kernels	477	
14.2.7	Pyramid match kernels	478	
14.2.8	Kernels derived from probabilistic generative models	479	
14.3	Using kernels inside GLMs	480	
14.3.1	Kernel machines	480	
14.3.2	LIVMs, RVMs, and other sparse kernel machines	481	
14.4	The kernel trick	482	
14.4.1	Kernelized nearest neighbor classification	483	
14.4.2	Kernelized K-medoids clustering	483	
14.4.3	Kernelized ridge regression	486	
14.4.4	Kernel PCA	487	
14.5	Support vector machines (SVMs)	490	
14.5.1	SVMs for regression	491	
14.5.2	SVMs for classification	492	
14.5.3	Choosing C	498	
14.5.4	Summary of key points	498	

14.5.5	A probabilistic interpretation of SVMs	499
14.6	Comparison of discriminative kernel methods	499
14.7	Kernels for building generative models	501
14.7.1	Smoothing kernels	501
14.7.2	Kernel density estimation (KDE)	502
14.7.3	From KDE to KNN	504
14.7.4	Kernel regression	504
14.7.5	Locally weighted regression	506
15	<i>Gaussian processes</i>	509
15.1	Introduction	509
15.2	GPs for regression	510
15.2.1	Predictions using noise-free observations	511
15.2.2	Predictions using noisy observations	512
15.2.3	Effect of the kernel parameters	513
15.2.4	Estimating the kernel parameters	515
15.2.5	Computational and numerical issues *	518
15.2.6	Semi-parametric GPs *	518
15.3	GPs meet GLMs	519
15.3.1	Binary classification	519
15.3.2	Multi-class classification	522
15.3.3	GPs for Poisson regression	525
15.4	Connection with other methods	526
15.4.1	Linear models compared to GPs	526
15.4.2	Linear smoothers compared to GPs	527
15.4.3	SVMs compared to GPs	528
15.4.4	L1VM and RVMs compared to GPs	528
15.4.5	Neural networks compared to GPs	529
15.4.6	Smoothing splines compared to GPs *	530
15.4.7	RKHS methods compared to GPs *	532
15.5	GP latent variable model	534
15.6	Approximation methods for large datasets	536
16	<i>Adaptive basis function models</i>	537
16.1	Introduction	537
16.2	Classification and regression trees (CART)	538
16.2.1	Basics	538
16.2.2	Growing a tree	540
16.2.3	Pruning a tree	543
16.2.4	Pros and cons of trees	544
16.2.5	Random forests	545
16.2.6	CART compared to hierarchical mixture of experts *	545
16.3	Generalized additive models	546
16.3.1	Backfitting	546
16.3.2	Computational efficiency	547

16.3.3	Multivariate adaptive regression splines (MARS)	547
16.4	Boosting	548
16.4.1	Forward stagewise additive modeling	549
16.4.2	L2boosting	552
16.4.3	AdaBoost	552
16.4.4	LogitBoost	554
16.4.5	Boosting as functional gradient descent	554
16.4.6	Sparse boosting	556
16.4.7	Multivariate adaptive regression trees (MART)	556
16.4.8	Why does boosting work so well?	557
16.4.9	A Bayesian view	557
16.5	Feedforward neural networks (multilayer perceptrons)	558
16.5.1	Convolutional neural networks	559
16.5.2	Other kinds of neural networks	562
16.5.3	A brief history of the field	563
16.5.4	The backpropagation algorithm	564
16.5.5	Identifiability	566
16.5.6	Regularization	566
16.5.7	Bayesian inference *	570
16.6	Ensemble learning	574
16.6.1	Stacking	574
16.6.2	Error-correcting output codes	575
16.6.3	Ensemble learning is not equivalent to Bayes model averaging	575
16.7	Experimental comparison	576
16.7.1	Low-dimensional features	576
16.7.2	High-dimensional features	577
16.8	Interpreting black-box models	579
17	Markov and hidden Markov Models	583
17.1	Introduction	583
17.2	Markov models	583
17.2.1	Transition matrix	583
17.2.2	Application: Language modeling	585
17.2.3	Stationary distribution of a Markov chain *	590
17.2.4	Application: Google's PageRank algorithm for web page ranking *	594
17.3	Hidden Markov models	597
17.3.1	Applications of HMMs	598
17.4	Inference in HMMs	600
17.4.1	Types of inference problems for temporal models	600
17.4.2	The forwards algorithm	603
17.4.3	The forwards-backwards algorithm	604
17.4.4	The Viterbi algorithm	606
17.4.5	Forwards filtering, backwards sampling	610
17.5	Learning for HMMs	611
17.5.1	Training with fully observed data	611

17.5.2	EM for HMMs (the Baum-Welch algorithm)	612
17.5.3	Bayesian methods for “fitting” HMMs *	614
17.5.4	Discriminative training	614
17.5.5	Model selection	615
17.6	Generalizations of HMMs	615
17.6.1	Variable duration (semi-Markov) HMMs	616
17.6.2	Hierarchical HMMs	618
17.6.3	Input-output HMMs	619
17.6.4	Auto-regressive and buried HMMs	620
17.6.5	Factorial HMM	621
17.6.6	Coupled HMM and the influence model	622
17.6.7	Dynamic Bayesian networks (DBNs)	622
18	<i>State space models</i>	625
18.1	Introduction	625
18.2	Applications of SSMs	626
18.2.1	SSMs for object tracking	626
18.2.2	Robotic SLAM	627
18.2.3	Online parameter learning using recursive least squares	630
18.2.4	SSM for time series forecasting *	631
18.3	Inference in LG-SSM	634
18.3.1	The Kalman filtering algorithm	634
18.3.2	The Kalman smoothing algorithm	637
18.4	Learning for LG-SSM	640
18.4.1	Identifiability and numerical stability	640
18.4.2	Training with fully observed data	641
18.4.3	EM for LG-SSM	641
18.4.4	Subspace methods	641
18.4.5	Bayesian methods for “fitting” LG-SSMs	641
18.5	Approximate online inference for non-linear, non-Gaussian SSMs	641
18.5.1	Extended Kalman filter (EKF)	642
18.5.2	Unscented Kalman filter (UKF)	644
18.5.3	Assumed density filtering (ADF)	646
18.6	Hybrid discrete/ continuous SSMs	649
18.6.1	Inference	650
18.6.2	Application: Data association and multi target tracking	652
18.6.3	Application: fault diagnosis	653
18.6.4	Application: econometric forecasting	654
19	<i>Undirected graphical models (Markov random fields)</i>	655
19.1	Introduction	655
19.2	Conditional independence properties of UGMs	655
19.2.1	Key properties	655
19.2.2	An undirected alternative to d-separation	657
19.2.3	Comparing directed and undirected graphical models	658

19.3	Parameterization of MRFs	659	
19.3.1	The Hammersley-Clifford theorem	659	
19.3.2	Representing potential functions	661	
19.4	Examples of MRFs	662	
19.4.1	Ising model	662	
19.4.2	Hopfield networks	663	
19.4.3	Potts model	665	
19.4.4	Gaussian MRFs	666	
19.4.5	Markov logic networks *	668	
19.5	Learning	670	
19.5.1	Training maxent models using gradient methods	670	
19.5.2	Training partially observed maxent models	671	
19.5.3	Approximate methods for computing the MLEs of MRFs	672	
19.5.4	Pseudo likelihood	672	
19.5.5	Stochastic Maximum Likelihood	673	
19.5.6	Feature induction for maxent models *	674	
19.5.7	Iterative proportional fitting (IPF) *	675	
19.6	Conditional random fields (CRFs)	678	
19.6.1	Chain-structured CRFs, MEMMs and the label-bias problem	678	
19.7	Applications of CRFs	680	
19.7.1	Handwriting recognition	680	
19.7.2	Noun phrase chunking	681	
19.7.3	Named entity recognition	682	
19.7.4	CRFs for protein side-chain prediction	682	
19.7.5	Stereo vision	683	
19.8	CRF training	685	
19.9	Max margin methods for structured output classifiers *	686	
20	<i>Exact inference for graphical models</i>	689	
20.1	Introduction	689	
20.2	Belief propagation for trees	689	
20.2.1	Serial protocol	689	
20.2.2	Parallel protocol	691	
20.2.3	Gaussian BP *	692	
20.2.4	Other BP variants *	694	
20.3	The variable elimination algorithm	696	
20.3.1	The generalized distributive law *	699	
20.3.2	Computational complexity of VE	699	
20.3.3	A weakness of VE	702	
20.4	The junction tree algorithm *	702	
20.4.1	Creating a junction tree	702	
20.4.2	Message passing on a junction tree	704	
20.4.3	Computational complexity of JTA	707	
20.4.4	JTA generalizations *	708	
20.5	Computational intractability of exact inference in the worst case	708	

20.5.1	Approximate inference	709	
21	<i>Variational inference</i>	713	
21.1	Introduction	713	
21.2	Variational inference	714	
21.2.1	Alternative interpretations of the variational objective	715	
21.2.2	Forward or reverse KL? *	715	
21.3	The mean field method	717	
21.3.1	Derivation of the mean field update equations	718	
21.3.2	Example: Mean field for the Ising model	719	
21.4	Structured mean field *	721	
21.4.1	Example: factorial HMM	722	
21.5	Variational Bayes	724	
21.5.1	Example: VB for a univariate Gaussian	724	
21.5.2	Example: VB for linear regression	728	
21.6	Variational Bayes EM	731	
21.6.1	Example: VBEM for mixtures of Gaussians *	732	
21.7	Variational message passing and VIBES	738	
21.8	Local variational bounds *	738	
21.8.1	Motivating applications	738	
21.8.2	Bohning's quadratic bound to the log-sum-exp function	740	
21.8.3	Bounds for the sigmoid function	742	
21.8.4	Other bounds and approximations to the log-sum-exp function *	744	
21.8.5	Variational inference based on upper bounds	745	
22	<i>More variational inference</i>	749	
22.1	Introduction	749	
22.2	Loopy belief propagation: algorithmic issues	749	
22.2.1	A brief history	749	
22.2.2	LBP on pairwise models	750	
22.2.3	LBP on a factor graph	751	
22.2.4	Convergence	753	
22.2.5	Accuracy of LBP	756	
22.2.6	Other speedup tricks for BP *	757	
22.3	Loopy belief propagation: theoretical issues *	758	
22.3.1	UGMs represented in exponential family form	758	
22.3.2	The marginal polytope	759	
22.3.3	Exact inference as a variational optimization problem	760	
22.3.4	Mean field as a variational optimization problem	761	
22.3.5	LBP as a variational optimization problem	761	
22.3.6	Loopy BP vs mean field	765	
22.4	Extensions of belief propagation *	765	
22.4.1	Generalized belief propagation	765	
22.4.2	Convex belief propagation	767	
22.5	Expectation propagation	769	

22.5.1	EP as a variational inference problem	770	
22.5.2	Optimizing the EP objective using moment matching		771
22.5.3	EP for the clutter problem	773	
22.5.4	LBP is a special case of EP	774	
22.5.5	Ranking players using TrueSkill	775	
22.5.6	Other applications	781	
22.6	MAP state estimation	781	
22.6.1	Linear programming relaxation	781	
22.6.2	Max-product belief propagation	782	
22.6.3	Graphcuts	783	
22.6.4	Experimental comparison of graphcuts and BP		786
22.6.5	Dual decomposition	788	
23	Monte Carlo inference	795	
23.1	Introduction	795	
23.2	Sampling from standard distributions	795	
23.2.1	Using the cdf	795	
23.2.2	Sampling from a Gaussian (Box-Muller method)		797
23.3	Rejection sampling	797	
23.3.1	Basic idea	797	
23.3.2	Example	798	
23.3.3	Application to Bayesian statistics	799	
23.3.4	Adaptive rejection sampling	799	
23.3.5	Rejection sampling in high dimensions	800	
23.4	Importance sampling	800	
23.4.1	Basic idea	800	
23.4.2	Handling unnormalized distributions	801	
23.4.3	Importance sampling for a DGM: Likelihood weighting		802
23.4.4	Sampling importance resampling (SIR)	802	
23.5	Particle filtering	803	
23.5.1	Sequential importance sampling	804	
23.5.2	The degeneracy problem	805	
23.5.3	The resampling step	805	
23.5.4	The proposal distribution	807	
23.5.5	Application: Robot localization	808	
23.5.6	Application: Visual object tracking	808	
23.5.7	Application: time series forecasting	811	
23.6	Rao-Blackwellised particle filtering (RBPF)	811	
23.6.1	RBPF for switching LG-SSMs	811	
23.6.2	Application: Tracking a maneuvering target		812
23.6.3	Application: Fast SLAM	814	
24	Markov Chain Monte Carlo (MCMC) inference	817	
24.1	Introduction	817	
24.2	Gibbs sampling	818	

24.2.1	Basic idea	818	
24.2.2	Example: Gibbs sampling for the Ising model	818	
24.2.3	Example: Gibbs sampling for inferring the parameters of a GMM	820	
24.2.4	Collapsed Gibbs sampling *	821	
24.2.5	Gibbs sampling for hierarchical GLMs	824	
24.2.6	BUGS and JAGS	826	
24.2.7	The Imputation Posterior (IP) algorithm	827	
24.2.8	Blocking Gibbs sampling	827	
24.3	Metropolis Hastings algorithm	828	
24.3.1	Basic idea	828	
24.3.2	Gibbs sampling is a special case of MH	829	
24.3.3	Proposal distributions	830	
24.3.4	Adaptive MCMC	833	
24.3.5	Initialization and mode hopping	834	
24.3.6	Why MH works *	834	
24.3.7	Reversible jump (trans-dimensional) MCMC *	835	
24.4	Speed and accuracy of MCMC	836	
24.4.1	The burn-in phase	836	
24.4.2	Mixing rates of Markov chains *	837	
24.4.3	Practical convergence diagnostics	838	
24.4.4	Accuracy of MCMC	840	
24.4.5	How many chains?	842	
24.5	Auxiliary variable MCMC *	843	
24.5.1	Auxiliary variable sampling for logistic regression	843	
24.5.2	Slice sampling	844	
24.5.3	Swendsen Wang	846	
24.5.4	Hybrid/ Hamiltonian MCMC *	848	
24.6	Annealing methods	848	
24.6.1	Simulated annealing	849	
24.6.2	Annealed importance sampling	851	
24.6.3	Parallel tempering	851	
24.7	Approximating the marginal likelihood	852	
24.7.1	The candidate method	852	
24.7.2	Harmonic mean estimate	852	
24.7.3	Annealed importance sampling	853	
25	Clustering	855	
25.1	Introduction	855	
25.1.1	Measuring (dis)similarity	855	
25.1.2	Evaluating the output of clustering methods *	856	
25.2	Dirichlet process mixture models	859	
25.2.1	From finite to infinite mixture models	859	
25.2.2	The Dirichlet process	862	
25.2.3	Applying Dirichlet processes to mixture modeling	865	
25.2.4	Fitting a DP mixture model	866	

25.3	Affinity propagation	867
25.4	Spectral clustering	870
25.4.1	Graph Laplacian	871
25.4.2	Normalized graph Laplacian	872
25.4.3	Example	873
25.5	Hierarchical clustering	873
25.5.1	Agglomerative clustering	875
25.5.2	Divisive clustering	878
25.5.3	Choosing the number of clusters	879
25.5.4	Bayesian hierarchical clustering	879
25.6	Clustering datapoints and features	881
25.6.1	Biclustering	883
25.6.2	Multi-view clustering	883
26	Graphical model structure learning	887
26.1	Introduction	887
26.2	Quick and dirty ways to learn graph structure	888
26.2.1	Relevance networks	888
26.2.2	Dependency networks	889
26.3	Learning tree structures	890
26.3.1	Directed or undirected tree?	891
26.3.2	Chow-Liu algorithm for finding the ML tree structure	892
26.3.3	Finding the MAP forest	892
26.3.4	Mixtures of trees	894
26.4	Learning DAG structures	894
26.4.1	Exact structural inference	894
26.4.2	Scaling up to larger graphs	900
26.5	Learning DAG structure with latent variables	902
26.5.1	Approximating the marginal likelihood when we have missing data	902
26.5.2	Structural EM	905
26.5.3	Discovering hidden variables	905
26.5.4	Case study: Google's Rephil	908
26.5.5	Structural equation models *	909
26.6	Learning causal DAGs	911
26.6.1	Causal interpretation of DAGs	911
26.6.2	Using causal DAGs to resolve Simpson's paradox	912
26.6.3	Learning causal DAG structures	915
26.7	Learning undirected Gaussian graphical models	918
26.7.1	MLE for a GRF	918
26.7.2	Graphical lasso	919
26.7.3	Bayesian inference for GRF structure	921
26.7.4	Handling non-Gaussian data *	923
26.8	Learning undirected discrete graphical models	923
26.8.1	Graphical lasso for MRFs/ CRFs	923
26.8.2	Thin junction trees	924

27	<i>Latent variable models for discrete data</i>	927
27.1	Introduction	927
27.2	Distributed state LVMs for discrete data	928
27.2.1	Mixture models	928
27.2.2	Exponential family PCA	929
27.2.3	LDA and mPCA	930
27.2.4	GaP model and non-negative matrix factorization	931
27.3	Latent Dirichlet allocation (LDA)	932
27.3.1	Basics	932
27.3.2	Unsupervised discovery of topics	935
27.3.3	Quantitatively evaluating LDA as a language model	935
27.3.4	Fitting using (collapsed) Gibbs sampling	937
27.3.5	Example	938
27.3.6	Fitting using batch variational inference	939
27.3.7	Fitting using online variational inference	941
27.3.8	Determining the number of topics	942
27.4	Extensions of LDA	943
27.4.1	Correlated topic model	943
27.4.2	Dynamic topic model	944
27.4.3	LDA-HMM	945
27.4.4	Supervised LDA	949
27.5	LVMs for graph-structured data	952
27.5.1	Stochastic block model	953
27.5.2	Mixed membership stochastic block model	955
27.5.3	Relational topic model	956
27.6	LVMs for relational data	957
27.6.1	Infinite relational model	958
27.6.2	Probabilistic matrix factorization for collaborative filtering	961
27.7	Restricted Boltzmann machines (RBMs)	965
27.7.1	Varieties of RBMs	967
27.7.2	Learning RBMs	969
27.7.3	Applications of RBMs	973
28	<i>Deep learning</i>	977
28.1	Introduction	977
28.2	Deep generative models	978
28.2.1	Deep sigmoid networks	978
28.2.2	Deep Boltzmann machines	979
28.2.3	Deep belief networks	980
28.3	Training deep networks	981
28.3.1	Greedy layer-wise learning of DBNs	981
28.3.2	Fitting deep neural nets	983
28.3.3	Fitting deep auto-encoders	983
28.3.4	Stacked denoising auto-encoders	984
28.4	Applications of deep networks	984

28.4.1	Handwritten digit classification using DBNs	984	
28.4.2	Data visualization using deep auto-encoders	986	
28.4.3	Information retrieval using deep autoencoders (semantic hashing)		986
28.4.4	Learning audio features using 1d convolutional DBNs	987	
28.4.5	Learning image features using 2d convolutional DBNs	988	
28.5	Discussion	989	

Bibliography **991**

Index to code	1021
Index to keywords	1025

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 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` (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`.

MATLAB is a high-level, interactive scripting language ideally suited to numerical computation and data visualization, and can be purchased from `www.mathworks.com`. (Additional toolboxes, such as the Statistics toolbox, can be purchased, too; we have tried to minimize our dependence on this toolbox, but it is nevertheless very useful to have.) 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 (see the PMTK website for a comparison).

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` website.

Details on *how to use* PMTK can be found on the PMTK 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, 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,

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, Tom Minka, Ian Nabney, Arthur Pope, Carl Rasmussen, Ryan Rifkin, Ruslan Salakhutdinov, Mark Schmidt, David Sontag, Erik Sudderth, Josh Tenenbaum, Kai Yu, Martin Wainwright, Yair Weiss.

Kevin Murphy
Palo Alto, California
March 2012