

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	14
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	Proof of the result *	113
4.4	Linear Gaussian systems	116
4.4.1	Statement of the result	117
4.4.2	Examples	117
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 *	124
4.6	Inferring the parameters of an MVN	124
4.6.1	Posterior distribution of μ	125
4.6.2	Posterior distribution of Σ *	126
4.6.3	Posterior distribution of μ and Σ *	129
4.6.4	Sensor fusion with unknown precisions *	133
5	<i>Bayesian statistics</i>	143

5.1	Introduction	143	
5.2	Summarizing posterior distributions	143	
5.2.1	MAP estimation	143	
5.2.2	Credible intervals	146	
5.2.3	Inference for a difference in proportions	148	
5.3	Bayesian model selection	150	
5.3.1	Bayesian Occam's razor	150	
5.3.2	Computing the marginal likelihood (evidence)	152	
5.3.3	Bayes factors	157	
5.3.4	Jeffreys-Lindley paradox *	158	
5.4	Priors	159	
5.4.1	Uninformative priors	159	
5.4.2	Jeffreys priors *	160	
5.4.3	Robust priors	162	
5.4.4	Mixtures of conjugate priors	162	
5.5	Hierarchical Bayes	165	
5.5.1	Example: modeling related cancer rates	165	
5.6	Empirical Bayes	166	
5.6.1	Example: Beta-Binomial model	167	
5.6.2	Example: Gaussian-Gaussian model	167	
5.7	Bayesian decision theory	170	
5.7.1	Bayes estimators for common loss functions	171	
5.7.2	The false positive vs false negative tradeoff	174	
5.7.3	More general action spaces	178	
5.7.4	Sequential decision theory	179	
6	<i>Frequentist statistics</i>	183	
6.1	Introduction	183	
6.2	Sampling distribution of an estimator	183	
6.2.1	Bootstrap	184	
6.2.2	Large sample theory for the MLE *	185	
6.2.3	Connection with Bayesian statistics *	186	
6.3	Frequentist decision theory	187	
6.3.1	Bayes risk	188	
6.3.2	Minimax risk	189	
6.3.3	Admissible estimators	190	
6.4	Desirable properties of estimators	193	
6.4.1	Consistent estimators	193	
6.4.2	Unbiased estimators	193	
6.4.3	Minimum variance estimators	194	
6.4.4	The bias-variance tradeoff	195	
6.5	Empirical risk minimization	197	
6.5.1	Regularized risk minimization	198	
6.5.2	Structural risk minimization	199	
6.5.3	Estimating the risk using cross validation	199	

6.5.4	Upper bounding the risk using statistical learning theory *	202
6.5.5	Surrogate loss functions	203
6.6	Pathologies of frequentist statistics *	204
6.6.1	Counter-intuitive behavior of confidence intervals	205
6.6.2	p-values considered harmful *	206
6.6.3	The likelihood principle	207
6.6.4	Why isn't everyone a Bayesian?	208
7	Linear regression	211
7.1	Introduction	211
7.2	Model specification	211
7.3	Maximum likelihood estimation (least squares)	211
7.3.1	Derivation of the MLE	213
7.3.2	Geometric interpretation	214
7.3.3	Convexity	215
7.4	Robust linear regression *	217
7.5	Ridge regression	219
7.5.1	Basic idea	219
7.5.2	Numerically stable computation *	221
7.5.3	Connection with PCA *	222
7.5.4	Regularization effects of big data	224
7.6	Bayesian linear regression	225
7.6.1	Computing the posterior	226
7.6.2	Computing the posterior predictive	227
7.6.3	Bayesian inference when σ^2 is unknown *	228
7.6.4	EB for linear regression (evidence procedure)	232
8	Logistic regression	239
8.1	Introduction	239
8.2	Model specification	239
8.3	Model fitting	239
8.3.1	MLE	240
8.3.2	Steepest descent	241
8.3.3	Newton's method	243
8.3.4	Iteratively reweighted least squares (IRLS)	244
8.3.5	Quasi-Newton (variable metric) methods	245
8.3.6	ℓ_2 regularization	246
8.3.7	Multi-class logistic regression	246
8.4	Bayesian logistic regression	248
8.4.1	Gaussian/ Laplace approximation in general	248
8.4.2	Derivation of the BIC	249
8.4.3	Gaussian approximation for logistic regression	249
8.4.4	Approximating the posterior predictive	251
8.4.5	Residual analysis (outlier detection) *	254
8.5	Online learning and stochastic optimization	255

8.5.1	Online learning and regret minimization	255
8.5.2	Stochastic optimization and risk minimization	256
8.5.3	The LMS algorithm	259
8.5.4	The perceptron algorithm	259
8.5.5	A Bayesian view	260
8.6	Generative vs discriminative classifiers	261
8.6.1	Pros and cons of each approach	261
8.6.2	Dealing with missing data	262
8.6.3	Fisher's linear discriminant analysis (FLDA) *	265
9	<i>Generalized linear models and the exponential family</i>	273
9.1	Introduction	273
9.2	The exponential family	273
9.2.1	Definition	274
9.2.2	Examples	274
9.2.3	Log partition function	276
9.2.4	MLE for the exponential family	278
9.2.5	Bayes for the exponential family *	279
9.2.6	Maximum entropy derivation of the exponential family *	281
9.3	Generalized linear models (GLMs)	282
9.3.1	Basics	282
9.3.2	ML and MAP estimation	284
9.3.3	Bayesian inference	285
9.4	Probit regression	285
9.4.1	ML/ MAP estimation using gradient-based optimization	286
9.4.2	Latent variable interpretation	286
9.4.3	Ordinal probit regression *	287
9.4.4	Multinomial probit models *	287
9.5	Multi-task learning and mixed effect GLMs *	289
9.5.1	Basic model	289
9.5.2	Example: semi-parametric GLMMs for medical data	290
9.5.3	Example: discrete choice modeling	290
9.5.4	Other kinds of prior	291
9.5.5	Computational issues	291
10	<i>Directed graphical models (Bayes nets)</i>	293
10.1	Introduction	293
10.1.1	Chain rule	293
10.1.2	Conditional independence	294
10.1.3	Graphical models	294
10.1.4	Graph terminology	295
10.1.5	Directed graphical models	296
10.2	Examples	297
10.2.1	Naive Bayes classifiers	297
10.2.2	Markov and hidden Markov models	298

10.2.3	Medical diagnosis	299
10.2.4	Genetic linkage analysis *	301
10.2.5	Directed Gaussian graphical models *	304
10.3	Inference	305
10.4	Learning	306
10.4.1	Plate notation	306
10.4.2	Learning from complete data	308
10.4.3	Learning with missing and/or latent variables	309
10.5	Conditional independence properties of DGMs	310
10.5.1	d-separation and the Bayes Ball algorithm (global Markov properties)	310
10.5.2	Other Markov properties of DGMs	313
10.5.3	Markov blanket and full conditionals	313
10.6	Influence (decision) diagrams *	314
11	<i>Mixture models and the EM algorithm</i>	321
11.1	Latent variable models	321
11.2	Mixture models	321
11.2.1	Mixtures of Gaussians	323
11.2.2	Mixture of multinoullis	324
11.2.3	Using mixture models for clustering	324
11.2.4	Mixtures of experts	326
11.3	Parameter estimation for mixture models	329
11.3.1	Unidentifiability	330
11.3.2	Computing a MAP estimate is non-convex	331
11.4	The EM algorithm	332
11.4.1	Basic idea	333
11.4.2	EM for GMMs	334
11.4.3	EM for mixture of experts	341
11.4.4	EM for DGMs with hidden variables	342
11.4.5	EM for the Student distribution *	343
11.4.6	EM for probit regression *	346
11.4.7	Theoretical basis for EM *	347
11.4.8	EM variants *	349
11.5	Model selection for latent variable models	351
11.5.1	Model selection for probabilistic models	352
11.5.2	Model selection for non-probabilistic methods	352
11.6	Fitting models with missing data	354
11.6.1	EM for the MLE of an MVN with missing data	355
12	<i>Latent linear models</i>	363
12.1	Factor analysis	363
12.1.1	FA is a low rank parameterization of an MVN	363
12.1.2	Inference of the latent factors	364
12.1.3	Unidentifiability	365

12.1.4	Mixtures of factor analysers	367
12.1.5	EM for factor analysis models	368
12.1.6	Fitting FA models with missing data	369
12.2	Principal components analysis (PCA)	369
12.2.1	Classical PCA: statement of the theorem	369
12.2.2	Proof *	371
12.2.3	Singular value decomposition (SVD)	374
12.2.4	Probabilistic PCA	377
12.2.5	EM algorithm for PCA	378
12.3	Choosing the number of latent dimensions	380
12.3.1	Model selection for FA/ PPCA	380
12.3.2	Model selection for PCA	381
12.4	PCA for categorical data	384
12.5	PCA for paired and multi-view data	386
12.5.1	Supervised PCA (latent factor regression)	387
12.5.2	Partial least squares	388
12.5.3	Canonical correlation analysis	389
12.6	Independent Component Analysis (ICA)	389
12.6.1	Maximum likelihood estimation	392
12.6.2	The FastICA algorithm	393
12.6.3	Using EM	396
12.6.4	Other estimation principles *	397

13 Sparse linear models 403

13.1	Introduction	403
13.2	Bayesian variable selection	404
13.2.1	The spike and slab model	406
13.2.2	From the Bernoulli-Gaussian model to ℓ_0 regularization	407
13.2.3	Algorithms	408
13.3	ℓ_1 regularization: basics	411
13.3.1	Why does ℓ_1 regularization yield sparse solutions?	412
13.3.2	Optimality conditions for lasso	413
13.3.3	Comparison of least squares, lasso, ridge and subset selection	417
13.3.4	Regularization path	418
13.3.5	Model selection	421
13.3.6	Bayesian inference for linear models with Laplace priors	422
13.4	ℓ_1 regularization: algorithms	423
13.4.1	Coordinate descent	423
13.4.2	LARS and other homotopy methods	423
13.4.3	Proximal and gradient projection methods	424
13.4.4	EM for lasso	429
13.5	ℓ_1 regularization: extensions	431
13.5.1	Group Lasso	431
13.5.2	Fused lasso	436
13.5.3	Elastic net (ridge and lasso combined)	437

13.6	Non-convex regularizers	439	
13.6.1	Bridge regression	440	
13.6.2	Hierarchical adaptive lasso	440	
13.6.3	Other hierarchical priors	444	
13.7	Automatic relevance determination (ARD)/ sparse Bayesian learning (SBL)		445
13.7.1	ARD for linear regression	445	
13.7.2	Whence sparsity?	447	
13.7.3	Connection to MAP estimation	447	
13.7.4	Algorithms for ARD *	448	
13.7.5	ARD for logistic regression	450	
13.8	Sparse coding *	450	
13.8.1	Learning a sparse coding dictionary	451	
13.8.2	Results of dictionary learning from image patches		452
13.8.3	Compressed sensing	454	
13.8.4	Image inpainting and denoising	454	

14 Kernels 461

14.1	Introduction	461	
14.2	Kernel functions	461	
14.2.1	RBF kernels	462	
14.2.2	Kernels for comparing documents	462	
14.2.3	Mercer (positive definite) kernels	463	
14.2.4	Linear kernels	464	
14.2.5	Matern kernels	464	
14.2.6	String kernels	465	
14.2.7	Pyramid match kernels	466	
14.2.8	Kernels derived from probabilistic generative models		467
14.3	Using kernels inside GLMs	468	
14.3.1	Kernel machines	468	
14.3.2	LIVMs, RVMs, and other sparse kernel machines		469
14.4	The kernel trick	470	
14.4.1	Kernelized nearest neighbor classification		471
14.4.2	Kernelized K-medoids clustering		471
14.4.3	Kernelized ridge regression	474	
14.4.4	Kernel PCA	475	
14.5	Support vector machines (SVMs)	478	
14.5.1	SVMs for regression	479	
14.5.2	SVMs for classification	480	
14.5.3	Choosing C	486	
14.5.4	Summary of key points	486	
14.5.5	A probabilistic interpretation of SVMs		487
14.6	Comparison of discriminative kernel methods		487
14.7	Kernels for building generative models	489	
14.7.1	Smoothing kernels	489	
14.7.2	Kernel density estimation (KDE)		490

14.7.3	From KDE to KNN	492
14.7.4	Kernel regression	492
14.7.5	Locally weighted regression	494
15	<i>Gaussian processes</i>	497
15.1	Introduction	497
15.2	GPs for regression	498
15.2.1	Predictions using noise-free observations	499
15.2.2	Predictions using noisy observations	500
15.2.3	Effect of the kernel parameters	501
15.2.4	Estimating the kernel parameters	503
15.2.5	Computational and numerical issues *	506
15.2.6	Semi-parametric GPs *	506
15.3	GPs meet GLMs	507
15.3.1	Binary classification	507
15.3.2	Multi-class classification	510
15.3.3	GPs for Poisson regression	513
15.4	Connection with other methods	514
15.4.1	Linear models compared to GPs	514
15.4.2	Linear smoothers compared to GPs	515
15.4.3	SVMs compared to GPs	516
15.4.4	L1VM and RVMs compared to GPs	516
15.4.5	Neural networks compared to GPs	517
15.4.6	Smoothing splines compared to GPs *	518
15.4.7	RKHS methods compared to GPs *	520
15.5	GP latent variable model	522
15.6	Approximation methods for large datasets	524
16	<i>Adaptive basis function models</i>	525
16.1	Introduction	525
16.2	Classification and regression trees (CART)	526
16.2.1	Basics	526
16.2.2	Growing a tree	528
16.2.3	Pruning a tree	531
16.2.4	Pros and cons of trees	532
16.2.5	Random forests	533
16.2.6	CART compared to hierarchical mixture of experts *	533
16.3	Generalized additive models	534
16.3.1	Backfitting	534
16.3.2	Computational efficiency	535
16.3.3	Multivariate adaptive regression splines (MARS)	535
16.4	Boosting	536
16.4.1	Forward stagewise additive modeling	537
16.4.2	L2boosting	540
16.4.3	AdaBoost	540

16.4.4	LogitBoost	542	
16.4.5	Boosting as functional gradient descent	542	
16.4.6	Sparse boosting	544	
16.4.7	Multivariate adaptive regression trees (MART)	544	
16.4.8	Why does boosting work so well?	545	
16.4.9	A Bayesian view	545	
16.5	Feedforward neural networks (multilayer perceptrons)	546	
16.5.1	Convolutional neural networks	547	
16.5.2	Other kinds of neural networks	550	
16.5.3	A brief history of the field	551	
16.5.4	The backpropagation algorithm	552	
16.5.5	Identifiability	554	
16.5.6	Regularization	554	
16.5.7	Bayesian inference *	558	
16.6	Ensemble learning	562	
16.6.1	Stacking	562	
16.6.2	Error-correcting output codes	563	
16.6.3	Ensemble learning is not equivalent to Bayes model averaging	563	
16.7	Experimental comparison	564	
16.7.1	Low-dimensional features	564	
16.7.2	High-dimensional features	565	
16.8	Interpreting black-box models	567	
17	Markov and hidden Markov Models	571	
17.1	Introduction	571	
17.2	Markov models	571	
17.2.1	Transition matrix	571	
17.2.2	Application: Language modeling	573	
17.2.3	Stationary distribution of a Markov chain *	578	
17.2.4	Application: Google's PageRank algorithm for web page ranking *	582	
17.3	Hidden Markov models	585	
17.3.1	Applications of HMMs	586	
17.4	Inference in HMMs	588	
17.4.1	Types of inference problems for temporal models	588	
17.4.2	The forwards algorithm	591	
17.4.3	The forwards-backwards algorithm	592	
17.4.4	The Viterbi algorithm	594	
17.4.5	Forwards filtering, backwards sampling	598	
17.5	Learning for HMMs	599	
17.5.1	Training with fully observed data	599	
17.5.2	EM for HMMs (the Baum-Welch algorithm)	600	
17.5.3	Bayesian methods for "fitting" HMMs *	602	
17.5.4	Discriminative training	602	
17.5.5	Model selection	603	
17.6	Generalizations of HMMs	603	

17.6.1	Variable duration (semi-Markov) HMMs	604
17.6.2	Hierarchical HMMs	606
17.6.3	Input-output HMMs	607
17.6.4	Auto-regressive and buried HMMs	608
17.6.5	Factorial HMM	609
17.6.6	Coupled HMM and the influence model	610
17.6.7	Dynamic Bayesian networks (DBNs)	610
18	<i>State space models</i>	613
18.1	Introduction	613
18.2	Applications of SSMs	614
18.2.1	SSMs for object tracking	614
18.2.2	Robotic SLAM	615
18.2.3	Online parameter learning using recursive least squares	618
18.2.4	SSM for time series forecasting *	619
18.3	Inference in LG-SSM	622
18.3.1	The Kalman filtering algorithm	622
18.3.2	The Kalman smoothing algorithm	625
18.4	Learning for LG-SSM	628
18.4.1	Identifiability and numerical stability	628
18.4.2	Training with fully observed data	628
18.4.3	EM for LG-SSM	629
18.4.4	Subspace methods	629
18.4.5	Bayesian methods for “fitting” LG-SSMs	629
18.5	Approximate online inference for non-linear, non-Gaussian SSMs	629
18.5.1	Extended Kalman filter (EKF)	630
18.5.2	Unscented Kalman filter (UKF)	632
18.5.3	Assumed density filtering (ADF)	634
18.6	Hybrid discrete/ continuous SSMs	637
18.6.1	Inference	638
18.6.2	Application: Data association and multi target tracking	640
18.6.3	Application: fault diagnosis	641
18.6.4	Application: econometric forecasting	641
19	<i>Undirected graphical models (Markov random fields)</i>	643
19.1	Introduction	643
19.2	Conditional independence properties of UGMs	643
19.2.1	Key properties	643
19.2.2	An undirected alternative to d-separation	645
19.2.3	Comparing directed and undirected graphical models	646
19.3	Parameterization of MRFs	647
19.3.1	The Hammersley-Clifford theorem	647
19.3.2	Representing potential functions	649
19.4	Examples of MRFs	650
19.4.1	Ising model	650

19.4.2	Hopfield networks	651	
19.4.3	Potts model	653	
19.4.4	Gaussian MRFs	654	
19.4.5	Markov logic networks *	656	
19.5	Learning	658	
19.5.1	Training maxent models using gradient methods	658	
19.5.2	Training partially observed maxent models	659	
19.5.3	Approximate methods for computing the MLEs of MRFs	660	
19.5.4	Pseudo likelihood	660	
19.5.5	Stochastic Maximum Likelihood	662	
19.5.6	Feature induction for maxent models *	662	
19.5.7	Iterative proportional fitting (IPF) *	664	
19.6	Conditional random fields (CRFs)	666	
19.6.1	Chain-structured CRFs, MEMMs and the label-bias problem	666	
19.7	Applications of CRFs	668	
19.7.1	Handwriting recognition	668	
19.7.2	Noun phrase chunking	669	
19.7.3	Named entity recognition	670	
19.7.4	CRFs for protein side-chain prediction	671	
19.7.5	Stereo vision	671	
19.8	CRF training	673	
19.9	Max margin methods for structured output classifiers *	674	
20	<i>Exact inference for graphical models</i>	677	
20.1	Introduction	677	
20.2	Belief propagation for trees	677	
20.2.1	Serial protocol	677	
20.2.2	Parallel protocol	679	
20.2.3	Gaussian BP *	680	
20.2.4	Other BP variants *	682	
20.3	The variable elimination algorithm	684	
20.3.1	The generalized distributive law *	687	
20.3.2	Computational complexity of VE	687	
20.3.3	A weakness of VE	690	
20.4	The junction tree algorithm *	690	
20.4.1	Creating a junction tree	690	
20.4.2	Message passing on a junction tree	692	
20.4.3	Computational complexity of JTA	695	
20.4.4	JTA generalizations *	696	
20.5	Computational intractability of exact inference in the worst case	696	
20.5.1	Approximate inference	697	
21	<i>Variational inference</i>	701	
21.1	Introduction	701	
21.2	Variational inference	702	

	21.2.1	Forward or reverse KL? *	703
21.3		The mean field method	705
	21.3.1	Derivation of the mean field update equations	706
	21.3.2	Example: Mean field for the Ising model	707
21.4		Structured mean field *	709
	21.4.1	Example: factorial HMM	709
21.5		Variational Bayes	711
	21.5.1	Example: VB for a univariate Gaussian	712
	21.5.2	Example: VB for linear regression	716
21.6		Variational Bayes EM	718
	21.6.1	Example: VBEM for mixtures of Gaussians *	720
21.7		Variational message passing and VIBES	725
21.8		Local variational bounds	726
	21.8.1	Motivating applications	726
	21.8.2	Bohning's quadratic bound to the log-sum-exp function	727
	21.8.3	Bounds for the sigmoid function	729
	21.8.4	Other bounds and approximations to the log-sum-exp function *	732
	21.8.5	Variational inference based on upper bounds	733
22		More variational inference	737
22.1		Introduction	737
22.2		Loopy belief propagation: algorithmic issues	737
	22.2.1	A brief history	737
	22.2.2	LBP on pairwise models	738
	22.2.3	LBP on a factor graph	739
	22.2.4	Convergence	741
	22.2.5	Other speedup tricks for BP	744
	22.2.6	Accuracy of LBP	746
22.3		Loopy belief propagation: theoretical issues *	746
	22.3.1	UGMs represented in exponential family form	746
	22.3.2	The marginal polytope	747
	22.3.3	Exact inference as a variational optimization problem	748
	22.3.4	Mean field as a variational optimization problem	749
	22.3.5	LBP as a variational optimization problem	749
	22.3.6	Loopy BP vs mean field	753
22.4		Extensions of belief propagation *	753
	22.4.1	Generalized belief propagation	753
	22.4.2	Convex belief propagation	755
22.5		Expectation propagation	757
22.6		MAP state estimation	758
	22.6.1	Linear programming relaxation	758
	22.6.2	Max-product belief propagation	759
	22.6.3	Dual decomposition	760
	22.6.4	Submodularity	763
	22.6.5	Graphcuts	763

23	Monte Carlo inference	767
23.1	Introduction	767
23.2	Sampling from standard distributions	767
23.2.1	Using the cdf	767
23.2.2	Sampling from a Gaussian (Box-Muller method)	769
23.3	Rejection sampling	769
23.3.1	Basic idea	769
23.3.2	Example	770
23.3.3	Application to Bayesian statistics	771
23.3.4	Adaptive rejection sampling	771
23.3.5	Rejection sampling in high dimensions	772
23.4	Importance sampling	772
23.4.1	Basic idea	772
23.4.2	Handling unnormalized distributions	773
23.4.3	Importance sampling for a DGM: Likelihood weighting	774
23.4.4	Sampling importance resampling (SIR)	774
23.5	Particle filtering	775
23.5.1	Sequential importance sampling	776
23.5.2	The degeneracy problem	777
23.5.3	The resampling step	777
23.5.4	The proposal distribution	779
23.5.5	Application: Robot localization	780
23.5.6	Application: Visual object tracking	780
23.5.7	Application: time series forecasting	783
23.6	Rao-Blackwellised particle filtering (RBPF)	783
23.6.1	RBPF for switching LG-SSMs	783
23.6.2	Application: Tracking a maneuvering target	784
23.6.3	Application: Fast SLAM	786
24	Markov Chain Monte Carlo (MCMC) inference	789
24.1	Introduction	789
24.2	Gibbs sampling	790
24.2.1	Basic idea	790
24.2.2	Example: Gibbs sampling for the Ising model	790
24.2.3	Example: Gibbs sampling for inferring the parameters of a GMM	792
24.2.4	Collapsed Gibbs sampling *	793
24.2.5	Gibbs sampling for hierarchical GLMs	795
24.2.6	BUGS and JAGS	798
24.2.7	The Imputation Posterior (IP) algorithm	799
24.2.8	Blocking Gibbs sampling	799
24.3	Metropolis Hastings algorithm	800
24.3.1	Basic idea	800
24.3.2	Gibbs sampling is a special case of MH	802
24.3.3	Proposal distributions	802
24.3.4	Adaptive MCMC	805

24.3.5	Initialization and mode hopping	805	
24.3.6	Why MH works *	806	
24.3.7	Reversible jump (trans-dimensional) MCMC *	807	
24.4	Speed and accuracy of MCMC	807	
24.4.1	The burn-in phase	807	
24.4.2	Mixing rates of Markov chains *	809	
24.4.3	Practical convergence diagnostics	810	
24.4.4	Accuracy of MCMC	812	
24.4.5	How many chains?	814	
24.5	Auxiliary variable MCMC *	815	
24.5.1	Auxiliary variable sampling for logistic regression	815	
24.5.2	Slice sampling	816	
24.5.3	Swendsen Wang	818	
24.5.4	Hybrid/ Hamiltonian MCMC *	820	
24.6	Simulated annealing	820	
24.7	Approximating the marginal likelihood	822	
24.7.1	The candidate method	823	
24.7.2	Harmonic mean estimate	823	
24.7.3	Annealed importance sampling	823	
25	Clustering	827	
25.1	Introduction	827	
25.1.1	Measuring (dis)similarity	827	
25.1.2	Evaluating the output of clustering methods *	828	
25.2	Dirichlet process mixture models	831	
25.2.1	From finite to infinite mixture models	831	
25.2.2	The Dirichlet process	834	
25.2.3	Applying Dirichlet processes to mixture modeling	837	
25.2.4	Fitting a DP mixture model	838	
25.3	Affinity propagation	839	
25.4	Spectral clustering	842	
25.4.1	Graph Laplacian	843	
25.4.2	Normalized graph Laplacian	844	
25.4.3	Example	845	
25.5	Hierarchical clustering	845	
25.5.1	Agglomerative clustering	847	
25.5.2	Divisive clustering	850	
25.5.3	Choosing the number of clusters	851	
25.5.4	Bayesian hierarchical clustering	851	
25.6	Clustering datapoints and features	853	
25.6.1	Biclustering	855	
25.6.2	Multi-view clustering	855	
26	Graphical model structure learning	859	
26.1	Introduction	859	

26.2	Quick and dirty ways to learn graph structure	860
26.2.1	Relevance networks	860
26.2.2	Dependency networks	861
26.3	Learning tree structures	862
26.3.1	Directed or undirected tree?	863
26.3.2	Chow-Liu algorithm for finding the ML tree structure	864
26.3.3	Finding the MAP forest	864
26.3.4	Mixtures of trees	866
26.4	Learning DAG structures	866
26.4.1	Exact structural inference	866
26.4.2	Scaling up to larger graphs	872
26.5	Learning DAG structure with latent variables	874
26.5.1	Approximating the marginal likelihood when we have missing data	874
26.5.2	Structural EM	877
26.5.3	Discovering hidden variables	877
26.5.4	Case study: Google's Rephil	880
26.5.5	Structural equation models *	881
26.6	Learning causal DAGs	883
26.6.1	Causal interpretation of DAGs	883
26.6.2	Using causal DAGs to resolve Simpson's paradox	884
26.6.3	Learning causal DAG structures	887
26.7	Learning undirected Gaussian graphical models	890
26.7.1	MLE for a GRF	890
26.7.2	Graphical lasso	891
26.7.3	Bayesian inference for GRF structure	893
26.7.4	Handling non-Gaussian data *	895
26.8	Learning undirected discrete graphical models	895
26.8.1	Graphical lasso for MRFs/ CRFs	895
26.8.2	Thin junction trees	896
27	<i>Latent variable models for discrete data</i>	899
27.1	Introduction	899
27.2	Distributed state LVMs for discrete data	900
27.2.1	Mixture models	900
27.2.2	Exponential family PCA	901
27.2.3	LDA and mPCA	902
27.2.4	GaP model and non-negative matrix factorization	903
27.3	Latent Dirichlet allocation (LDA)	904
27.3.1	Basics	904
27.3.2	Unsupervised discovery of topics	907
27.3.3	Quantitatively evaluating LDA as a language model	907
27.3.4	Fitting using (collapsed) Gibbs sampling	909
27.3.5	Example	910
27.3.6	Fitting using batch variational inference	911
27.3.7	Fitting using online variational inference	913

27.3.8	Determining the number of topics	914	
27.4	Extensions of LDA	915	
27.4.1	Correlated topic model	915	
27.4.2	Dynamic topic model	916	
27.4.3	LDA-HMM	917	
27.4.4	Supervised LDA	919	
27.5	LVMs for graph-structured data	924	
27.5.1	Stochastic block model	925	
27.5.2	Mixed membership stochastic block model	927	
27.5.3	Relational topic model	929	
27.6	LVMs for relational data	930	
27.6.1	Infinite relational model	931	
27.6.2	Probabilistic matrix factorization for collaborative filtering	933	
27.7	Restricted Boltzmann machines (RBMs)	937	
27.7.1	Varieties of RBMs	939	
27.7.2	Learning RBMs	941	
27.7.3	Applications of RBMs	945	
28	Deep learning	949	
28.1	Introduction	949	
28.2	Deep generative models	950	
28.2.1	Deep sigmoid networks	950	
28.2.2	Deep Boltzmann machines	951	
28.2.3	Deep belief networks	952	
28.3	Training deep networks	953	
28.3.1	Greedy layer-wise learning of DBNs	953	
28.3.2	Fitting deep neural nets	955	
28.3.3	Fitting deep auto-encoders	955	
28.3.4	Stacked denoising auto-encoders	956	
28.4	Applications of deep networks	956	
28.4.1	Handwritten digit classification using DBNs	956	
28.4.2	Data visualization using deep auto-encoders	958	
28.4.3	Information retrieval using deep autoencoders (semantic hashing)	958	
28.4.4	Learning audio features using 1d convolutional DBNs	959	
28.4.5	Learning image features using 2d convolutional DBNs	960	
28.5	Discussion	961	
	Bibliography	963	
	Index to code	993	
	Index to keywords	997	

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, Carl Rasmussen, Ryan Rifkin, Ruslan Salakhutdinov, Mark Schmidt, Erik Sudderth, Josh Tenenbaum, Kai Yu, Martin Wainwright, Yair Weiss.

Kevin Murphy
Palo Alto, California
March 2012

