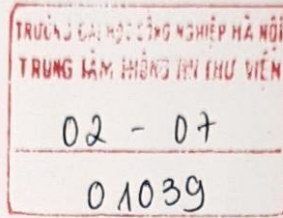




Machine Learning

A Probabilistic Perspective

Kevin P. Murphy



Machine Learning

A Probabilistic Perspective

Kevin P. Murphy

The MIT Press
Cambridge, Massachusetts
London, England

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's t distribution	39
	2.4.4	The Laplace distribution	41
	2.4.5	The gamma distribution	41
	2.4.6	The beta distribution	43
	2.4.7	Pareto distribution	43
2.5		Joint probability distributions	44
	2.5.1	Covariance and correlation	45
	2.5.2	The multivariate Gaussian	46
	2.5.3	Multivariate Student t distribution	47
	2.5.4	Dirichlet distribution	49
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	52
2.7		Monte Carlo approximation	53
	2.7.1	Example: change of variables, the MC way	53
	2.7.2	Example: estimating π by Monte Carlo integration	54
	2.7.3	Accuracy of Monte Carlo approximation	54
2.8		Information theory	56
	2.8.1	Entropy	57
	2.8.2	KL divergence	58
	2.8.3	Mutual information	59
3		Generative models for discrete data	67
	3.1	Introduction	67
	3.2	Bayesian concept learning	67
	3.2.1	Likelihood	69
	3.2.2	Prior	69
	3.2.3	Posterior	70
	3.2.4	Posterior predictive distribution	73
	3.2.5	A more complex prior	74
	3.3	The beta-binomial model	74
	3.3.1	Likelihood	75
	3.3.2	Prior	76
	3.3.3	Posterior	77

3.3.4	Posterior predictive distribution	79
3.4	The Dirichlet-multinomial model	80
3.4.1	Likelihood	81
3.4.2	Prior	81
3.4.3	Posterior	81
3.4.4	Posterior predictive	83
3.5	Naive Bayes classifiers	84
3.5.1	Model fitting	85
3.5.2	Using the model for prediction	87
3.5.3	The log-sum-exp trick	88
3.5.4	Feature selection using mutual information	89
3.5.5	Classifying documents using bag of words	90
4	Gaussian models	99
4.1	Introduction	99
4.1.1	Notation	99
4.1.2	Basics	99
4.1.3	MLE for an MVN	101
4.1.4	Maximum entropy derivation of the Gaussian *	103
4.2	Gaussian discriminant analysis	103
4.2.1	Quadratic discriminant analysis (QDA)	104
4.2.2	Linear discriminant analysis (LDA)	105
4.2.3	Two-class LDA	106
4.2.4	MLE for discriminant analysis	108
4.2.5	Strategies for preventing overfitting	108
4.2.6	Regularized LDA *	109
4.2.7	Diagonal LDA	110
4.2.8	Nearest shrunken centroids classifier *	111
4.3	Inference in jointly Gaussian distributions	112
4.3.1	Statement of the result	113
4.3.2	Examples	113
4.3.3	Information form	117
4.3.4	Proof of the result *	118
4.4	Linear Gaussian systems	121
4.4.1	Statement of the result	122
4.4.2	Examples	122
4.4.3	Proof of the result *	127
4.5	Digression: The Wishart distribution *	128
4.5.1	Inverse Wishart distribution	129
4.5.2	Visualizing the Wishart distribution *	129
4.6	Inferring the parameters of an MVN	129
4.6.1	Posterior distribution of μ	130
4.6.2	Posterior distribution of Σ *	131
4.6.3	Posterior distribution of μ and Σ *	134
4.6.4	Sensor fusion with unknown precisions *	140

5	Bayesian statistics	151
5.1	Introduction	151
5.2	Summarizing posterior distributions	151
5.2.1	MAP estimation	151
5.2.2	Credible intervals	154
5.2.3	Inference for a difference in proportions	156
5.3	Bayesian model selection	157
5.3.1	Bayesian Occam's razor	158
5.3.2	Computing the marginal likelihood (evidence)	160
5.3.3	Bayes factors	165
5.3.4	Jeffreys-Lindley paradox *	166
5.4	Priors	167
5.4.1	Uninformative priors	167
5.4.2	Jeffreys priors *	168
5.4.3	Robust priors	170
5.4.4	Mixtures of conjugate priors	171
5.5	Hierarchical Bayes	173
5.5.1	Example: modeling related cancer rates	173
5.6	Empirical Bayes	174
5.6.1	Example: beta-binomial model	175
5.6.2	Example: Gaussian-Gaussian model	176
5.7	Bayesian decision theory	178
5.7.1	Bayes estimators for common loss functions	179
5.7.2	The false positive vs false negative tradeoff	182
5.7.3	Other topics *	186
6	Frequentist statistics	193
6.1	Introduction	193
6.2	Sampling distribution of an estimator	193
6.2.1	Bootstrap	194
6.2.2	Large sample theory for the MLE *	195
6.3	Frequentist decision theory	197
6.3.1	Bayes risk	197
6.3.2	Minimax risk	198
6.3.3	Admissible estimators	199
6.4	Desirable properties of estimators	202
6.4.1	Consistent estimators	202
6.4.2	Unbiased estimators	203
6.4.3	Minimum variance estimators	203
6.4.4	The bias-variance tradeoff	204
6.5	Empirical risk minimization	207
6.5.1	Regularized risk minimization	208
6.5.2	Structural risk minimization	208
6.5.3	Estimating the risk using cross validation	209
6.5.4	Upper bounding the risk using statistical learning theory *	211

6.5.5	Surrogate loss functions	213
6.6	Pathologies of frequentist statistics *	214
6.6.1	Counter-intuitive behavior of confidence intervals	214
6.6.2	p-values considered harmful	215
6.6.3	The likelihood principle	217
6.6.4	Why isn't everyone a Bayesian?	217
7	Linear regression	219
7.1	Introduction	219
7.2	Model specification	219
7.3	Maximum likelihood estimation (least squares)	219
7.3.1	Derivation of the MLE	221
7.3.2	Geometric interpretation	222
7.3.3	Convexity	223
7.4	Robust linear regression *	225
7.5	Ridge regression	227
7.5.1	Basic idea	227
7.5.2	Numerically stable computation *	229
7.5.3	Connection with PCA *	230
7.5.4	Regularization effects of big data	232
7.6	Bayesian linear regression	233
7.6.1	Computing the posterior	234
7.6.2	Computing the posterior predictive	235
7.6.3	Bayesian inference when σ^2 is unknown *	236
7.6.4	EB for linear regression (evidence procedure)	240
8	Logistic regression	247
8.1	Introduction	247
8.2	Model specification	247
8.3	Model fitting	248
8.3.1	MLE	249
8.3.2	Steepest descent	249
8.3.3	Newton's method	251
8.3.4	Iteratively reweighted least squares (IRLS)	253
8.3.5	Quasi-Newton (variable metric) methods	253
8.3.6	ℓ_2 regularization	254
8.3.7	Multi-class logistic regression	255
8.4	Bayesian logistic regression	257
8.4.1	Laplace approximation	257
8.4.2	Derivation of the Bayesian information criterion (BIC)	258
8.4.3	Gaussian approximation for logistic regression	258
8.4.4	Approximating the posterior predictive	260
8.4.5	Residual analysis (outlier detection) *	263
8.5	Online learning and stochastic optimization	264
8.5.1	Online learning and regret minimization	264

8.5.2	Stochastic optimization and risk minimization	265
8.5.3	The LMS algorithm	267
8.5.4	The perceptron algorithm	268
8.5.5	A Bayesian view	270
8.6	Generative vs discriminative classifiers	270
8.6.1	Pros and cons of each approach	271
8.6.2	Dealing with missing data	271
8.6.3	Fisher's linear discriminant analysis (FLDA) *	274
9	Generalized linear models and the exponential family	283
9.1	Introduction	283
9.2	The exponential family	283
9.2.1	Definition	284
9.2.2	Examples	284
9.2.3	Log partition function	286
9.2.4	MLE for the exponential family	288
9.2.5	Bayes for the exponential family *	289
9.2.6	Maximum entropy derivation of the exponential family *	291
9.3	Generalized linear models (GLMs)	292
9.3.1	Basics	292
9.3.2	ML and MAP estimation	294
9.3.3	Bayesian inference	295
9.4	Probit regression	295
9.4.1	ML/MAP estimation using gradient-based optimization	296
9.4.2	Latent variable interpretation	296
9.4.3	Ordinal probit regression *	297
9.4.4	Multinomial probit models *	297
9.5	Multi-task learning	298
9.5.1	Hierarchical Bayes for multi-task learning	298
9.5.2	Application to personalized email spam filtering	298
9.5.3	Application to domain adaptation	299
9.5.4	Other kinds of prior	299
9.6	Generalized linear mixed models *	300
9.6.1	Example: semi-parametric GLMMs for medical data	300
9.6.2	Computational issues	302
9.7	Learning to rank *	302
9.7.1	The pointwise approach	303
9.7.2	The pairwise approach	303
9.7.3	The listwise approach	304
9.7.4	Loss functions for ranking	305
10	Directed graphical models (Bayes nets)	309
10.1	Introduction	309
10.1.1	Chain rule	309
10.1.2	Conditional independence	310

10.1.3	Graphical models	310
10.1.4	Graph terminology	311
10.1.5	Directed graphical models	312
10.2	Examples	313
10.2.1	Naive Bayes classifiers	313
10.2.2	Markov and hidden Markov models	314
10.2.3	Medical diagnosis	315
10.2.4	Genetic linkage analysis *	317
10.2.5	Directed Gaussian graphical models *	320
10.3	Inference	321
10.4	Learning	322
10.4.1	Plate notation	322
10.4.2	Learning from complete data	324
10.4.3	Learning with missing and/or latent variables	325
10.5	Conditional independence properties of DGMs	326
10.5.1	d-separation and the Bayes Ball algorithm (global Markov properties)	326
10.5.2	Other Markov properties of DGMs	329
10.5.3	Markov blanket and full conditionals	329
10.6	Influence (decision) diagrams *	330
II	Mixture models and the EM algorithm	339
11.1	Latent variable models	339
11.2	Mixture models	339
11.2.1	Mixtures of Gaussians	341
11.2.2	Mixture of multinoullis	342
11.2.3	Using mixture models for clustering	342
11.2.4	Mixtures of experts	344
11.3	Parameter estimation for mixture models	347
11.3.1	Unidentifiability	348
11.3.2	Computing a MAP estimate is non-convex	349
11.4	The EM algorithm	350
11.4.1	Basic idea	351
11.4.2	EM for GMMs	352
11.4.3	EM for mixture of experts	359
11.4.4	EM for DGMs with hidden variables	360
11.4.5	EM for the Student distribution *	361
11.4.6	EM for probit regression *	364
11.4.7	Theoretical basis for EM *	365
11.4.8	Online EM	367
11.4.9	Other EM variants *	369
11.5	Model selection for latent variable models	372
11.5.1	Model selection for probabilistic models	372
11.5.2	Model selection for non-probabilistic methods	372
11.6	Fitting models with missing data	374

11.6.1	EM for the MLE of an MVN with missing data	375
12	Latent linear models	383
12.1	Factor analysis	383
12.1.1	FA is a low rank parameterization of an MVN	383
12.1.2	Inference of the latent factors	384
12.1.3	Unidentifiability	386
12.1.4	Mixtures of factor analysers	387
12.1.5	EM for factor analysis models	388
12.1.6	Fitting FA models with missing data	389
12.2	Principal components analysis (PCA)	389
12.2.1	Classical PCA: statement of the theorem	390
12.2.2	Proof *	392
12.2.3	Singular value decomposition (SVD)	394
12.2.4	Probabilistic PCA	397
12.2.5	EM algorithm for PCA	398
12.3	Choosing the number of latent dimensions	400
12.3.1	Model selection for FA/PPCA	400
12.3.2	Model selection for PCA	401
12.4	PCA for categorical data	404
12.5	PCA for paired and multi-view data	406
12.5.1	Supervised PCA (latent factor regression)	406
12.5.2	Partial least squares	408
12.5.3	Canonical correlation analysis	409
12.6	Independent Component Analysis (ICA)	409
12.6.1	Maximum likelihood estimation	412
12.6.2	The FastICA algorithm	413
12.6.3	Using EM	416
12.6.4	Other estimation principles *	417
13	Sparse linear models	423
13.1	Introduction	423
13.2	Bayesian variable selection	424
13.2.1	The spike and slab model	426
13.2.2	From the Bernoulli-Gaussian model to ℓ_0 regularization	428
13.2.3	Algorithms	429
13.3	ℓ_1 regularization: basics	431
13.3.1	Why does ℓ_1 regularization yield sparse solutions?	432
13.3.2	Optimality conditions for lasso	434
13.3.3	Comparison of least squares, lasso, ridge and subset selection	437
13.3.4	Regularization path	438
13.3.5	Model selection	441
13.3.6	Bayesian inference for linear models with Laplace priors	442
13.4	ℓ_1 regularization: algorithms	443
13.4.1	Coordinate descent	443

13.4.2	LARS and other homotopy methods	443	
13.4.3	Proximal and gradient projection methods	444	
13.4.4	EM for lasso	448	
13.5	ℓ_1 regularization: extensions	451	
13.5.1	Group lasso	451	
13.5.2	Fused lasso	456	
13.5.3	Elastic net (ridge and lasso combined)	457	
13.6	Non-convex regularizers	459	
13.6.1	Bridge regression	460	
13.6.2	Hierarchical adaptive lasso	460	
13.6.3	Other hierarchical priors	464	
13.7	Automatic relevance determination (ARD)/sparse Bayesian learning (SBL)	465	
13.7.1	ARD for linear regression	465	
13.7.2	Whence sparsity?	467	
13.7.3	Connection to MAP estimation	467	
13.7.4	Algorithms for ARD *	468	
13.7.5	ARD for logistic regression	470	
13.8	Sparse coding *	470	
13.8.1	Learning a sparse coding dictionary	471	
13.8.2	Results of dictionary learning from image patches	472	
13.8.3	Compressed sensing	474	
13.8.4	Image inpainting and denoising	474	
14	Kernels	481	
14.1	Introduction	481	
14.2	Kernel functions	481	
14.2.1	RBF kernels	482	
14.2.2	Kernels for comparing documents	482	
14.2.3	Mercer (positive definite) kernels	483	
14.2.4	Linear kernels	484	
14.2.5	Matern kernels	484	
14.2.6	String kernels	485	
14.2.7	Pyramid match kernels	486	
14.2.8	Kernels derived from probabilistic generative models	487	
14.3	Using kernels inside GLMs	488	
14.3.1	Kernel machines	488	
14.3.2	LIVMs, RVMs, and other sparse vector machines	489	
14.4	The kernel trick	490	
14.4.1	Kernelized nearest neighbor classification	491	
14.4.2	Kernelized K-medoids clustering	492	
14.4.3	Kernelized ridge regression	494	
14.4.4	Kernel PCA	495	
14.5	Support vector machines (SVMs)	498	
14.5.1	SVMs for regression	499	
14.5.2	SVMs for classification	500	

14.5.3	Choosing C	506
14.5.4	Summary of key points	506
14.5.5	A probabilistic interpretation of SVMs	507
14.6	Comparison of discriminative kernel methods	507
14.7	Kernels for building generative models	509
14.7.1	Smoothing kernels	509
14.7.2	Kernel density estimation (KDE)	510
14.7.3	From KDE to KNN	511
14.7.4	Kernel regression	512
14.7.5	Locally weighted regression	514
15	Gaussian processes	517
15.1	Introduction	517
15.2	GPs for regression	518
15.2.1	Predictions using noise-free observations	519
15.2.2	Predictions using noisy observations	520
15.2.3	Effect of the kernel parameters	521
15.2.4	Estimating the kernel parameters	523
15.2.5	Computational and numerical issues *	526
15.2.6	Semi-parametric GPs *	526
15.3	GPs meet GLMs	527
15.3.1	Binary classification	527
15.3.2	Multi-class classification	530
15.3.3	GPs for Poisson regression	533
15.4	Connection with other methods	534
15.4.1	Linear models compared to GPs	534
15.4.2	Linear smoothers compared to GPs	535
15.4.3	SVMs compared to GPs	536
15.4.4	LIVM and RVMs compared to GPs	536
15.4.5	Neural networks compared to GPs	537
15.4.6	Smoothing splines compared to GPs *	538
15.4.7	RKHS methods compared to GPs *	540
15.5	GP latent variable model	542
15.6	Approximation methods for large datasets	544
16	Adaptive basis function models	545
16.1	Introduction	545
16.2	Classification and regression trees (CART)	546
16.2.1	Basics	546
16.2.2	Growing a tree	547
16.2.3	Pruning a tree	551
16.2.4	Pros and cons of trees	552
16.2.5	Random forests	552
16.2.6	CART compared to hierarchical mixture of experts *	553
16.3	Generalized additive models	554

16.3.1	Backfitting	554	
16.3.2	Computational efficiency	555	
16.3.3	Multivariate adaptive regression splines (MARS)	555	
16.4	Boosting	556	
16.4.1	Forward stagewise additive modeling	557	
16.4.2	L2boosting	559	
16.4.3	AdaBoost	560	
16.4.4	LogitBoost	561	
16.4.5	Boosting as functional gradient descent	562	
16.4.6	Sparse boosting	563	
16.4.7	Multivariate adaptive regression trees (MART)	564	
16.4.8	Why does boosting work so well?	564	
16.4.9	A Bayesian view	565	
16.5	Feedforward neural networks (multilayer perceptrons)	565	
16.5.1	Convolutional neural networks	566	
16.5.2	Other kinds of neural networks	570	
16.5.3	A brief history of the field	571	
16.5.4	The backpropagation algorithm	572	
16.5.5	Identifiability	574	
16.5.6	Regularization	574	
16.5.7	Bayesian inference *	578	
16.6	Ensemble learning	582	
16.6.1	Stacking	582	
16.6.2	Error-correcting output codes	583	
16.6.3	Ensemble learning is not equivalent to Bayes model averaging	583	
16.7	Experimental comparison	584	
16.7.1	Low-dimensional features	584	
16.7.2	High-dimensional features	585	
16.8	Interpreting black-box models	587	
17	Markov and hidden Markov models	591	
17.1	Introduction	591	
17.2	Markov models	591	
17.2.1	Transition matrix	591	
17.2.2	Application: Language modeling	593	
17.2.3	Stationary distribution of a Markov chain *	598	
17.2.4	Application: Google's PageRank algorithm for web page ranking *	602	
17.3	Hidden Markov models	606	
17.3.1	Applications of HMMs	606	
17.4	Inference in HMMs	608	
17.4.1	Types of inference problems for temporal models	608	
17.4.2	The forwards algorithm	611	
17.4.3	The forwards-backwards algorithm	612	
17.4.4	The Viterbi algorithm	614	
17.4.5	Forwards filtering, backwards sampling	619	

17.5	Learning for HMMs	619	
17.5.1	Training with fully observed data	620	
17.5.2	EM for HMMs (the Baum-Welch algorithm)	620	
17.5.3	Bayesian methods for "fitting" HMMs *	622	
17.5.4	Discriminative training	623	
17.5.5	Model selection	623	
17.6	Generalizations of HMMs	624	
17.6.1	Variable duration (semi-Markov) HMMs	624	
17.6.2	Hierarchical HMMs	626	
17.6.3	Input-output HMMs	628	
17.6.4	Auto-regressive and buried HMMs	628	
17.6.5	Factorial HMM	629	
17.6.6	Coupled HMM and the influence model	630	
17.6.7	Dynamic Bayesian networks (DBNs)	631	
18	State space models	633	
18.1	Introduction	633	
18.2	Applications of SSMs	634	
18.2.1	SSMs for object tracking	634	
18.2.2	Robotic SLAM	635	
18.2.3	Online parameter learning using recursive least squares	638	
18.2.4	SSM for time series forecasting *	639	
18.3	Inference in LG-SSM	642	
18.3.1	The Kalman filtering algorithm	642	
18.3.2	The Kalman smoothing algorithm	645	
18.4	Learning for LG-SSM	648	
18.4.1	Identifiability and numerical stability	648	
18.4.2	Training with fully observed data	649	
18.4.3	EM for LG-SSM	649	
18.4.4	Subspace methods	649	
18.4.5	Bayesian methods for "fitting" LG-SSMs	649	
18.5	Approximate online inference for non-linear, non-Gaussian SSMs	649	
18.5.1	Extended Kalman filter (EKF)	650	
18.5.2	Unscented Kalman filter (UKF)	652	
18.5.3	Assumed density filtering (ADF)	654	
18.6	Hybrid discrete/continuous SSMs	657	
18.6.1	Inference	658	
18.6.2	Application: data association and multi-target tracking	660	
18.6.3	Application: fault diagnosis	661	
18.6.4	Application: econometric forecasting	662	
19	Undirected graphical models (Markov random fields)	663	
19.1	Introduction	663	
19.2	Conditional independence properties of UGMs	663	
19.2.1	Key properties	663	

19.2.2	An undirected alternative to d-separation	665
19.2.3	Comparing directed and undirected graphical models	666
19.3	Parameterization of MRFs	667
19.3.1	The Hammersley-Clifford theorem	667
19.3.2	Representing potential functions	669
19.4	Examples of MRFs	670
19.4.1	Ising model	670
19.4.2	Hopfield networks	671
19.4.3	Potts model	673
19.4.4	Gaussian MRFs	674
19.4.5	Markov logic networks *	676
19.5	Learning	678
19.5.1	Training maxent models using gradient methods	678
19.5.2	Training partially observed maxent models	679
19.5.3	Approximate methods for computing the MLEs of MRFs	680
19.5.4	Pseudo likelihood	680
19.5.5	Stochastic maximum likelihood	682
19.5.6	Feature induction for maxent models *	682
19.5.7	Iterative proportional fitting (IPF) *	684
19.6	Conditional random fields (CRFs)	686
19.6.1	Chain-structured CRFs, MEMMs and the label-bias problem	687
19.6.2	Applications of CRFs	688
19.6.3	CRF training	694
19.7	Structural SVMs	696
19.7.1	SSVMs: a probabilistic view	696
19.7.2	SSVMs: a non-probabilistic view	698
19.7.3	Cutting plane methods for fitting SSVMs	700
19.7.4	Online algorithms for fitting SSVMs	703
19.7.5	Latent structural SVMs	704
20 Exact inference for graphical models		709
20.1	Introduction	709
20.2	Belief propagation for trees	709
20.2.1	Serial protocol	709
20.2.2	Parallel protocol	711
20.2.3	Gaussian belief propagation *	712
20.2.4	Other BP variants *	714
20.3	The variable elimination algorithm	716
20.3.1	The generalized distributive law *	719
20.3.2	Computational complexity of VE	719
20.3.3	A weakness of VE	722
20.4	The junction tree algorithm *	722
20.4.1	Creating a junction tree	722
20.4.2	Message passing on a junction tree	724
20.4.3	Computational complexity of JTA	727

20.4.4	JTA generalizations *	728	
20.5	Computational intractability of exact inference in the worst case		728
20.5.1	Approximate inference	729	
21	Variational inference	733	
21.1	Introduction	733	
21.2	Variational inference	733	
21.2.1	Alternative interpretations of the variational objective		735
21.2.2	Forward or reverse KL? *	735	
21.3	The mean field method	737	
21.3.1	Derivation of the mean field update equations		738
21.3.2	Example: mean field for the Ising model	739	
21.4	Structured mean field *	741	
21.4.1	Example: factorial HMM	742	
21.5	Variational Bayes	744	
21.5.1	Example: VB for a univariate Gaussian		744
21.5.2	Example: VB for linear regression	748	
21.6	Variational Bayes EM	751	
21.6.1	Example: VBEM for mixtures of Gaussians *		752
21.7	Variational message passing and VIBES	758	
21.8	Local variational bounds *	758	
21.8.1	Motivating applications	758	
21.8.2	Bohning's quadratic bound to the log-sum-exp function		760
21.8.3	Bounds for the sigmoid function	762	
21.8.4	Other bounds and approximations to the log-sum-exp function *		764
21.8.5	Variational inference based on upper bounds	765	
22	More variational inference	769	
22.1	Introduction	769	
22.2	Loopy belief propagation: algorithmic issues		769
22.2.1	A brief history	769	
22.2.2	LBP on pairwise models	770	
22.2.3	LBP on a factor graph	771	
22.2.4	Convergence	773	
22.2.5	Accuracy of LBP	776	
22.2.6	Other speedup tricks for LBP *		777
22.3	Loopy belief propagation: theoretical issues *		778
22.3.1	UGMs represented in exponential family form		778
22.3.2	The marginal polytope	779	
22.3.3	Exact inference as a variational optimization problem		780
22.3.4	Mean field as a variational optimization problem		781
22.3.5	LBP as a variational optimization problem	781	
22.3.6	Loopy BP vs mean field	785	
22.4	Extensions of belief propagation *	785	
22.4.1	Generalized belief propagation	785	

22.4.2	Convex belief propagation	787
22.5	Expectation propagation	789
22.5.1	EP as a variational inference problem	790
22.5.2	Optimizing the EP objective using moment matching	791
22.5.3	EP for the clutter problem	793
22.5.4	LBP is a special case of EP	794
22.5.5	Ranking players using TrueSkill	795
22.5.6	Other applications of EP	801
22.6	MAP state estimation	801
22.6.1	Linear programming relaxation	801
22.6.2	Max-product belief propagation	802
22.6.3	Graphcuts	803
22.6.4	Experimental comparison of graphcuts and BP	806
22.6.5	Dual decomposition	808
23	Monte Carlo inference	817
23.1	Introduction	817
23.2	Sampling from standard distributions	817
23.2.1	Using the cdf	817
23.2.2	Sampling from a Gaussian (Box-Muller method)	819
23.3	Rejection sampling	819
23.3.1	Basic idea	819
23.3.2	Example	820
23.3.3	Application to Bayesian statistics	821
23.3.4	Adaptive rejection sampling	821
23.3.5	Rejection sampling in high dimensions	822
23.4	Importance sampling	822
23.4.1	Basic idea	822
23.4.2	Handling unnormalized distributions	823
23.4.3	Importance sampling for a DGM: likelihood weighting	824
23.4.4	Sampling importance resampling (SIR)	825
23.5	Particle filtering	825
23.5.1	Sequential importance sampling	826
23.5.2	The degeneracy problem	827
23.5.3	The resampling step	827
23.5.4	The proposal distribution	829
23.5.5	Application: robot localization	830
23.5.6	Application: visual object tracking	830
23.5.7	Application: time series forecasting	833
23.6	Rao-Blackwellised particle filtering (RBPF)	833
23.6.1	RBPF for switching LG-SSMs	833
23.6.2	Application: tracking a maneuvering target	834
23.6.3	Application: Fast SLAM	836
24	Markov chain Monte Carlo (MCMC) inference	839

24.1	Introduction	839	
24.2	Gibbs sampling	840	
24.2.1	Basic idea	840	
24.2.2	Example: Gibbs sampling for the Ising model	840	
24.2.3	Example: Gibbs sampling for inferring the parameters of a GMM	842	
24.2.4	Collapsed Gibbs sampling *	843	
24.2.5	Gibbs sampling for hierarchical GLMs	846	
24.2.6	BUGS and JAGS	848	
24.2.7	The Imputation Posterior (IP) algorithm	849	
24.2.8	Blocking Gibbs sampling	849	
24.3	Metropolis Hastings algorithm	850	
24.3.1	Basic idea	850	
24.3.2	Gibbs sampling is a special case of MH	851	
24.3.3	Proposal distributions	852	
24.3.4	Adaptive MCMC	855	
24.3.5	Initialization and mode hopping	856	
24.3.6	Why MH works *	856	
24.3.7	Reversible jump (trans-dimensional) MCMC *	857	
24.4	Speed and accuracy of MCMC	858	
24.4.1	The burn-in phase	858	
24.4.2	Mixing rates of Markov chains *	859	
24.4.3	Practical convergence diagnostics	860	
24.4.4	Accuracy of MCMC	862	
24.4.5	How many chains?	864	
24.5	Auxiliary variable MCMC *	865	
24.5.1	Auxiliary variable sampling for logistic regression	865	
24.5.2	Slice sampling	866	
24.5.3	Swendsen Wang	868	
24.5.4	Hybrid/Hamiltonian MCMC *	870	
24.6	Annealing methods	870	
24.6.1	Simulated annealing	871	
24.6.2	Annealed importance sampling	873	
24.6.3	Parallel tempering	873	
24.7	Approximating the marginal likelihood	874	
24.7.1	The candidate method	874	
24.7.2	Harmonic mean estimate	874	
24.7.3	Annealed importance sampling	875	
25	Clustering	877	
25.1	Introduction	877	
25.1.1	Measuring (dis)similarity	877	
25.1.2	Evaluating the output of clustering methods *	878	
25.2	Dirichlet process mixture models	881	
25.2.1	From finite to infinite mixture models	881	
25.2.2	The Dirichlet process	884	

25.2.3	Applying Dirichlet processes to mixture modeling	887
25.2.4	Fitting a DP mixture model	888
25.3	Affinity propagation	889
25.4	Spectral clustering	892
25.4.1	Graph Laplacian	893
25.4.2	Normalized graph Laplacian	894
25.4.3	Example	895
25.5	Hierarchical clustering	895
25.5.1	Agglomerative clustering	897
25.5.2	Divisive clustering	900
25.5.3	Choosing the number of clusters	901
25.5.4	Bayesian hierarchical clustering	901
25.6	Clustering datapoints and features	903
25.6.1	Biclustering	905
25.6.2	Multi-view clustering	905
26	Graphical model structure learning	909
26.1	Introduction	909
26.2	Structure learning for knowledge discovery	910
26.2.1	Relevance networks	910
26.2.2	Dependency networks	911
26.3	Learning tree structures	912
26.3.1	Directed or undirected tree?	913
26.3.2	Chow-Liu algorithm for finding the ML tree structure	914
26.3.3	Finding the MAP forest	914
26.3.4	Mixtures of trees	916
26.4	Learning DAG structures	916
26.4.1	Markov equivalence	916
26.4.2	Exact structural inference	918
26.4.3	Scaling up to larger graphs	922
26.5	Learning DAG structure with latent variables	924
26.5.1	Approximating the marginal likelihood when we have missing data	924
26.5.2	Structural EM	927
26.5.3	Discovering hidden variables	928
26.5.4	Case study: Google's Rephil	930
26.5.5	Structural equation models *	931
26.6	Learning causal DAGs	933
26.6.1	Causal interpretation of DAGs	933
26.6.2	Using causal DAGs to resolve Simpson's paradox	935
26.6.3	Learning causal DAG structures	938
26.7	Learning undirected Gaussian graphical models	940
26.7.1	MLE for a GGM	940
26.7.2	Graphical lasso	941
26.7.3	Bayesian inference for GGM structure *	943
26.7.4	Handling non-Gaussian data using copulas *	944

26.8	Learning undirected discrete graphical models	944
26.8.1	Graphical lasso for MRFs/CRFs	944
26.8.2	Thin junction trees	945
27	<i>Latent variable models for discrete data</i>	949
27.1	Introduction	949
27.2	Distributed state LVMs for discrete data	950
27.2.1	Mixture models	950
27.2.2	Exponential family PCA	951
27.2.3	LDA and mPCA	952
27.2.4	GaP model and non-negative matrix factorization	953
27.3	Latent Dirichlet allocation (LDA)	954
27.3.1	Basics	954
27.3.2	Unsupervised discovery of topics	957
27.3.3	Quantitatively evaluating LDA as a language model	957
27.3.4	Fitting using (collapsed) Gibbs sampling	959
27.3.5	Example	960
27.3.6	Fitting using batch variational inference	961
27.3.7	Fitting using online variational inference	963
27.3.8	Determining the number of topics	964
27.4	Extensions of LDA	965
27.4.1	Correlated topic model	965
27.4.2	Dynamic topic model	966
27.4.3	LDA-HMM	967
27.4.4	Supervised LDA	971
27.5	LVMs for graph-structured data	974
27.5.1	Stochastic block model	975
27.5.2	Mixed membership stochastic block model	977
27.5.3	Relational topic model	978
27.6	LVMs for relational data	979
27.6.1	Infinite relational model	980
27.6.2	Probabilistic matrix factorization for collaborative filtering	983
27.7	Restricted Boltzmann machines (RBMs)	987
27.7.1	Varieties of RBMs	989
27.7.2	Learning RBMs	991
27.7.3	Applications of RBMs	995
28	<i>Deep learning</i>	999
28.1	Introduction	999
28.2	Deep generative models	999
28.2.1	Deep directed networks	1000
28.2.2	Deep Boltzmann machines	1000
28.2.3	Deep belief networks	1001
28.2.4	Greedy layer-wise learning of DBNs	1002
28.3	Deep neural networks	1003

28.3.1	Deep multi-layer perceptrons	1003	
28.3.2	Deep auto-encoders	1004	
28.3.3	Stacked denoising auto-encoders	1005	
28.4	Applications of deep networks	1005	
28.4.1	Handwritten digit classification using DBNs	1005	
28.4.2	Data visualization and feature discovery using deep auto-encoders	1006	
28.4.3	Information retrieval using deep auto-encoders (semantic hashing)	1007	
28.4.4	Learning audio features using 1d convolutional DBNs	1008	
28.4.5	Learning image features using 2d convolutional DBNs	1009	
28.5	Discussion	1010	
Notation		1013	
Bibliography		1019	
Indexes		1051	
	Index to code	1051	
	Index to keywords	1054	

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 anyone 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. 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. 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 website.

Details on *how to use* PMTK can be found on its website. 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, Jutta Degener, 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

First printing: August 2012
Second printing: November 2012 (same as first)
Third printing: February 2013 (fixed some typos)
Fourth printing: August 2013 (fixed many typos)