

Machine Learning: A Probabilistic Perspective

Machine Learning

A Probabilistic Perspective

Kevin P. Murphy

The MIT Press
Cambridge, Massachusetts
London, England

Brief Contents

1	<i>Introduction</i>	1
2	<i>Probability</i>	25
3	<i>Generative models for discrete data</i>	61
4	<i>Gaussian models</i>	89
5	<i>Bayesian statistics</i>	133
6	<i>Frequentist statistics</i>	173
7	<i>Linear regression</i>	197
8	<i>Logistic regression</i>	225
9	<i>Generalized linear models and the exponential family</i>	259
10	<i>Graphical models</i>	279
11	<i>Mixture models and the EM algorithm</i>	317
12	<i>Latent linear models</i>	359
13	<i>Sparse linear models</i>	399
14	<i>Kernels</i>	457
15	<i>Gaussian processes</i>	493
16	<i>Adaptive basis function models</i>	521
17	<i>Markov and hidden Markov Models</i>	565
18	<i>State space models</i>	609
19	<i>Markov random fields revisited</i>	639
20	<i>Exact inference algorithms for graphical models</i>	669
21	<i>Variational inference</i>	693
22	<i>Belief propagation revisited</i>	729
23	<i>Monte Carlo inference</i>	759
24	<i>Markov Chain Monte Carlo (MCMC) inference</i>	783
25	<i>Clustering</i>	819
26	<i>Graphical model structure learning</i>	851
27	<i>Latent variable models for discrete data</i>	891
28	<i>Deep learning</i>	941

Contents

1	<i>Introduction</i>	1
1.1	What is machine learning?	1
1.2	Supervised learning	2
1.2.1	Classification	2
1.2.2	Regression	8
1.3	Unsupervised learning	8
1.3.1	Discovering clusters	9
1.3.2	Discovering latent factors	10
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	15
1.4.3	The curse of dimensionality	17
1.4.4	Parametric models for classification and regression	18
1.4.5	Linear regression	18
1.4.6	Logistic regression	20
1.4.7	Overfitting	21
1.4.8	Model selection	21
1.4.9	No free lunch theorem	23
2	<i>Probability</i>	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
2.5	Joint probability distributions	41
2.5.1	Covariance and correlation	41
2.5.2	The multivariate Gaussian	43
2.5.3	Multivariate Student t distribution	44
2.5.4	Dirichlet distribution	44
2.6	Transformations of random variables	46
2.6.1	Linear transformations	46
2.6.2	General transformations	46
2.6.3	Central limit theorem	47
2.7	Monte Carlo approximation	48
2.7.1	Example: change of variables, the MC way	49
2.7.2	Example: estimating π by Monte Carlo integration	49
2.7.3	Accuracy of Monte Carlo approximation	50
2.8	Information theory	51
2.8.1	Entropy	52
2.8.2	KL divergence	52
2.8.3	Mutual information	54
3	<i>Generative models for discrete data</i>	61
3.1	Introduction	61
3.2	Bayesian concept learning	61
3.2.1	Likelihood	63
3.2.2	Prior	63
3.2.3	Posterior	64
3.2.4	Posterior predictive distribution	65
3.2.5	A more complex prior	67
3.2.6	Philosophical significance of the result	68
3.3	The Beta-Binomial model	68
3.3.1	Likelihood	69
3.3.2	Prior	69
3.3.3	Posterior	70
3.3.4	Posterior predictive distribution	72
3.4	The Dirichlet-multinomial model	74

3.4.1	Likelihood	74	
3.4.2	Prior	74	
3.4.3	Posterior	75	
3.4.4	Posterior predictive	76	
3.5	Naive Bayes classifiers	77	
3.5.1	Model fitting	78	
3.5.2	Using the model for prediction	80	
3.5.3	The log-sum-exp trick	81	
3.5.4	Feature selection using mutual information	81	
3.5.5	Classifying documents using bag of words	82	
4	<i>Gaussian models</i>	89	
4.1	Introduction	89	
4.1.1	Notation	89	
4.1.2	Basics	89	
4.1.3	MLE for an MVN	91	
4.1.4	Maximum entropy derivation of the Gaussian *	93	
4.2	Gaussian Discriminant analysis	93	
4.2.1	Quadratic discriminant analysis (QDA)	94	
4.2.2	Linear discriminant analysis (LDA)	94	
4.2.3	Two-class LDA	96	
4.2.4	MLE for discriminant analysis	98	
4.2.5	Strategies for preventing overfitting	98	
4.2.6	Regularized LDA *	99	
4.2.7	Diagonal LDA	100	
4.2.8	Nearest shrunken centroids classifier *	100	
4.3	Inference in jointly Gaussian distributions	101	
4.3.1	Statement of the result	102	
4.3.2	Examples	103	
4.3.3	Proof of the result *	107	
4.4	Linear Gaussian systems	110	
4.4.1	Statement of the result	110	
4.4.2	Examples	111	
4.4.3	Proof of the result *	115	
4.5	Inferring the parameters of an MVN	116	
4.5.1	Posterior distribution of μ	116	
4.5.2	Posterior distribution of Σ *	117	
4.5.3	Posterior distribution of μ and Σ *	120	
4.5.4	Sensor fusion with unknown precisions *	126	
5	<i>Bayesian statistics</i>	133	
5.1	Introduction	133	
5.2	Why not just use MAP estimation?	133	
5.2.1	No measure of uncertainty	134	
5.2.2	Plugging in the MAP estimate can result in overfitting	134	

5.2.3	The mode is an untypical point	134	
5.2.4	MAP estimation is not invariant to reparameterization *	135	
5.3	Computing our confidence in various quantities of interest	136	
5.3.1	Credible intervals	136	
5.3.2	Inference for a difference in proportions	138	
5.4	Bayesian model selection	139	
5.4.1	Bayesian Occam's razor	140	
5.4.2	Computing the marginal likelihood (evidence)	142	
5.4.3	Bayes factors	146	
5.4.4	Jeffreys-Lindley paradox *	148	
5.5	Priors	149	
5.5.1	Uninformative priors	149	
5.5.2	Jeffreys priors *	150	
5.5.3	Robust priors	152	
5.5.4	Mixtures of conjugate priors	152	
5.6	Hierarchical Bayes	154	
5.6.1	Example: modeling related cancer rates	154	
5.7	Empirical Bayes	156	
5.7.1	Example: Beta-Binomial model	156	
5.7.2	Example: Gaussian-Gaussian model	157	
5.8	Bayesian decision theory	160	
5.8.1	Bayes estimators for common loss functions	160	
5.8.2	The false positive vs false negative tradeoff	163	
5.8.3	More general action spaces	167	
5.8.4	Sequential decision theory	168	
6	<i>Frequentist statistics</i>	173	
6.1	Introduction	173	
6.2	Sampling distribution of an estimator	173	
6.2.1	Bootstrap	174	
6.2.2	Large sample theory for the MLE *	175	
6.3	Frequentist decision theory	176	
6.3.1	Bayes risk	177	
6.3.2	Minimax risk	178	
6.3.3	Admissable estimators	179	
6.4	Desirable properties of estimators	182	
6.4.1	Consistent estimators	182	
6.4.2	Unbiased estimators	182	
6.4.3	Minimum variance estimators	183	
6.4.4	The bias-variance tradeoff	183	
6.5	Empirical risk minimization	186	
6.5.1	Regularized risk minimization	187	
6.5.2	Structural risk minimization	188	
6.5.3	Estimating the risk using cross validation	188	
6.5.4	Upper bounding the risk using statistical learning theory *	191	

6.5.5	Surrogate loss functions	192
6.6	Pathologies of frequentist statistics *	193
6.6.1	Undesirable properties of confidence intervals	194
6.6.2	Why isn't everyone a Bayesian?	195
7	Linear regression	197
7.1	Introduction	197
7.2	Model specification	197
7.3	Maximum likelihood estimation (least squares)	197
7.3.1	Derivation of the MLE	199
7.3.2	Geometric interpretation	200
7.3.3	Convexity	201
7.4	Robust linear regression *	203
7.5	Ridge regression	205
7.5.1	Basic idea	205
7.5.2	Numerically stable computation *	207
7.5.3	Connection with PCA *	208
7.5.4	Regularization effects of big data	210
7.6	Bayesian linear regression	211
7.6.1	Computing the posterior	212
7.6.2	Computing the posterior predictive	213
7.6.3	Bayesian inference when σ^2 is unknown *	214
7.6.4	EB for linear regression (evidence procedure)	218
8	Logistic regression	225
8.1	Introduction	225
8.2	Model specification	225
8.3	Model fitting	225
8.3.1	MLE	226
8.3.2	Steepest descent	227
8.3.3	Newton's method	229
8.3.4	Iteratively reweighted least squares (IRLS)	230
8.3.5	Quasi-Newton (variable metric) methods	231
8.3.6	ℓ_2 regularization	232
8.3.7	Multi-class logistic regression	232
8.4	Bayesian logistic regression	234
8.4.1	Gaussian/ Laplace approximation in general	234
8.4.2	Derivation of the BIC	235
8.4.3	Gaussian approximation for logistic regression	235
8.4.4	Approximating the posterior predictive	237
8.4.5	Residual analysis (outlier detection) *	240
8.5	Online learning and stochastic optimization	241
8.5.1	Online learning and regret minimization	241
8.5.2	Stochastic optimization and risk minimization	242
8.5.3	The LMS algorithm	243

8.5.4	The perceptron algorithm	244
8.5.5	A Bayesian view	245
8.6	Generative vs discriminative classifiers	246
8.6.1	Pros and cons of each approach	247
8.6.2	Dealing with missing data	248
8.6.3	Fisher's linear discriminant analysis (FLDA) *	250
9	<i>Generalized linear models and the exponential family</i>	259
9.1	Introduction	259
9.2	The exponential family	259
9.2.1	Definition	260
9.2.2	Examples	260
9.2.3	Log partition function	262
9.2.4	MLE for the exponential family	264
9.2.5	Bayes for the exponential family *	265
9.2.6	Maximum entropy derivation of the exponential family *	267
9.3	Generalized linear models (GLMs)	268
9.3.1	Basics	268
9.3.2	ML and MAP estimation	270
9.3.3	Bayesian inference	271
9.4	Probit regression	271
9.4.1	ML/ MAP estimation using gradient-based optimization	272
9.4.2	Latent variable interpretation	272
9.4.3	Ordinal probit regression *	273
9.4.4	Multinomial probit models *	273
9.5	Multi-task learning and mixed effect GLMs *	274
9.5.1	Basic model	275
9.5.2	Example: semi-parametric GLMMs for medical data	275
9.5.3	Example: discrete choice modeling	276
9.5.4	Other kinds of prior	276
9.5.5	Computational issues	277
10	<i>Graphical models</i>	279
10.1	Introduction	279
10.1.1	Chain rule	279
10.1.2	Conditional independence	280
10.1.3	Graphical models	280
10.1.4	Graph terminology	281
10.2	Directed graphical models (Bayes nets)	282
10.2.1	Representing the joint distribution	282
10.2.2	Example: Naive Bayes classifiers	283
10.2.3	Example: Markov and hidden Markov models	284
10.2.4	Example: Genetic linkage analysis	285
10.2.5	Example: Medical diagnosis	288
10.2.6	Influence (decision) diagrams	289

10.3	Undirected graphical models (Markov random fields)	293
10.3.1	Representing the joint distribution	293
10.3.2	Example: Ising and Potts models	295
10.3.3	Markov logic networks *	298
10.4	Inference	300
10.5	Learning	301
10.5.1	Plate notation	301
10.5.2	Learning DAGs with no latent variables	302
10.5.3	Learning DAGs with latent variables	304
10.5.4	Learning UGMs	304
10.6	CI properties of GMs	304
10.6.1	CI properties of UGMs	305
10.6.2	CI properties of DGMs	306
10.6.3	Comparing directed and undirected graphical models *	311
11	Mixture models and the EM algorithm	317
11.1	Latent variable models	317
11.2	Mixture models	317
11.2.1	Mixtures of Gaussians	319
11.2.2	Mixture of multinoullis	320
11.2.3	Using mixture models for clustering	320
11.2.4	Mixtures of experts	322
11.3	Parameter estimation for mixture models	325
11.3.1	Unidentifiability	326
11.3.2	Computing a MAP estimate is non-convex	327
11.4	The EM algorithm	328
11.4.1	Basic idea	329
11.4.2	EM for GMMs	330
11.4.3	EM for mixture of experts	337
11.4.4	EM for DGMs with hidden variables	338
11.4.5	EM for the Student distribution *	339
11.4.6	EM for probit regression *	342
11.4.7	Theoretical basis for EM *	343
11.4.8	EM variants *	345
11.5	Model selection for latent variable models	347
11.5.1	Model selection for probabilistic models	348
11.5.2	Model selection for non-probabilistic methods	348
11.6	Fitting models with missing data	350
11.6.1	EM for the MLE of an MVN with missing data	351
12	Latent linear models	359
12.1	Factor analysis	359
12.1.1	FA is a low rank parameterization of an MVN	359
12.1.2	Inference of the latent factors	360
12.1.3	Unidentifiability	361

12.1.4	Mixtures of factor analysers	363
12.1.5	EM for factor analysis models	364
12.1.6	Fitting FA models with missing data	365
12.2	Principal components analysis (PCA)	365
12.2.1	Classical PCA: statement of the theorem	365
12.2.2	Proof *	367
12.2.3	Singular value decomposition (SVD)	369
12.2.4	Probabilistic PCA	373
12.2.5	EM algorithm for PCA	374
12.3	Choosing the number of latent dimensions	376
12.3.1	Model selection for FA/ PPCA	376
12.3.2	Model selection for PCA	377
12.4	PCA for categorical data	379
12.5	PCA for paired and multi-view data	381
12.5.1	Supervised PCA (latent factor regression)	382
12.5.2	Partial least squares	384
12.5.3	Canonical correlation analysis	384
12.6	Independent Component Analysis (ICA)	385
12.6.1	Maximum likelihood estimation	388
12.6.2	The FastICA algorithm	388
12.6.3	Using EM	391
12.6.4	Other estimation principles *	392
13	<i>Sparse linear models</i>	399
13.1	Introduction	399
13.2	Bayesian variable selection	400
13.2.1	The spike and slab model	402
13.2.2	From the Bernoulli-Gaussian model to ℓ_0 regularization	403
13.2.3	Algorithms	404
13.3	ℓ_1 regularization: basics	407
13.3.1	Why does ℓ_1 regularization yield sparse solutions?	408
13.3.2	Optimality conditions for lasso	409
13.3.3	Comparison of least squares, lasso, ridge and subset selection	413
13.3.4	Regularization path	414
13.3.5	Model selection	417
13.3.6	Bayesian inference for linear models with Laplace priors	418
13.4	ℓ_1 regularization: algorithms	419
13.4.1	Coordinate descent	419
13.4.2	LARS and other homotopy methods	419
13.4.3	Proximal and gradient projection methods	420
13.4.4	EM for lasso	425
13.5	ℓ_1 regularization: extensions	427
13.5.1	Group Lasso	427
13.5.2	Fused lasso	432
13.5.3	Elastic net (ridge and lasso combined)	433

13.6	Non-convex regularizers	435
13.6.1	Bridge regression	436
13.6.2	Hierarchical adaptive lasso	436
13.6.3	Other hierarchical priors	440
13.7	Automatic relevance determination (ARD)/ sparse Bayesian learning (SBL)	441
13.7.1	ARD for linear regression	441
13.7.2	Whence sparsity?	443
13.7.3	Connection to MAP estimation	443
13.7.4	Algorithms for ARD *	444
13.7.5	ARD for logistic regression	446
13.8	Sparse coding *	446
13.8.1	Learning a sparse coding dictionary	447
13.8.2	Results of dictionary learning from image patches	448
13.8.3	Compressed sensing	450
13.8.4	Image inpainting and denoising	450

14 Kernels 457

14.1	Introduction	457
14.2	Kernel functions	457
14.2.1	RBF kernels	458
14.2.2	Kernels for comparing documents	458
14.2.3	Mercer (positive definite) kernels	459
14.2.4	Linear kernels	460
14.2.5	Matern kernels	460
14.2.6	String kernels	461
14.2.7	Pyramid match kernels	462
14.2.8	Kernels derived from probabilistic generative models	463
14.3	Using kernels inside GLMs	464
14.3.1	Kernel machines	464
14.3.2	LIVMs, RVMs, and other sparse kernel machines	465
14.4	The kernel trick	466
14.4.1	Kernelized nearest neighbor classification	467
14.4.2	Kernelized K-medoids clustering	467
14.4.3	Kernelized ridge regression	470
14.4.4	Kernel PCA	471
14.5	Support vector machines (SVMs)	474
14.5.1	SVMs for regression	475
14.5.2	SVMs for classification	476
14.5.3	Choosing C	482
14.5.4	Summary of key points	482
14.5.5	A probabilistic interpretation of SVMs	483
14.6	Comparison of discriminative kernel methods	483
14.7	Kernels for building generative models	485
14.7.1	Smoothing kernels	485
14.7.2	Kernel density estimation (KDE)	486

14.7.3	From KDE to KNN	488
14.7.4	Kernel regression	488
14.7.5	Locally weighted regression	490
15	<i>Gaussian processes</i>	493
15.1	Introduction	493
15.2	GPs for regression	494
15.2.1	Predictions using noise-free observations	495
15.2.2	Predictions using noisy observations	496
15.2.3	Effect of the kernel parameters	497
15.2.4	Estimating the kernel parameters	499
15.2.5	Computational and numerical issues *	502
15.2.6	Semi-parametric GPs *	502
15.3	GPs meet GLMs	503
15.3.1	Binary classification	503
15.3.2	Multi-class classification	506
15.3.3	GPs for Poisson regression	509
15.4	Connection with other methods	510
15.4.1	Linear models compared to GPs	510
15.4.2	Linear smoothers compared to GPs	511
15.4.3	SVMs compared to GPs	512
15.4.4	L1VM and RVMs compared to GPs	512
15.4.5	Neural networks compared to GPs	513
15.4.6	Smoothing splines compared to GPs *	514
15.4.7	RKHS methods compared to GPs *	516
15.5	GP latent variable model	518
15.6	Approximation methods for large datasets	520
16	<i>Adaptive basis function models</i>	521
16.1	Introduction	521
16.2	Classification and regression trees (CART)	522
16.2.1	Basics	522
16.2.2	Growing a tree	524
16.2.3	Pruning a tree	527
16.2.4	Pros and cons of trees	528
16.2.5	Random forests	528
16.2.6	CART compared to hierarchical mixture of experts *	529
16.3	Generalized additive models	530
16.3.1	Backfitting	530
16.3.2	Computational efficiency	531
16.3.3	Multivariate adaptive regression splines (MARS)	531
16.4	Boosting	532
16.4.1	Forward stagewise additive modeling	533
16.4.2	L2boosting	536
16.4.3	AdaBoost	536

16.4.4	LogitBoost	537	
16.4.5	Boosting as functional gradient descent	538	
16.4.6	Sparse boosting	539	
16.4.7	Multivariate adaptive regression trees (MART)	540	
16.4.8	Why does boosting work so well?	540	
16.4.9	A Bayesian view	541	
16.5	Feedforward neural networks (multilayer perceptrons)	541	
16.5.1	Convolutional neural networks	542	
16.5.2	Other kinds of neural networks	546	
16.5.3	A brief history of the field	546	
16.5.4	The backpropagation algorithm	548	
16.5.5	Identifiability	550	
16.5.6	Regularization	550	
16.5.7	Bayesian inference *	554	
16.6	Experimental comparison	558	
16.6.1	Low-dimensional features	558	
16.6.2	High-dimensional features	559	
16.7	Interpreting black-box models	561	
17	Markov and hidden Markov Models	565	
17.1	Introduction	565	
17.2	Markov models	565	
17.2.1	Transition matrix	565	
17.2.2	Application: Language modeling	567	
17.2.3	Stationary distribution of a Markov chain *	572	
17.2.4	Application: Google's PageRank algorithm for web page ranking *	576	
17.3	Hidden Markov models	579	
17.3.1	Applications of HMMs	580	
17.4	Inference in HMMs	582	
17.4.1	Types of inference problems for temporal models	582	
17.4.2	The forwards algorithm	585	
17.4.3	The forwards-backwards algorithm	586	
17.4.4	The Viterbi algorithm	588	
17.4.5	Forwards filtering, backwards sampling	592	
17.5	Learning for HMMs	593	
17.5.1	Training with fully observed data	593	
17.5.2	EM for HMMs (the Baum-Welch algorithm)	594	
17.5.3	Bayesian methods for "fitting" HMMs *	596	
17.5.4	Discriminative training	596	
17.5.5	Model selection	597	
17.6	Generalizations of HMMs	597	
17.6.1	Variable duration (semi-Markov) HMMs	597	
17.6.2	Hierarchical HMMs	600	
17.6.3	Input-output HMMs	601	
17.6.4	Auto-regressive and buried HMMs	602	

17.6.5	Factorial HMM	604	
17.6.6	Coupled HMM and the influence model	605	
17.6.7	Dynamic Bayesian networks (DBNs)	605	
18	<i>State space models</i>	609	
18.1	Introduction	609	
18.2	Applications of SSMs	610	
18.2.1	SSMs for tracking	610	
18.2.2	Robotic SLAM	611	
18.2.3	Online parameter learning using recursive least squares	614	
18.2.4	SSM for time series forecasting	615	
18.3	Inference in LG-SSM	620	
18.3.1	The Kalman filtering algorithm	620	
18.3.2	The Kalman smoothing algorithm	622	
18.4	Learning for LG-SSM	624	
18.4.1	Identifiability and numerical stability	625	
18.4.2	Training with fully observed data	625	
18.4.3	EM for LG-SSM	625	
18.4.4	Subspace methods	625	
18.4.5	Bayesian methods for “fitting” LG-SSMs	626	
18.5	Approximate online inference for non-linear, non-Gaussian SSMs	626	
18.5.1	Extended Kalman filter (EKF)	627	
18.5.2	Unscented Kalman filter (UKF)	629	
18.5.3	Assumed density filtering (ADF)	631	
18.6	Hybrid discrete/ continuous SSMs	633	
18.6.1	Inference	634	
18.6.2	Application: Data association and multi target tracking	636	
18.6.3	Application: fault diagnosis	637	
18.6.4	Application: econometric forecasting	638	
19	<i>Markov random fields revisited</i>	639	
19.1	Introduction	639	
19.1.1	Chain-structured CRFs, MEMMs and the label-bias problem	640	
19.1.2	Representing the potentials	642	
19.2	Applications of CRFs	643	
19.2.1	CRFs for natural language processing	644	
19.2.2	CRFs for computational biology	646	
19.2.3	CRFs for computer vision	646	
19.3	Inference	651	
19.4	Learning	651	
19.4.1	Training fully observed CRFs	652	
19.4.2	Training partially observed CRFs	654	
19.4.3	Pseudo likelihood	655	
19.4.4	Stochastic Maximum Likelihood	656	
19.4.5	Feature induction for maxent models	657	

19.4.6	Iterative proportional fitting (IPF) for fitting MRFs *	659
19.5	Structural SVMs and max margin Markov networks *	661
19.6	Gaussian graphical models *	661
19.6.1	Undirected GGMs	662
19.6.2	Directed GGMs	662
19.6.3	Comparing Gaussian DAGs and MRFs	663
19.6.4	Inference	665
19.6.5	Learning	665
20	<i>Exact inference algorithms for graphical models</i>	669
20.1	Introduction	669
20.2	Belief propagation for trees	669
20.2.1	Serial protocol	669
20.2.2	Parallel protocol	671
20.2.3	Gaussian BP *	672
20.2.4	Other BP variants *	674
20.3	The variable elimination algorithm	676
20.3.1	The generalized distributive law *	679
20.3.2	Computational complexity of VE	680
20.3.3	A weakness of VE	682
20.4	The junction tree algorithm *	682
20.4.1	Creating a junction tree	684
20.4.2	Message passing on a junction tree	684
20.4.3	Computational complexity of JTA	687
20.4.4	JTA generalizations *	688
20.5	Computational intractability of exact inference in the worst case	688
20.5.1	Approximate inference	689
21	<i>Variational inference</i>	693
21.1	Introduction	693
21.2	Variational inference	694
21.2.1	Forward or reverse KL? *	695
21.3	The mean field method	697
21.3.1	Derivation of the mean field update equations	698
21.3.2	Example: Mean field for the Ising model	699
21.4	Structured mean field *	700
21.4.1	Example: factorial HMM	701
21.5	Variational Bayes	703
21.5.1	Example: VB for a univariate Gaussian	703
21.5.2	Example: VB for linear regression	707
21.6	Variational Bayes EM	710
21.6.1	Example: VBEM for mixtures of Gaussians *	711
21.7	Variational message passing and VIBES	717
21.8	Local variational bounds	717
21.8.1	Motivating applications	717

21.8.2	Bohning's quadratic bound to the log-sum-exp function	718
21.8.3	Bounds for the sigmoid function	721
21.8.4	Other bounds and approximations to the log-sum-exp function *	723
21.8.5	Variational inference based on upper bounds	724
22	<i>Belief propagation revisited</i>	729
22.1	Introduction	729
22.2	Loopy belief propagation: algorithmic issues	729
22.2.1	LBP on pairwise models	730
22.2.2	LBP on a factor graph	730
22.2.3	Convergence	732
22.2.4	Other speedup tricks for BP	736
22.2.5	Accuracy of LBP	737
22.2.6	Application: LBP for error correcting codes	737
22.3	Loopy belief propagation: theoretical issues *	738
22.3.1	UGMs represented in exponential family form	738
22.3.2	The marginal polytope	739
22.3.3	Exact inference as a variational optimization problem	740
22.3.4	Mean field as a variational optimization problem	741
22.3.5	LBP as a variational optimization problem	742
22.3.6	Loopy BP vs mean field	745
22.4	Extensions of belief propagation *	746
22.4.1	Generalized belief propagation	746
22.4.2	Convex belief propagation	747
22.5	Expectation propagation	750
22.6	MAP state estimation	750
22.6.1	Linear programming relaxation	751
22.6.2	Max-product belief propagation	751
22.6.3	Dual decomposition	752
22.6.4	Submodularity	755
22.6.5	Graphcuts	756
23	<i>Monte Carlo inference</i>	759
23.1	Introduction	759
23.2	Sampling from standard distributions	759
23.2.1	Using the cdf	759
23.2.2	Sampling from a Gaussian (Box-Muller method)	760
23.3	Rejection sampling	761
23.3.1	Basic idea	761
23.3.2	Example	762
23.3.3	Application to Bayesian statistics	763
23.3.4	Adaptive rejection sampling	763
23.3.5	Rejection sampling in high dimensions	764
23.4	Importance sampling	764
23.4.1	Basic idea	764

23.4.2	Handling unnormalized distributions	765	
23.4.3	Importance sampling for a DGM: Likelihood weighting	765	
23.4.4	Sampling importance resampling (SIR)	766	
23.5	Particle filtering	767	
23.5.1	Sequential importance sampling	767	
23.5.2	The degeneracy problem	768	
23.5.3	The resampling step	769	
23.5.4	The proposal distribution	771	
23.5.5	Application: Robot localization	772	
23.5.6	Application: Visual object tracking	773	
23.5.7	Application: time series forecasting	775	
23.6	Rao-Blackwellised particle filtering (RBPF)	775	
23.6.1	RBPF for switching LG-SSMs	775	
23.6.2	Application: Tracking a maneuvering target	777	
23.6.3	Application: Fast SLAM	779	
23.7	Approximating the marginal likelihood	779	
23.7.1	The candidate method	780	
23.7.2	Annealed importance sampling	780	
24	Markov Chain Monte Carlo (MCMC) inference	783	
24.1	Introduction	783	
24.2	Gibbs sampling	783	
24.2.1	Basic idea	783	
24.2.2	Example: Gibbs sampling for the Ising model	784	
24.2.3	Example: Gibbs sampling for inferring the parameters of a GMM	785	
24.2.4	Collapsed Gibbs sampling *	787	
24.2.5	Gibbs sampling for hierarchical GLMs	790	
24.2.6	BUGS and JAGS	792	
24.2.7	The Imputation Posterior (IP) algorithm	792	
24.2.8	Blocking Gibbs sampling	793	
24.3	Metropolis Hastings algorithm	793	
24.3.1	Basic idea	793	
24.3.2	Gibbs sampling is a special case of MH	794	
24.3.3	Proposal distributions	795	
24.3.4	Adaptive MCMC	798	
24.3.5	Initialization and mode hopping	799	
24.3.6	Why MH works *	799	
24.3.7	Reversible jump (trans-dimensional) MCMC *	800	
24.4	Speed and accuracy of MCMC	800	
24.4.1	The burn-in phase	801	
24.4.2	Mixing rates of Markov chains *	802	
24.4.3	Practical convergence diagnostics	803	
24.4.4	Accuracy of MCMC	805	
24.4.5	How many chains?	807	
24.5	Auxiliary variable MCMC *	808	

24.5.1	Auxiliary variable sampling for logistic regression	808
24.5.2	Slice sampling	809
24.5.3	Swendsen Wang	811
24.5.4	Hybrid/ Hamiltonian MCMC *	813
24.6	Simulated annealing	813
25	Clustering	819
25.1	Introduction	819
25.1.1	Measuring (dis)similarity	819
25.1.2	Evaluating the output of clustering methods *	820
25.2	Dirichlet process mixture models	823
25.2.1	From finite to infinite mixture models	823
25.2.2	The Dirichlet process	826
25.2.3	Applying Dirichlet processes to mixture modeling	829
25.2.4	Fitting a DP mixture model	830
25.3	Affinity propagation	831
25.4	Spectral clustering	834
25.4.1	Graph Laplacian	835
25.4.2	Normalized graph Laplacian	836
25.4.3	Example	837
25.5	Hierarchical clustering	837
25.5.1	Agglomerative clustering	838
25.5.2	Divisive clustering	841
25.5.3	Choosing the number of clusters	842
25.5.4	Bayesian hierarchical clustering	842
25.6	Clustering datapoints and features	846
25.6.1	Biclustering	846
25.6.2	Multi-view clustering	847
26	Graphical model structure learning	851
26.1	Introduction	851
26.2	Quick and dirty ways to learn graph structure	852
26.2.1	Relevance networks	852
26.2.2	Dependency networks	853
26.3	Learning tree structures	854
26.3.1	Directed or undirected tree?	855
26.3.2	Chow-Liu algorithm for finding the ML tree structure	856
26.3.3	Finding the MAP forest	856
26.3.4	Mixtures of trees	858
26.4	Learning DAG structures	858
26.4.1	Exact structural inference	858
26.4.2	Scaling up to larger graphs	864
26.5	Learning DAG structure with latent variables	866
26.5.1	Approximating the marginal likelihood when we have missing data	866
26.5.2	Structural EM	869

26.5.3	Discovering hidden variables	869
26.5.4	Case study: Google's Rephil	871
26.5.5	Structural equation models *	873
26.6	Learning causal DAGs	875
26.6.1	Causal interpretation of DAGs	875
26.6.2	Using causal DAGs to resolve Simpson's paradox	877
26.6.3	Learning causal DAG structures	879
26.7	Learning undirected Gaussian graphical models	882
26.7.1	MLE for a GRF	882
26.7.2	Graphical lasso	884
26.7.3	Bayesian inference for GRF structure	885
26.8	Learning undirected discrete graphical models	887
26.8.1	Graphical lasso for MRFs/ CRFs	887
26.8.2	Thin junction trees	888
27	<i>Latent variable models for discrete data</i>	891
27.1	Introduction	891
27.2	IVMs for modeling vectors and bags of counts and tokens	892
27.2.1	Mixture models	892
27.2.2	Exponential family PCA	893
27.2.3	LDA and mPCA	894
27.2.4	GaP model and non-negative matrix factorization	895
27.3	Latent Dirichlet allocation (LDA)	896
27.3.1	Basics	896
27.3.2	Quantitatively evaluating LDA as a language model	899
27.3.3	Fitting using (collapsed) Gibbs sampling	901
27.3.4	Example	902
27.3.5	Fitting using batch variational inference	903
27.3.6	Fitting using online variational inference	905
27.3.7	Determining the number of topics	906
27.4	Extensions of LDA	907
27.4.1	Correlated topic model	907
27.4.2	Dynamic topic model	908
27.4.3	LDA-HMM	909
27.4.4	Supervised LDA *	911
27.5	IVMs for graph-structured data	916
27.5.1	Stochastic block model	917
27.5.2	Mixed membership stochastic block model	919
27.5.3	Relational topic model	920
27.6	IVMs for relational data	921
27.6.1	Infinite relational model	922
27.6.2	Probabilistic matrix factorization for collaborative filtering	925
27.7	Restricted Boltzmann machines (RBMs)	929
27.7.1	Varieties of RBMs	931
27.7.2	Learning RBMs	933

27.7.3	Applications of RBMs	937	
28	Deep learning	941	
28.1	Introduction	941	
28.2	Deep generative models	942	
28.2.1	Deep sigmoid networks	942	
28.2.2	Deep Boltzmann machines	943	
28.2.3	Deep belief networks	944	
28.3	Training deep networks	945	
28.3.1	Greedy layer-wise learning of DBNs	945	
28.3.2	Fitting deep neural nets	947	
28.3.3	Fitting deep auto-encoders	947	
28.3.4	Stacked denoising auto-encoders	948	
28.4	Applications of deep networks	948	
28.4.1	Handwritten digit classification using DBNs	948	
28.4.2	Data visualization using deep auto-encoders	949	
28.4.3	Information retrieval using deep autoencoders (semantic hashing)	950	
28.4.4	Learning audio features using 1d convolutional DBNs	951	
28.4.5	Learning image features using 2d convolutional DBNs	952	
28.5	Discussion	953	
Bibliography	955		
Index to code	955		
Index to code	955		
Index to keywords	957		