

Contents

Contents	v	2.5.3	Multivariate Student's t-distribution	10
Notation	ix	2.5.4	Dirichlet distribution	10
1 Introduction	1	2.6	Transformations of random variables ...	11
1.1 Types of machine learning	1	2.6.1	Linear transformations	11
1.2 Three elements of a machine learning model	1	2.6.2	General transformations	11
1.2.1 Representation	1	2.6.3	Central limit theorem	13
1.2.2 Evaluation	1	2.7	Monte Carlo approximation	13
1.2.3 Optimization	2	2.8	Information theory	14
1.3 Some basic concepts	2	2.8.1	Entropy	14
1.3.1 Parametric vs non-parametric models	2	2.8.2	KL divergence	14
1.3.2 A simple non-parametric classifier: K-nearest neighbours	2	2.8.3	Mutual information	14
1.3.3 Overfitting	2	3 Generative models for discrete data		17
1.3.4 Cross validation	2	3.1	Generative classifier	17
1.3.5 Model selection	2	3.2	Bayesian concept learning	17
2 Probability	3	3.2.1	Likelihood	17
2.1 Frequentists vs. Bayesians	3	3.2.2	Prior	17
2.2 A brief review of probability theory ...	3	3.2.3	Posterior	17
2.2.1 Basic concepts	3	3.2.4	Posterior predictive distribution	18
2.2.2 Mutivariate random variables ..	3	3.3	The beta-binomial model	18
2.2.3 Bayes rule	4	3.3.1	Likelihood	18
2.2.4 Independence and conditional independence	4	3.3.2	Prior	18
2.2.5 Quantiles	4	3.3.3	Posterior	18
2.2.6 Mean and variance	4	3.3.4	Posterior predictive distribution	19
2.3 Some common discrete distributions ...	5	3.4	The Dirichlet-multinomial model	19
2.3.1 The Bernoulli and binomial distributions	5	3.4.1	Likelihood	20
2.3.2 The multinoulli and multinomial distributions	5	3.4.2	Prior	20
2.3.3 The Poisson distribution	5	3.4.3	Posterior	20
2.3.4 The empirical distribution ...	5	3.4.4	Posterior predictive distribution	20
2.4 Some common continuous distributions ..	6	3.5	Naive Bayes classifiers	20
2.4.1 Gaussian (normal) distribution ..	6	3.5.1	Optimization	21
2.4.2 Student's t-distribution	6	3.5.2	Using the model for prediction	21
2.4.3 The Laplace distribution	7	3.5.3	The log-sum-exp trick	21
2.4.4 The gamma distribution	8	3.5.4	Feature selection using mutual information	22
2.4.5 The beta distribution	8	3.5.5	Classifying documents using bag of words	22
2.4.6 Pareto distribution	8	4 Gaussian Models		25
2.5 Joint probability distributions	9	4.1	Basics	25
2.5.1 Covariance and correlation ...	9	4.1.1	MLE for a MVN	25
2.5.2 Multivariate Gaussian distribution	10	4.1.2	Maximum entropy derivation of the Gaussian *	26
		4.2	Gaussian discriminant analysis	26
		4.2.1	Quadratic discriminant analysis (QDA)	26

4.2.2	Linear discriminant analysis (LDA)	27	6.2	Frequentist decision theory	39
4.2.3	Two-class LDA	28	6.3	Desirable properties of estimators	39
4.2.4	MLE for discriminant analysis	28	6.4	Empirical risk minimization	39
4.2.5	Strategies for preventing overfitting	29	6.4.1	Regularized risk minimization	39
4.2.6	Regularized LDA *	29	6.4.2	Structural risk minimization	39
4.2.7	Diagonal LDA	29	6.4.3	Estimating the risk using cross validation	39
4.2.8	Nearest shrunken centroids classifier *	29	6.4.4	Upper bounding the risk using statistical learning theory *	39
4.3	Inference in jointly Gaussian distributions	29	6.4.5	Surrogate loss functions	39
4.3.1	Statement of the result	29	6.5	Pathologies of frequentist statistics *	39
4.3.2	Examples	30	7	Linear Regression	41
4.4	Linear Gaussian systems	30	7.1	Introduction	41
4.4.1	Statement of the result	30	7.2	Representation	41
4.5	Digression: The Wishart distribution *	30	7.3	MLE	41
4.6	Inferring the parameters of an MVN	30	7.3.1	OLS	41
4.6.1	Posterior distribution of μ	30	7.3.2	SGD	42
4.6.2	Posterior distribution of Σ *	30	7.4	Ridge regression(MAP)	42
4.6.3	Posterior distribution of μ and Σ *	30	7.4.1	Basic idea	43
4.6.4	Sensor fusion with unknown precisions *	30	7.4.2	Numerically stable computation *	43
5	Bayesian statistics	31	7.4.3	Connection with PCA *	43
5.1	Introduction	31	7.4.4	Regularization effects of big data	43
5.2	Summarizing posterior distributions	31	7.5	Bayesian linear regression	43
5.2.1	MAP estimation	31	8	Logistic Regression	45
5.2.2	Credible intervals	32	8.1	Representation	45
5.2.3	Inference for a difference in proportions	33	8.2	Optimization	45
5.3	Bayesian model selection	33	8.2.1	MLE	45
5.3.1	Bayesian Occam's razor	33	8.2.2	MAP	45
5.3.2	Computing the marginal likelihood (evidence)	34	8.3	Multinomial logistic regression	45
5.3.3	Bayes factors	36	8.3.1	Representation	45
5.4	Priors	36	8.3.2	MLE	46
5.4.1	Uninformative priors	36	8.3.3	MAP	46
5.4.2	Robust priors	36	8.4	Bayesian logistic regression	46
5.4.3	Mixtures of conjugate priors	36	8.4.1	Laplace approximation	47
5.5	Hierarchical Bayes	36	8.4.2	Derivation of the BIC	47
5.6	Empirical Bayes	36	8.4.3	Gaussian approximation for logistic regression	47
5.7	Bayesian decision theory	36	8.4.4	Approximating the posterior predictive	47
5.7.1	Bayes estimators for common loss functions	37	8.4.5	Residual analysis (outlier detection) *	47
5.7.2	The false positive vs false negative tradeoff	38	8.5	Online learning and stochastic optimization	47
6	Frequentist statistics	39	8.5.1	The perceptron algorithm	47
6.1	Sampling distribution of an estimator	39	8.6	Generative vs discriminative classifiers	48
6.1.1	Bootstrap	39	8.6.1	Pros and cons of each approach	48
6.1.2	Large sample theory for the MLE *	39	8.6.2	Dealing with missing data	48
			8.6.3	Fishers linear discriminant analysis (FLDA) *	50

9	Generalized linear models and the exponential family	51	11.3.2	Computing a MAP estimate is non-convex	60
9.1	The exponential family	51	11.4	The EM algorithm	60
9.1.1	Definition	51	11.4.1	Introduction	60
9.1.2	Examples	51	11.4.2	Basic idea	62
9.1.3	Log partition function	52	11.4.3	EM for GMMs	62
9.1.4	MLE for the exponential family	53	11.4.4	EM for K-means	64
9.1.5	Bayes for the exponential family	53	11.4.5	EM for mixture of experts	64
9.1.6	Maximum entropy derivation of the exponential family *	53	11.4.6	EM for DGMs with hidden variables	64
9.2	Generalized linear models (GLMs)	53	11.4.7	EM for the Student distribution *	64
9.2.1	Basics	53	11.4.8	EM for probit regression *	64
9.3	Probit regression	53	11.4.9	Derivation of the Q function	64
9.4	Multi-task learning	53	11.4.10	Convergence of the EM Algorithm *	65
10	Directed graphical models (Bayes nets)	55	11.4.11	Generalization of EM Algorithm *	65
10.1	Introduction	55	11.4.12	Online EM	66
10.1.1	Chain rule	55	11.4.13	Other EM variants *	66
10.1.2	Conditional independence	55	11.5	Model selection for latent variable models	66
10.1.3	Graphical models	55	11.5.1	Model selection for probabilistic models	67
10.1.4	Directed graphical model	55	11.5.2	Model selection for non-probabilistic methods	67
10.2	Examples	56	11.6	Fitting models with missing data	67
10.2.1	Naive Bayes classifiers	56	11.6.1	EM for the MLE of an MVN with missing data	67
10.2.2	Markov and hidden Markov models	56	12	Latent linear models	69
10.3	Inference	56	12.1	Factor analysis	69
10.4	Learning	56	12.1.1	FA is a low rank parameterization of an MVN	69
10.4.1	Learning from complete data	56	12.1.2	Inference of the latent factors	69
10.4.2	Learning with missing and/or latent variables	57	12.1.3	Unidentifiability	70
10.5	Conditional independence properties of DGMs	57	12.1.4	Mixtures of factor analysers	70
10.5.1	d-separation and the Bayes Ball algorithm (global Markov properties)	57	12.1.5	EM for factor analysis models	71
10.5.2	Other Markov properties of DGMs	57	12.1.6	Fitting FA models with missing data	71
10.5.3	Markov blanket and full conditionals	57	12.2	Principal components analysis (PCA)	71
10.5.4	Multinoulli Learning	57	12.2.1	Classical PCA	71
10.6	Influence (decision) diagrams *	57	12.2.2	Singular value decomposition (SVD)	72
11	Mixture models and the EM algorithm	59	12.2.3	Probabilistic PCA	73
11.1	Latent variable models	59	12.2.4	EM algorithm for PCA	74
11.2	Mixture models	59	12.3	Choosing the number of latent dimensions	74
11.2.1	Mixtures of Gaussians	59	12.3.1	Model selection for FA/PPCA	74
11.2.2	Mixtures of multinoullis	60	12.3.2	Model selection for PCA	74
11.2.3	Using mixture models for clustering	60	12.4	PCA for categorical data	74
11.2.4	Mixtures of experts	60	12.5	PCA for paired and multi-view data	75
11.3	Parameter estimation for mixture models	60	12.5.1	Supervised PCA (latent factor regression)	75
11.3.1	Unidentifiability	60			

12.5.2	Discriminative supervised PCA	75	16 Adaptive basis function models	89
12.5.3	Canonical correlation analysis	75	16.1 AdaBoost	89
12.6	Independent Component Analysis (ICA)	75	16.1.1 Representation	89
12.6.1	Maximum likelihood estimation	75	16.1.2 Evaluation	89
12.6.2	The FastICA algorithm	76	16.1.3 Optimization	89
12.6.3	Using EM	76	16.1.4 The upper bound of the training error of AdaBoost	89
12.6.4	Other estimation principles *	76		
13	Sparse linear models	77	17 Hidden markov Model	91
14	Kernels	79	17.1 Introduction	91
14.1	Introduction	79	17.2 Markov models	91
14.2	Kernel functions	79	18 State space models	93
14.2.1	RBF kernels	79	19 Undirected graphical models (Markov random fields)	95
14.2.2	TF-IDF kernels	79	20 Exact inference for graphical models	97
14.2.3	Mercer (positive definite) kernels	79	21 Variational inference	99
14.2.4	Linear kernels	80	22 More variational inference	101
14.2.5	Matern kernels	80	23 Monte Carlo inference	103
14.2.6	String kernels	80	24 Markov chain Monte Carlo (MCMC) inference	105
14.2.7	Pyramid match kernels	81	24.1 Introduction	105
14.2.8	Kernels derived from probabilistic generative models	81	24.2 Metropolis Hastings algorithm	105
14.3	Using kernels inside GLMs	81	24.3 Gibbs sampling	105
14.3.1	Kernel machines	81	24.4 Speed and accuracy of MCMC	105
14.3.2	L1VMs, RVMs, and other sparse vector machines	81	24.5 Auxiliary variable MCMC *	105
14.4	The kernel trick	81	25 Clustering	107
14.4.1	Kernelized KNN	82	26 Graphical model structure learning	109
14.4.2	Kernelized K-medoids clustering	82	27 Latent variable models for discrete data	111
14.4.3	Kernelized ridge regression	82	27.1 Introduction	111
14.4.4	Kernel PCA	83	27.2 Distributed state LVMs for discrete data	111
14.5	Support vector machines (SVMs)	83	28 Deep learning	113
14.5.1	SVMs for classification	83	A Optimization methods	115
14.5.2	SVMs for regression	84	A.1 Convexity	115
14.5.3	Choosing C	85	A.2 Gradient descent	115
14.5.4	A probabilistic interpretation of SVMs	85	A.2.1 Stochastic gradient descent	115
14.5.5	Summary of key points	85	A.2.2 Batch gradient descent	115
14.6	Comparison of discriminative kernel methods	86	A.2.3 Line search	115
14.7	Kernels for building generative models	86	A.2.4 Momentum term	116
15	Gaussian processes	87	A.3 Lagrange duality	116
15.1	Introduction	87	A.3.1 Primal form	116
15.2	GPs for regression	87	A.3.2 Dual form	116
15.3	GPs meet GLMs	87	A.4 Newton's method	116
15.4	Connection with other methods	87	A.5 Quasi-Newton method	116
15.5	GP latent variable model	87	A.5.1 DFP	116
15.6	Approximation methods for large datasets	87		

A.5.2	BFGS	116	Glossary	119
A.5.3	Broyden	117		