
Contents

List of Tables	xi
List of Figures	xiii
Preface	xix
1 Linear Modelling: A Least Squares Approach	1
1.1 Linear modelling	1
1.1.1 Defining the model	2
1.1.2 Modelling assumptions	3
1.1.3 Defining what a <i>good</i> model is	4
1.1.4 The least squares solution – a worked example	6
1.1.5 Worked example	9
1.1.6 Least squares fit to the Olympics data	10
1.1.7 Summary	11
1.2 Making predictions	12
1.2.1 A second Olympics dataset	12
1.2.2 Summary	15
1.3 Vector/matrix notation	15
1.3.1 Example	22
1.3.2 Numerical example	23
1.3.3 Making predictions	24
1.3.4 Summary	24
1.4 Nonlinear response from a linear model	25
1.5 Generalisation and over-fitting	28
1.5.1 Validation data	29
1.5.2 Cross-validation	29
1.5.3 Computational scaling of K -fold cross-validation	32
1.6 Regularised least squares	33
1.7 Exercises	35
Further reading	37
2 Linear Modelling: A Maximum Likelihood Approach	39
2.1 Errors as noise	39
2.1.1 Thinking generatively	40
2.2 Random variables and probability	41

2.2.1	Random variables	41
2.2.2	Probability and distributions	42
2.2.3	Adding probabilities	44
2.2.4	Conditional probabilities	44
2.2.5	Joint probabilities	45
2.2.6	Marginalisation	47
2.2.7	Aside – Bayes' rule	49
2.2.8	Expectations	50
2.3	Popular discrete distributions	53
2.3.1	Bernoulli distribution	53
2.3.2	Binomial distribution	53
2.3.3	Multinomial distribution	54
2.4	Continuous random variables – density functions	55
2.5	Popular continuous density functions	58
2.5.1	The uniform density function	58
2.5.2	The beta density function	60
2.5.3	The Gaussian density function	61
2.5.4	Multivariate Gaussian	62
2.5.5	Summary	65
2.6	Thinking generatively...continued	65
2.7	Likelihood	67
2.7.1	Dataset likelihood	68
2.7.2	Maximum likelihood	69
2.7.3	Characteristics of the maximum likelihood solution	71
2.7.4	Maximum likelihood favours complex models	74
2.8	The bias-variance tradeoff	75
2.8.1	Summary	75
2.9	Effect of noise on parameter estimates	76
2.9.1	Uncertainty in estimates	78
2.9.2	Comparison with empirical values	81
2.9.3	Variability in model parameters – Olympics data	82
2.10	Variability in predictions	83
2.10.1	Predictive variability – an example	85
2.10.2	Expected values of the estimators	86
2.10.3	Summary	90
2.11	Exercises	90
	Further reading	93
3	The Bayesian Approach to Machine Learning	95
3.1	A coin game	95
3.1.1	Counting heads	97
3.1.2	The Bayesian way	98
3.2	The exact posterior	103
3.3	The three scenarios	104
3.3.1	No prior knowledge	104

3.3.2	The fair coin scenario	111
3.3.3	A biased coin	114
3.3.4	The three scenarios – a summary	116
3.3.5	Adding more data	116
3.4	Marginal likelihoods	117
3.4.1	Model comparison with the marginal likelihood	118
3.5	Hyper-parameters	119
3.6	Graphical models	120
3.6.1	Summary	121
3.7	A Bayesian treatment of the Olympics 100 m data	122
3.7.1	The model	122
3.7.2	The likelihood	124
3.7.3	The prior	124
3.7.4	The posterior	124
3.7.5	A first-order polynomial	126
3.7.6	Making predictions	129
3.8	Marginal likelihood for polynomial model order selection	131
3.9	Chapter summary	133
3.10	Exercises	133
	Further reading	137
4	Bayesian Inference	139
4.1	Non-conjugate models	139
4.2	Binary responses	140
4.2.1	A model for binary responses	140
4.3	A point estimate – the MAP solution	143
4.4	The Laplace approximation	149
4.4.1	Laplace approximation example: Approximating a gamma density	150
4.4.2	Laplace approximation for the binary response model	151
4.5	Sampling techniques	154
4.5.1	Playing darts	154
4.5.2	The Metropolis–Hastings algorithm	156
4.5.3	The art of sampling	164
4.6	Summary	165
4.7	Exercises	165
	Further reading	167
5	Classification	169
5.1	The general problem	169
5.2	Probabilistic classifiers	170
5.2.1	The Bayes classifier	170
5.2.1.1	Likelihood – class-conditional distributions	171
5.2.1.2	Prior class distribution	171
5.2.1.3	Example – Gaussian class-conditionals	172

5.2.1.4	Making predictions	173
5.2.1.5	The naive Bayes assumption	175
5.2.1.6	Example – classifying text	175
5.2.1.7	Smoothing	177
5.2.2	Logistic regression	179
5.2.2.1	Motivation	180
5.2.2.2	Nonlinear decision functions	181
5.2.2.3	Nonparametric models – the Gaussian process	182
5.3	Nonprobabilistic classifiers	183
5.3.1	K -nearest neighbours	183
5.3.1.1	Choosing K	184
5.3.2	Support vector machines and other kernel methods	186
5.3.2.1	The margin	186
5.3.2.2	Maximising the margin	187
5.3.2.3	Making predictions	190
5.3.2.4	Support vectors	191
5.3.2.5	Soft margins	192
5.3.2.6	Kernels	193
5.3.3	Summary	197
5.4	Assessing classification performance	198
5.4.1	Accuracy – 0/1 loss	198
5.4.2	Sensitivity and specificity	198
5.4.3	The area under the ROC curve	199
5.4.4	Confusion matrices	201
5.5	Discriminative and generative classifiers	203
5.6	Summary	203
5.7	Exercises	203
	Further reading	205
6	Clustering	207
6.1	The general problem	207
6.2	K -means clustering	208
6.2.1	Choosing the number of clusters	210
6.2.2	Where K -means fails	212
6.2.3	Kernelised K -means	212
6.2.4	Summary	214
6.3	Mixture models	215
6.3.1	A generative process	216
6.3.2	Mixture model likelihood	217
6.3.3	The EM algorithm	219
6.3.3.1	Updating π_k	220
6.3.3.2	Updating μ_k	221
6.3.3.3	Updating Σ_k	222
6.3.3.4	Updating q_{nk}	223
6.3.3.5	Some intuition	224

6.3.4	Example	225
6.3.5	EM finds local optima	226
6.3.6	Choosing the number of components	228
6.3.7	Other forms of mixture components	230
6.3.8	MAP estimates with EM	232
6.3.9	Bayesian mixture models	233
6.4	Summary	234
6.5	Exercises	234
	Further reading	237
7	Principal Components Analysis and Latent Variable Models	239
7.1	The general problem	239
7.1.1	Variance as a proxy for interest	239
7.2	Principal components analysis	242
7.2.1	Choosing D	247
7.2.2	Limitations of PCA	247
7.3	Latent variable models	248
7.3.1	Mixture models as latent variable models	248
7.3.2	Summary	249
7.4	Variational Bayes	249
7.4.1	Choosing $Q(\theta)$	251
7.4.2	Optimising the bound	252
7.5	A probabilistic model for PCA	252
7.5.1	$Q_{\tau}(\tau)$	254
7.5.2	$Q_{\mathbf{x}_n}(\mathbf{x}_n)$	256
7.5.3	$Q_{\mathbf{w}_m}(\mathbf{w}_m)$	257
7.5.4	The required expectations	258
7.5.5	The algorithm	258
7.5.6	An example	260
7.6	Missing values	260
7.6.1	Missing values as latent variables	262
7.6.2	Predicting missing values	264
7.7	Non-real-valued data	264
7.7.1	Probit PPCA	264
7.7.2	Visualising parliamentary data	268
7.7.2.1	Aside – relationship to classification	272
7.8	Summary	273
7.9	Exercises	273
	Further reading	275
	Glossary	277
	Index	283