

Natural Networks for Pattern Recognition

CHRISTOPHER M. BISHOP

*Microsoft Research
Cambridge, UK*

OXFORD
UNIVERSITY PRESS

CONTENTS

1 Statistical Pattern Recognition	1
1.1 An example – character recognition	1
1.2 Classification and regression	5
1.3 Pre-processing and feature extraction	6
1.4 The curse of dimensionality	7
1.5 Polynomial curve fitting	9
1.6 Model complexity	14
1.7 Multivariate non-linear functions	15
1.8 Bayes' theorem	17
1.9 Decision boundaries	23
1.10 Minimising risk	27
Exercises	28
2 Probability Density Estimation	33
2.1 Parametric methods	34
2.2 Maximum likelihood	39
2.3 Bayesian inference	42
2.4 Sequential parameter estimation	46
2.5 Non-parametric methods	49
2.6 Mixture models	59
Exercises	73
3 Single-Layer Networks	77
3.1 Linear discriminant functions	77
3.2 Linear separability	85
3.3 Generalised linear discriminants	88
3.4 Least-squares techniques	89
3.5 The perceptron	98
3.6 Fisher's linear discriminant	105
Exercises	112
4 The Multi-layer Perceptron	116
4.1 Feed-forward network mappings	116
4.2 Threshold units	121
4.3 Sigmoidal units	126
4.4 Weight-space symmetries	133
4.5 Higher-order networks	133
4.6 Projection pursuit regression	135
4.7 Kolmogorov's theorem	137

4.8	Error back-propagation	140
4.9	The Jacobian matrix	148
4.10	The Hessian matrix	150
	Exercises	161
5	Radial Basis Functions	164
5.1	Exact interpolation	164
5.2	Radial basis function networks	167
5.3	Network training	170
5.4	Regularization theory	171
5.5	Noisy interpolation theory	176
5.6	Relation to kernel regression	177
5.7	Radial basis function networks for classification	179
5.8	Comparison with the multi-layer perceptron	182
5.9	Basis function optimization	183
5.10	Supervised training	190
	Exercises	191
6	Error Functions	194
6.1	Sum-of-squares error	195
6.2	Minkowski error	208
6.3	Input-dependent variance	211
6.4	Modelling conditional distributions	212
6.5	Estimating posterior probabilities	222
6.6	Sum-of-squares for classification	225
6.7	Cross-entropy for two classes	230
6.8	Multiple independent attributes	236
6.9	Cross-entropy for multiple classes	237
6.10	Entropy	240
6.11	General conditions for outputs to be probabilities	245
	Exercises	248
7	Parameter Optimization Algorithms	253
7.1	Error surfaces	254
7.2	Local quadratic approximation	257
7.3	Linear output units	259
7.4	Optimization in practice	260
7.5	Gradient descent	263
7.6	Line search	272
7.7	Conjugate gradients	274
7.8	Scaled conjugate gradients	282
7.9	Newton's method	285
7.10	Quasi-Newton methods	287
7.11	The Levenberg-Marquardt algorithm	290
	Exercises	292

8	Pre-processing and Feature Extraction	295
8.1	Pre-processing and post-processing	296
8.2	Input normalization and encoding	298
8.3	Missing data	301
8.4	Time series prediction	302
8.5	Feature selection	304
8.6	Principal component analysis	310
8.7	Invariances and prior knowledge	319
	Exercises	329
9	Learning and Generalization	332
9.1	Bias and variance	333
9.2	Regularization	338
9.3	Training with noise	346
9.4	Soft weight sharing	349
9.5	Growing and pruning algorithms	353
9.6	Committees of networks	364
9.7	Mixtures of experts	369
9.8	Model order selection	371
9.9	Vapnik-Chervonenkis dimension	377
	Exercises	380
10	Bayesian Techniques	385
10.1	Bayesian learning of network weights	387
10.2	Distribution of network outputs	398
10.3	Application to classification problems	403
10.4	The evidence framework for α and β	406
10.5	Integration over hyperparameters	415
10.6	Bayesian model comparison	418
10.7	Committees of networks	422
10.8	Practical implementation of Bayesian techniques	424
10.9	Monte Carlo methods	425
10.10	Minimum description length	429
	Exercises	433
A	Symmetric Matrices	440
B	Gaussian Integrals	444
C	Lagrange Multipliers	448
D	Calculus of Variations	451
E	Principal Components	454
	References	457
	Index	477