Contents

Preface vii

Mathematical notation xi

Contents xiii

1 Introduction 1
- 1.1 Example: Polynomial Curve Fitting . . . 4
- 1.2 Probability Theory . . . 12
  - 1.2.1 Probability densities . . . 17
  - 1.2.2 Expectations and covariances . . . 19
  - 1.2.3 Bayesian probabilities . . . 21
  - 1.2.4 The Gaussian distribution . . . 24
  - 1.2.5 Curve fitting re-visited . . . 28
  - 1.2.6 Bayesian curve fitting . . . 30
- 1.3 Model Selection . . . 32
- 1.4 The Curse of Dimensionality . . . 33
- 1.5 Decision Theory . . . 38
  - 1.5.1 Minimizing the misclassification rate . . . 39
  - 1.5.2 Minimizing the expected loss . . . 41
  - 1.5.3 The reject option . . . 42
  - 1.5.4 Inference and decision . . . 42
  - 1.5.5 Loss functions for regression . . . 46
- 1.6 Information Theory . . . 48
  - 1.6.1 Relative entropy and mutual information . . . 55

Exercises . . . 58

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

2 Probability Distributions **67**
2.1 Binary Variables . . . 68
2.1.1 The beta distribution . . . 71
2.2 Multinomial Variables . . . 74
2.2.1 The Dirichlet distribution . . . 76
2.3 The Gaussian Distribution . . . 78
2.3.1 Conditional Gaussian distributions . . . 85
2.3.2 Marginal Gaussian distributions . . . 88
2.3.3 Bayes' theorem for Gaussian variables . . . 90
2.3.4 Maximum likelihood for the Gaussian . . . 93
2.3.5 Sequential estimation . . . 94
2.3.6 Bayesian inference for the Gaussian . . . 97
2.3.7 Student's t-distribution . . . 102
2.3.8 Periodic variables . . . 105
2.3.9 Mixtures of Gaussians . . . 110
2.4 The Exponential Family . . . 113
2.4.1 Maximum likelihood and sufficient statistics . . . 116
2.4.2 Conjugate priors . . . 117
2.4.3 Noninformative priors . . . 117
2.5 Nonparametric Methods . . . 120
2.5.1 Kernel density estimators . . . 122
2.5.2 Nearest-neighbour methods . . . 124
Exercises . . . 127

3 Linear Models for Regression **137**
3.1 Linear Basis Function Models . . . 138
3.1.1 Maximum likelihood and least squares . . . 140
3.1.2 Geometry of least squares . . . 143
3.1.3 Sequential learning . . . 143
3.1.4 Regularized least squares . . . 144
3.1.5 Multiple outputs . . . 146
3.2 The Bias-Variance Decomposition . . . 147
3.3 Bayesian Linear Regression . . . 152
3.3.1 Parameter distribution . . . 152
3.3.2 Predictive distribution . . . 156
3.3.3 Equivalent kernel . . . 159
3.4 Bayesian Model Comparison . . . 161
3.5 The Evidence Approximation . . . 165
3.5.1 Evaluation of the evidence function . . . 166
3.5.2 Maximizing the evidence function . . . 168
3.5.3 Effective number of parameters . . . 170
3.6 Limitations of Fixed Basis Functions . . . 172
Exercises . . . 173

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

4 Linear Models for Classification **179**
4.1 Discriminant Functions . . . 181
4.1.1 Two classes . . . 181
4.1.2 Multiple classes . . . 182
4.1.3 Least squares for classification . . . 184
4.1.4 Fisher's linear discriminant . . . 186
4.1.5 Relation to least squares . . . 189
4.1.6 Fisher's discriminant for multiple classes . . . 191
4.1.7 The perceptron algorithm . . . 192
4.2 Probabilistic Generative Models . . . 196
4.2.1 Continuous inputs . . . 198
4.2.2 Maximum likelihood solution . . . 200
4.2.3 Discrete features . . . 202
4.2.4 Exponential family . . . 202
4.3 Probabilistic Discriminative Models . . . 203
4.3.1 Fixed basis functions . . . 204
4.3.2 Logistic regression . . . 205
4.3.3 Iterative reweighted least squares . . . 207
4.3.4 Multiclass logistic regression . . . 209
4.3.5 Probit regression . . . 210
4.3.6 Canonical link functions . . . 212
4.4 The Laplace Approximation . . . 213
4.4.1 Model comparison and BIC . . . 216
4.5 Bayesian Logistic Regression . . . 217
4.5.1 Laplace approximation . . . 217
4.5.2 Predictive distribution . . . 218
Exercises . . . 220

5 Neural Networks **225**
5.1 Feed-forward Network Functions . . . 227
5.1.1 Weight-space symmetries . . . 231
5.2 Network Training . . . 232
5.2.1 Parameter optimization . . . 236
5.2.2 Local quadratic approximation . . . 237
5.2.3 Use of gradient information . . . 239
5.2.4 Gradient descent optimization . . . 240
5.3 Error Backpropagation . . . 241
5.3.1 Evaluation of error-function derivatives . . . 242
5.3.2 A simple example . . . 245
5.3.3 Efficiency of backpropagation . . . 246
5.3.4 The Jacobian matrix . . . 247
5.4 The Hessian Matrix . . . 249
5.4.1 Diagonal approximation . . . 250
5.4.2 Outer product approximation . . . 251
5.4.3 Inverse Hessian . . . 252

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

5.4.4 Finite differences . . . 252
5.4.5 Exact evaluation of the Hessian . . . 253
5.4.6 Fast multiplication by the Hessian . . . 254
5.5 Regularization in Neural Networks . . . 256
5.5.1 Consistent Gaussian priors . . . 257
5.5.2 Early stopping . . . 259
5.5.3 Invariances . . . 261
5.5.4 Tangent propagation . . . 263
5.5.5 Training with transformed data . . . 265
5.5.6 Convolutional networks . . . 267
5.5.7 Soft weight sharing . . . 269
5.6 Mixture Density Networks . . . 272
5.7 Bayesian Neural Networks . . . 277
5.7.1 Posterior parameter distribution . . . 278
5.7.2 Hyperparameter optimization . . . 280
5.7.3 Bayesian neural networks for classification . . . 281
Exercises . . . 284

6 Kernel Methods 291
6.1 Dual Representations . . . 293
6.2 Constructing Kernels . . . 294
6.3 Radial Basis Function Networks . . . 299
6.3.1 Nadaraya-Watson model . . . 301
6.4 Gaussian Processes . . . 303
6.4.1 Linear regression revisited . . . 304
6.4.2 Gaussian processes for regression . . . 306
6.4.3 Learning the hyperparameters . . . 311
6.4.4 Automatic relevance determination . . . 312
6.4.5 Gaussian processes for classification . . . 313
6.4.6 Laplace approximation . . . 315
6.4.7 Connection to neural networks . . . 319
Exercises . . . 320

7 Sparse Kernel Machines 325
7.1 Maximum Margin Classifiers . . . 326
7.1.1 Overlapping class distributions . . . 331
7.1.2 Relation to logistic regression . . . 336
7.1.3 Multiclass SVMs . . . 338
7.1.4 SVMs for regression . . . 339
7.1.5 Computational learning theory . . . 344
7.2 Relevance Vector Machines . . . 345
7.2.1 RVM for regression . . . 345
7.2.2 Analysis of sparsity . . . 349
7.2.3 RVM for classification . . . 353
Exercises . . . 357

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

8 Graphical Models **359**
8.1 Bayesian Networks . . . 360
8.1.1 Example: Polynomial regression . . . 362
8.1.2 Generative models . . . 365
8.1.3 Discrete variables . . . 366
8.1.4 Linear-Gaussian models . . . 370
8.2 Conditional Independence . . . 372
8.2.1 Three example graphs . . . 373
8.2.2 D-separation . . . 378
8.3 Markov Random Fields . . . 383
8.3.1 Conditional independence properties . . . 383
8.3.2 Factorization properties . . . 384
8.3.3 Illustration: Image de-noising . . . 387
8.3.4 Relation to directed graphs . . . 390
8.4 Inference in Graphical Models . . . 393
8.4.1 Inference on a chain . . . 394
8.4.2 Trees . . . 398
8.4.3 Factor graphs . . . 399
8.4.4 The sum-product algorithm . . . 402
8.4.5 The max-sum algorithm . . . 411
8.4.6 Exact inference in general graphs . . . 416
8.4.7 Loopy belief propagation . . . 417
8.4.8 Learning the graph structure . . . 418
Exercises . . . 418

9 Mixture Models and EM **423**
9.1 K-means Clustering . . . 424
9.1.1 Image segmentation and compression . . . 428
9.2 Mixtures of Gaussians . . . 430
9.2.1 Maximum likelihood . . . 432
9.2.2 EM for Gaussian mixtures . . . 435
9.3 An Alternative View of EM . . . 439
9.3.1 Gaussian mixtures revisited . . . 441
9.3.2 Relation to K-means . . . 443
9.3.3 Mixtures of Bernoulli distributions . . . 444
9.3.4 EM for Bayesian linear regression . . . 448
9.4 The EM Algorithm in General . . . 450
Exercises . . . 455

10 Approximate Inference **461**
10.1 Variational Inference . . . 462
10.1.1 Factorized distributions . . . 464
10.1.2 Properties of factorized approximations . . . 466
10.1.3 Example: The univariate Gaussian . . . 470
10.1.4 Model comparison . . . 473
10.2 Illustration: Variational Mixture of Gaussians . . . 474

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

10.2.1 Variational distribution 475
10.2.2 Variational lower bound 481
10.2.3 Predictive density 482
10.2.4 Determining the number of components 483
10.2.5 Induced factorizations 485
10.3 Variational Linear Regression 486
10.3.1 Variational distribution 486
10.3.2 Predictive distribution 488
10.3.3 Lower bound 489
10.4 Exponential Family Distributions 490
10.4.1 Variational message passing 491
10.5 Local Variational Methods 493
10.6 Variational Logistic Regression 498
10.6.1 Variational posterior distribution 498
10.6.2 Optimizing the variational parameters 500
10.6.3 Inference of hyperparameters 502
10.7 Expectation Propagation 505
10.7.1 Example: The clutter problem 511
10.7.2 Expectation propagation on graphs 513
Exercises 517

11 Sampling Methods **523**
11.1 Basic Sampling Algorithms 526
11.1.1 Standard distributions 526
11.1.2 Rejection sampling 528
11.1.3 Adaptive rejection sampling 530
11.1.4 Importance sampling 532
11.1.5 Sampling-importance-resampling 534
11.1.6 Sampling and the EM algorithm 536
11.2 Markov Chain Monte Carlo 537
11.2.1 Markov chains 539
11.2.2 The Metropolis-Hastings algorithm 541
11.3 Gibbs Sampling 542
11.4 Slice Sampling 546
11.5 The Hybrid Monte Carlo Algorithm 548
11.5.1 Dynamical systems 548
11.5.2 Hybrid Monte Carlo 552
11.6 Estimating the Partition Function 554
Exercises 556

12 Continuous Latent Variables **559**
12.1 Principal Component Analysis 561
12.1.1 Maximum variance formulation 561
12.1.2 Minimum-error formulation 563
12.1.3 Applications of PCA 565
12.1.4 PCA for high-dimensional data 569

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

12.2 Probabilistic PCA . . . 570
12.2.1 Maximum likelihood PCA . . . 574
12.2.2 EM algorithm for PCA . . . 577
12.2.3 Bayesian PCA . . . 580
12.2.4 Factor analysis . . . 583
12.3 Kernel PCA . . . 586
12.4 Nonlinear Latent Variable Models . . . 591
12.4.1 Independent component analysis . . . 591
12.4.2 Autoassociative neural networks . . . 592
12.4.3 Modelling nonlinear manifolds . . . 595
Exercises . . . 599

13 Sequential Data **605**
13.1 Markov Models . . . 607
13.2 Hidden Markov Models . . . 610
13.2.1 Maximum likelihood for the HMM . . . 615
13.2.2 The forward-backward algorithm . . . 618
13.2.3 The sum-product algorithm for the HMM . . . 625
13.2.4 Scaling factors . . . 627
13.2.5 The Viterbi algorithm . . . 629
13.2.6 Extensions of the hidden Markov model . . . 631
13.3 Linear Dynamical Systems . . . 635
13.3.1 Inference in LDS . . . 638
13.3.2 Learning in LDS . . . 642
13.3.3 Extensions of LDS . . . 644
13.3.4 Particle filters . . . 645
Exercises . . . 646

14 Combining Models **653**
14.1 Bayesian Model Averaging . . . 654
14.2 Committees . . . 655
14.3 Boosting . . . 657
14.3.1 Minimizing exponential error . . . 659
14.3.2 Error functions for boosting . . . 661
14.4 Tree-based Models . . . 663
14.5 Conditional Mixture Models . . . 666
14.5.1 Mixtures of linear regression models . . . 667
14.5.2 Mixtures of logistic models . . . 670
14.5.3 Mixtures of experts . . . 672
Exercises . . . 674

Appendix A Data Sets **677**

Appendix B Probability Distributions **685**

Appendix C Properties of Matrices **695**

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

Appendix D Calculus of Variations **703**

Appendix E Lagrange Multipliers **707**

References **711**

© Christpher M. Bishop (2002–2006). Springer, 2006. First printing.
Further information available at http://research.microsoft.com/∼cmbishop/PRML

http://www.springer.com/978-0-387-31073-2

Pattern Recognition and Machine Learning
Bishop, C.M.
2006, XX, 740 p. 304 illus. in color., Hardcover
ISBN: 978-0-387-31073-2