

Contents

Series Foreword	xi
Preface	xiii
Symbols and Notation	xvii
1 Introduction	1
1.1 A Pictorial Introduction to Bayesian Modelling	3
1.2 Roadmap	5
2 Regression	7
2.1 Weight-space View	7
2.1.1 The Standard Linear Model	8
2.1.2 Projections of Inputs into Feature Space	11
2.2 Function-space View	13
2.3 Varying the Hyperparameters	19
2.4 Decision Theory for Regression	21
2.5 An Example Application	22
2.6 Smoothing, Weight Functions and Equivalent Kernels	24
* 2.7 Incorporating Explicit Basis Functions	27
2.7.1 Marginal Likelihood	29
2.8 History and Related Work	29
2.9 Exercises	30
3 Classification	33
3.1 Classification Problems	34
3.1.1 Decision Theory for Classification	35
3.2 Linear Models for Classification	37
3.3 Gaussian Process Classification	39
3.4 The Laplace Approximation for the Binary GP Classifier	41
3.4.1 Posterior	42
3.4.2 Predictions	44
3.4.3 Implementation	45
3.4.4 Marginal Likelihood	47
* 3.5 Multi-class Laplace Approximation	48
3.5.1 Implementation	51
3.6 Expectation Propagation	52
3.6.1 Predictions	56
3.6.2 Marginal Likelihood	57
3.6.3 Implementation	57
3.7 Experiments	60
3.7.1 A Toy Problem	60
3.7.2 One-dimensional Example	62
3.7.3 Binary Handwritten Digit Classification Example	63
3.7.4 10-class Handwritten Digit Classification Example	70
3.8 Discussion	72

*Sections marked by an asterisk contain advanced material that may be omitted on a first reading.

* 3.9	Appendix: Moment Derivations	74
3.10	Exercises	75
4	Covariance Functions	79
4.1	Preliminaries	79
* 4.1.1	Mean Square Continuity and Differentiability	81
4.2	Examples of Covariance Functions	81
4.2.1	Stationary Covariance Functions	82
4.2.2	Dot Product Covariance Functions	89
4.2.3	Other Non-stationary Covariance Functions	90
4.2.4	Making New Kernels from Old	94
4.3	Eigenfunction Analysis of Kernels	96
* 4.3.1	An Analytic Example	97
4.3.2	Numerical Approximation of Eigenfunctions	98
4.4	Kernels for Non-vectorial Inputs	99
4.4.1	String Kernels	100
4.4.2	Fisher Kernels	101
4.5	Exercises	102
5	Model Selection and Adaptation of Hyperparameters	105
5.1	The Model Selection Problem	106
5.2	Bayesian Model Selection	108
5.3	Cross-validation	111
5.4	Model Selection for GP Regression	112
5.4.1	Marginal Likelihood	112
5.4.2	Cross-validation	116
5.4.3	Examples and Discussion	118
5.5	Model Selection for GP Classification	124
* 5.5.1	Derivatives of the Marginal Likelihood for Laplace’s Approximation	125
* 5.5.2	Derivatives of the Marginal Likelihood for EP	127
5.5.3	Cross-validation	127
5.5.4	Example	128
5.6	Exercises	128
6	Relationships between GPs and Other Models	129
6.1	Reproducing Kernel Hilbert Spaces	129
6.2	Regularization	132
* 6.2.1	Regularization Defined by Differential Operators	133
6.2.2	Obtaining the Regularized Solution	135
6.2.3	The Relationship of the Regularization View to Gaussian Process Prediction	135
6.3	Spline Models	136
* 6.3.1	A 1-d Gaussian Process Spline Construction	138
* 6.4	Support Vector Machines	141
6.4.1	Support Vector Classification	141
6.4.2	Support Vector Regression	145
* 6.5	Least-squares Classification	146
6.5.1	Probabilistic Least-squares Classification	147

* 6.6	Relevance Vector Machines	149
6.7	Exercises	150
7	Theoretical Perspectives	151
7.1	The Equivalent Kernel	151
7.1.1	Some Specific Examples of Equivalent Kernels	153
* 7.2	Asymptotic Analysis	155
7.2.1	Consistency	155
7.2.2	Equivalence and Orthogonality	157
* 7.3	Average-case Learning Curves	159
* 7.4	PAC-Bayesian Analysis	161
7.4.1	The PAC Framework	162
7.4.2	PAC-Bayesian Analysis	163
7.4.3	PAC-Bayesian Analysis of GP Classification	164
7.5	Comparison with Other Supervised Learning Methods	165
* 7.6	Appendix: Learning Curve for the Ornstein-Uhlenbeck Process	168
7.7	Exercises	169
8	Approximation Methods for Large Datasets	171
8.1	Reduced-rank Approximations of the Gram Matrix	171
8.2	Greedy Approximation	174
8.3	Approximations for GPR with Fixed Hyperparameters	175
8.3.1	Subset of Regressors	175
8.3.2	The Nyström Method	177
8.3.3	Subset of Datapoints	177
8.3.4	Projected Process Approximation	178
8.3.5	Bayesian Committee Machine	180
8.3.6	Iterative Solution of Linear Systems	181
8.3.7	Comparison of Approximate GPR Methods	182
8.4	Approximations for GPC with Fixed Hyperparameters	185
* 8.5	Approximating the Marginal Likelihood and its Derivatives	185
* 8.6	Appendix: Equivalence of SR and GPR Using the Nyström Approximate Kernel	187
8.7	Exercises	187
9	Further Issues and Conclusions	189
9.1	Multiple Outputs	190
9.2	Noise Models with Dependencies	190
9.3	Non-Gaussian Likelihoods	191
9.4	Derivative Observations	191
9.5	Prediction with Uncertain Inputs	192
9.6	Mixtures of Gaussian Processes	192
9.7	Global Optimization	193
9.8	Evaluation of Integrals	193
9.9	Student’s t Process	194
9.10	Invariances	194
9.11	Latent Variable Models	196
9.12	Conclusions and Future Directions	196

Appendix A Mathematical Background	199
A.1 Joint, Marginal and Conditional Probability	199
A.2 Gaussian Identities	200
A.3 Matrix Identities	201
A.3.1 Matrix Derivatives	202
A.3.2 Matrix Norms	202
A.4 Cholesky Decomposition	202
A.5 Entropy and Kullback-Leibler Divergence	203
A.6 Limits	204
A.7 Measure and Integration	204
A.7.1 L_p Spaces	205
A.8 Fourier Transforms	205
A.9 Convexity	206
Appendix B Gaussian Markov Processes	207
B.1 Fourier Analysis	208
B.1.1 Sampling and Periodization	209
B.2 Continuous-time Gaussian Markov Processes	211
B.2.1 Continuous-time GMPs on \mathbb{R}	211
B.2.2 The Solution of the Corresponding SDE on the Circle	213
B.3 Discrete-time Gaussian Markov Processes	214
B.3.1 Discrete-time GMPs on \mathbb{Z}	214
B.3.2 The Solution of the Corresponding Difference Equation on \mathbb{P}_N	215
B.4 The Relationship Between Discrete-time and Sampled Continuous-time GMPs	217
B.5 Markov Processes in Higher Dimensions	218
Appendix C Datasets and Code	221
Bibliography	223
Author Index	239
Subject Index	245