Contents

Preface ................................................................. xvii

Part I Machine learning and kernel vector spaces .......... 1

1 Fundamentals of kernel-based machine learning .......... 3

1.1 Introduction ....................................................... 3

1.2 Feature representation and dimension reduction .......... 4

1.2.1 Feature representation in vector space ................. 6

1.2.2 Conventional similarity metric: Euclidean inner product 8

1.2.3 Feature dimension reduction ................................. 8

1.3 The learning subspace property (LSP) and “kernelization” of learning models ........................................... 9

1.3.1 The LSP ......................................................... 9

1.3.2 Kernelization of the optimization formulation for learning models ......................................................... 13

1.3.3 The LSP is necessary and sufficient for kernelization ... 14

1.4 Unsupervised learning for cluster discovery ............... 15

1.4.1 Characterization of similarity metrics ................. 15

1.4.2 The LSP and kernelization of K-means learning models ................................................................. 16

1.4.3 The LSP and kernelization of \( \ell_2 \) elastic nets ....... 18

1.5 Supervised learning for linear classifiers ................. 19

1.5.1 Learning and prediction phases ............................. 20

1.5.2 Learning models and linear system of equations ....... 21

1.5.3 Kernelized learning models for under-determined systems ......................................................... 23

1.5.4 The vital role of the \( \ell_2 \)-norm for the LSP .......... 24

1.5.5 The LSP condition of one-class SVM for outlier detection ......................................................... 25

1.6 Generalized inner products and kernel functions .......... 25

1.6.1 Mahalanobis inner products ................................ 26

1.6.2 Nonlinear inner product: Mercer kernel functions .... 27

1.6.3 Effective implementation of kernel methods .......... 30

1.7 Performance metrics .............................................. 31

1.7.1 Accuracy and error rate ..................................... 31

1.7.2 Sensitivity, specificity, and precision ..................... 32

1.7.3 The receiver operating characteristic (ROC) .......... 33
1.8 Highlights of chapters 35
1.9 Problems 38

2 Kernel-induced vector spaces 44
2.1 Introduction 44
2.2 Mercer kernels and kernel-induced similarity metrics 45
   2.2.1 Distance axioms in metric space 45
   2.2.2 Mercer kernels 46
   2.2.3 Construction of Mercer kernels 50
   2.2.4 Shift-invariant kernel functions 50
2.3 Training-data-independent intrinsic feature vectors 50
   2.3.1 Intrinsic spaces associated with kernel functions 52
   2.3.2 Intrinsic-space-based learning models 56
2.4 Training-data-dependent empirical feature vectors 60
   2.4.1 The LSP: from intrinsic space to empirical space 61
   2.4.2 Kernelized learning models 63
   2.4.3 Implementation cost comparison of two spaces 66
2.5 The kernel-trick for nonvectorial data analysis 67
   2.5.1 The LSP: from intrinsic space to empirical space 67
   2.5.2 The Mercer condition and kernel tricks 70
2.6 Summary 72
2.7 Problems 72

Part II Dimension-reduction: PCA/KPCA and feature selection 77

3 PCA and kernel PCA 79
3.1 Introduction 79
3.2 Why dimension reduction? 79
3.3 Subspace projection and PCA 81
   3.3.1 Optimality criteria for subspace projection 81
   3.3.2 PCA via spectral decomposition of the covariance matrix 82
   3.3.3 The optimal PCA solution: the mean-square-error criterion 83
   3.3.4 The optimal PCA solution: the maximum-entropy criterion 87
3.4 Numerical methods for computation of PCA 89
   3.4.1 Singular value decomposition of the data matrix 90
   3.4.2 Spectral decomposition of the scatter matrix 90
   3.4.3 Spectral decomposition of the kernel matrix 91
   3.4.4 Application studies of the subspace projection approach 94
3.5 Kernel principal component analysis (KPCA) 95
   3.5.1 The intrinsic-space approach to KPCA 95
   3.5.2 The kernelization of KPCA learning models 99
   3.5.3 PCA versus KPCA 105
   3.5.4 Center-adjusted versus unadjusted KPCAs 106
   3.5.5 Spectral vector space 110
3.6 Summary 113
3.7 Problems 113

4 Feature selection 118
4.1 Introduction 118
4.2 The filtering approach to feature selection 119
  4.2.1 Supervised filtering methods 120
  4.2.2 Feature-weighted linear classifiers 122
  4.2.3 Unsupervised filtering methods 124
  4.2.4 Consecutive search methods 124
4.3 The wrapper approach to feature selection 127
  4.3.1 Supervised wrapper methods 127
  4.3.2 Unsupervised wrapper methods 129
  4.3.3 The least absolute shrinkage and selection operator 130
4.4 Application studies of the feature selection approach 131
4.5 Summary 134
4.6 Problems 134

Part III Unsupervised learning models for cluster analysis 139

5 Unsupervised learning for cluster discovery 141
5.1 Introduction 141
5.2 The similarity metric and clustering strategy 141
5.3 $K$-means clustering models 144
  5.3.1 $K$-means clustering criterion 144
  5.3.2 The $K$-means algorithm 146
  5.3.3 Monotonic convergence of $K$-means 148
  5.3.4 The local optimum problem of $K$-means 151
  5.3.5 The evaluation criterion for multiple trials of $K$-means 152
  5.3.6 The optimal number of clusters 152
  5.3.7 Application examples 152
5.4 Expectation-maximization (EM) learning models 153
  5.4.1 EM clustering criterion 153
  5.4.2 The iterative EM algorithm for basic GMM 155
  5.4.3 Convergence of the EM algorithm with fixed $\sigma$ 156
  5.4.4 Annealing EM (AEM) 158
5.5 Self-organizing-map (SOM) learning models 159
  5.5.1 Input and output spaces in the SOM 161
  5.5.2 The SOM learning algorithm 162
  5.5.3 The evaluation criterion for multiple-trial SOM 165
  5.5.4 Applications of SOM learning models 166
5.6 Bi-clustering data analysis 169
  5.6.1 Coherence models for bi-clustering 170
  5.6.2 Applications of bi-clustering methods 171
5.7 Summary 173
5.8 Problems 174

6 Kernel methods for cluster analysis 178
  6.1 Introduction 178
  6.2 Kernel-based $K$-means learning models 179
    6.2.1 Kernel $K$-means in intrinsic space 180
    6.2.2 The $K$-means clustering criterion in terms of kernel matrix 181
  6.3 Kernel $K$-means for nonvectorial data analysis 183
    6.3.1 The similarity matrix for nonvectorial training datasets 184
    6.3.2 Clustering criteria for network segmentation 185
    6.3.3 The Mercer condition and convergence of kernel $K$-means 187
  6.4 $K$-means learning models in kernel-induced spectral space 190
    6.4.1 Discrepancy on optimal solution due to spectral truncation 191
    6.4.2 Computational complexities 193
  6.5 Kernelized $K$-means learning models 194
    6.5.1 Solution invariance of spectral-shift on the kernel matrix 194
    6.5.2 Kernelized $K$-means algorithms 195
    6.5.3 A recursive algorithm modified to exploit sparsity 197
  6.6 Kernel-induced SOM learning models 201
    6.6.1 SOM learning models in intrinsic or spectral space 201
    6.6.2 Kernelized SOM learning models 202
  6.7 Neighbor-joining hierarchical cluster analysis 204
    6.7.1 Divisive and agglomerative approaches 204
    6.7.2 An NJ method that is based on centroid update 206
    6.7.3 Kernelized hierarchical clustering algorithm 207
    6.7.4 Case studies: hierarchical clustering of microarray data 212
  6.8 Summary 213
  6.9 Problems 215

Part IV Kernel ridge regressors and variants 219

7 Kernel-based regression and regularization analysis 221
  7.1 Introduction 221
  7.2 Linear least-squares-error analysis 222
    7.2.1 Linear-least-MSE and least-squares-error (LSE) regressors 223
    7.2.2 Ridge regression analysis 225
  7.3 Kernel-based regression analysis 225
    7.3.1 LSE regression analysis: intrinsic space 227
    7.3.2 Kernel ridge regression analysis: intrinsic space 228
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3.3</td>
<td>The learning subspace property (LSP): from intrinsic to empirical space</td>
<td>228</td>
</tr>
<tr>
<td>7.3.4</td>
<td>KRR learning models: empirical space</td>
<td>228</td>
</tr>
<tr>
<td>7.3.5</td>
<td>Comparison of KRRs in intrinsic and empirical spaces</td>
<td>230</td>
</tr>
<tr>
<td>7.4</td>
<td>Radial basis function (RBF) networks for regression analysis</td>
<td>230</td>
</tr>
<tr>
<td>7.4.1</td>
<td>RBF approximation networks</td>
<td>230</td>
</tr>
<tr>
<td>7.4.2</td>
<td>The Nadaraya–Watson regression estimator (NWRE)</td>
<td>232</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Back-propagation neural networks</td>
<td>234</td>
</tr>
<tr>
<td>7.5</td>
<td>Multi-kernel regression analysis</td>
<td>240</td>
</tr>
<tr>
<td>7.6</td>
<td>Summary</td>
<td>244</td>
</tr>
<tr>
<td>7.7</td>
<td>Problems</td>
<td>244</td>
</tr>
</tbody>
</table>

8 Linear regression and discriminant analysis for supervised classification

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>248</td>
</tr>
<tr>
<td>8.2</td>
<td>Characterization of supervised learning models</td>
<td>249</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Binary and multiple classification</td>
<td>249</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Learning, evaluation, and prediction phases</td>
<td>250</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Off-line and inductive learning models</td>
<td>251</td>
</tr>
<tr>
<td>8.2.4</td>
<td>Linear and nonlinear learning models</td>
<td>252</td>
</tr>
<tr>
<td>8.2.5</td>
<td>Basic supervised learning strategies</td>
<td>252</td>
</tr>
<tr>
<td>8.3</td>
<td>Supervised learning models: over-determined formulation</td>
<td>253</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Direct derivation of LSE solution</td>
<td>254</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Fisher’s discriminant analysis (FDA)</td>
<td>258</td>
</tr>
<tr>
<td>8.4</td>
<td>Supervised learning models: under-determined formulation</td>
<td>263</td>
</tr>
<tr>
<td>8.5</td>
<td>A regularization method for robust classification</td>
<td>266</td>
</tr>
<tr>
<td>8.5.1</td>
<td>The ridge regression approach to linear classification</td>
<td>266</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Perturbational discriminant analysis (PDA): an extension of FDA</td>
<td>268</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Equivalence between RR and PDA</td>
<td>269</td>
</tr>
<tr>
<td>8.5.4</td>
<td>Regularization effects of the ridge parameter $\rho$</td>
<td>270</td>
</tr>
<tr>
<td>8.6</td>
<td>Kernelized learning models in empirical space: linear kernels</td>
<td>273</td>
</tr>
<tr>
<td>8.6.1</td>
<td>Kernelized learning models for under-determined systems</td>
<td>273</td>
</tr>
<tr>
<td>8.6.2</td>
<td>Kernelized formulation of KRR in empirical space</td>
<td>276</td>
</tr>
<tr>
<td>8.6.3</td>
<td>Comparison of formulations in original versus empirical spaces</td>
<td>277</td>
</tr>
<tr>
<td>8.7</td>
<td>Summary</td>
<td>278</td>
</tr>
<tr>
<td>8.8</td>
<td>Problems</td>
<td>278</td>
</tr>
</tbody>
</table>

9 Kernel ridge regression for supervised classification

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>282</td>
</tr>
<tr>
<td>9.2</td>
<td>Kernel-based discriminant analysis (KDA)</td>
<td>284</td>
</tr>
<tr>
<td>9.3</td>
<td>Kernel ridge regression (KRR) for supervised classification</td>
<td>287</td>
</tr>
<tr>
<td>9.3.1</td>
<td>KRR and LS-SVM models: the intrinsic-space approach</td>
<td>287</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Kernelized learning models: the empirical-space approach</td>
<td>288</td>
</tr>
</tbody>
</table>
9.3.3 A proof of equivalence of two formulations 289
9.3.4 Complexities of intrinsic and empirical models 290
9.4 Perturbational discriminant analysis (PDA) 290
9.5 Robustness and the regression ratio in spectral space 292
  9.5.1 The decision vector of KDA in spectral space 293
  9.5.2 Resilience of the decision components of KDA classifiers 293
  9.5.3 Component magnitude and component resilience 298
  9.5.4 Regression ratio: KDA versus KRR 299
9.6 Application studies: KDA versus KRR 300
  9.6.1 Experiments on UCI data 300
  9.6.2 Experiments on microarray cancer diagnosis 301
  9.6.3 Experiments on subcellular localization 302
9.7 Trimming detrimental (anti-support) vectors in KRR learning models 303
  9.7.1 A pruned-KRR learning model: pruned PDA (PPDA) 304
  9.7.2 Case study: ECG arrhythmia detection 306
9.8 Multi-class and multi-label supervised classification 307
  9.8.1 Multi-class supervised classification 307
  9.8.2 Multi-label classification 310
9.9 Supervised subspace projection methods 313
  9.9.1 Successively optimized discriminant analysis (SODA) 313
  9.9.2 Trace-norm optimization for subspace projection 318
  9.9.3 Discriminant component analysis (DCA) 325
  9.9.4 Comparisons between PCA, DCA, PC-DCA, and SODA 331
  9.9.5 Kernelized DCA and SODA learning models 333
9.10 Summary 335
9.11 Problems 336

Part V Support vector machines and variants 341

10 Support vector machines 343
  10.1 Introduction 343
  10.2 Linear support vector machines 344
    10.2.1 The optimization formulation in original vector space 345
    10.2.2 The Wolfe dual optimizer in empirical space 345
    10.2.3 The Karush–Kuhn–Tucker (KKT) condition 348
    10.2.4 Support vectors 349
    10.2.5 Comparison between separation margins of LSE and SVM 351
  10.3 SVM with fuzzy separation: roles of slack variables 353
    10.3.1 Optimization in original space 354
    10.3.2 The learning subspace property and optimization in empirical space 354
    10.3.3 Characterization of support vectors and WEC analysis 356
### 10.4 Kernel-induced support vector machines
- **10.4.1 Primal optimizer in intrinsic space**
- **10.4.2 Dual optimizer in empirical space**
- **10.4.3 Multi-class SVM learning models**
- **10.4.4 SVM learning softwares**

### 10.5 Application case studies
- **10.5.1 SVM for cancer data analysis**
- **10.5.2 Prediction performances w.r.t. size of training datasets**
- **10.5.3 KRR versus SVM: application studies**

### 10.6 Empirical-space SVM for trimming of support vectors
- **10.6.1 $\ell_1$-Norm SVM in empirical space**
- **10.6.2 $\ell_2$-Norm SVM in empirical space**
- **10.6.3 Empirical learning models for vectorial and nonvectorial data analysis**
- **10.6.4 Wrapper methods for empirical learning models**
- **10.6.5 Fusion of filtering and wrapper methods**

### 10.7 Summary

### 10.8 Problems

### 11 Support vector learning models for outlier detection
- **11.1 Introduction**
- **11.2 Support vector regression (SVR)**
- **11.3 Hyperplane-based one-class SVM learning models**
  - **11.3.1 Hyperplane-based $\nu$-SV classifiers**
  - **11.3.2 Hyperplane-based one-class SVM**
- **11.4 Hypersphere-based one-class SVM**
- **11.5 Support vector clustering**
- **11.6 Summary**
- **11.7 Problems**

### 12 Ridge-SVM learning models
- **12.1 Introduction**
- **12.2 Roles of $\rho$ and $C$ on WECs of KRR and SVM**
  - **12.2.1 Roles of $\rho$ and $C$**
  - **12.2.2 WECs of KDA, KRR, and SVM**
- **12.3 Ridge-SVM learning models**
  - **12.3.1 Ridge-SVM: a unifying supervised learning model**
  - **12.3.2 Important special cases of Ridge-SVM models**
  - **12.3.3 Subset selection: KKT and the termination condition**
- **12.4 Impacts of design parameters on the WEC of Ridge-SVM**
  - **12.4.1 Transition ramps and the number of support vectors**
  - **12.4.2 Effects of $\rho$ and $C_{\text{min}}$ on the transition ramp**
  - **12.4.3 The number of support vectors w.r.t. $C_{\text{min}}$**
12.5 Prediction accuracy versus training time 408
12.5.1 The tuning of the parameter $C$ 409
12.5.2 The tuning of the parameter $C_{\text{min}}$ 409
12.5.3 The tuning of the parameter $\rho$ 411
12.6 Application case studies 412
12.6.1 Experiments on UCI data 412
12.6.2 Experiments on microarray cancer diagnosis 413
12.6.3 Experiments on subcellular localization 414
12.6.4 Experiments on the ischemic stroke dataset 415
12.7 Summary 416
12.8 Problems 417

Part VI Kernel methods for green machine learning technologies 419

13 Efficient kernel methods for learning and classification 421
13.1 Introduction 421
13.2 System design considerations 423
13.2.1 Green processing technologies for local or client computing 423
13.2.2 Cloud computing platforms 423
13.2.3 Local versus centralized processing 424
13.3 Selection of cost-effective kernel functions 424
13.3.1 The intrinsic degree $J$ 426
13.3.2 Truncated-RBF (TRBF) kernels 428
13.4 Classification complexities: empirical and intrinsic degrees 430
13.4.1 The discriminant function in the empirical representation 432
13.4.2 The discriminant function in the intrinsic representation 433
13.4.3 Tensor representation of discriminant functions 436
13.4.4 Complexity comparison of RBF and TRBF classifiers 438
13.4.5 Case study: ECG arrhythmia detection 438
13.5 Learning complexities: empirical and intrinsic degrees 439
13.5.1 Learning complexities for KRR and SVM 439
13.5.2 A scatter-matrix-based KRR algorithm 440
13.5.3 KRR learning complexity: RBF versus TRBF kernels 440
13.5.4 A learning and classification algorithms for big data size $N$ 440
13.5.5 Case study: ECG arrhythmia detection 442
13.6 The tradeoff between complexity and prediction performance 444
13.6.1 Comparison of prediction accuracies 444
13.6.2 Prediction–complexity tradeoff analysis 446
13.7 Time-adaptive updating algorithms for KRR learning models 447
13.7.1 Time-adaptive recursive KRR algorithms 448
13.7.2 The intrinsic-space recursive KRR algorithm 449
13.7.3 A time-adaptive KRR algorithm with a forgetting factor 452
14 Statistical regression analysis and errors-in-variables models
14.1 Introduction 459
14.2 Statistical regression analysis 460
14.2.1 The minimum mean-square-error (MMSE) estimator/regressor 461
14.2.2 Linear regression analysis 462
14.3 Kernel ridge regression (KRR) 463
14.3.1 Orthonormal basis functions: single-variate cases 463
14.3.2 Orthonormal basis functions: multivariate cases 466
14.4 The perturbation-regulated regressor (PRR) for errors-in-variables models 467
14.4.1 MMSE solution for errors-in-variables models 468
14.4.2 Linear perturbation-regulated regressors 470
14.4.3 Kernel-based perturbation-regulated regressors 471
14.5 The kernel-based perturbation-regulated regressor (PRR): Gaussian cases 472
14.5.1 Orthonormal basis functions: single-variate cases 472
14.5.2 Single-variate Hermite estimators 473
14.5.3 Error–order tradeoff 475
14.5.4 Simulation results 477
14.5.5 Multivariate PRR estimators: Gaussian distribution 480
14.6 Two-projection theorems 482
14.6.1 The two-projection theorem: general case 483
14.6.2 The two-projection theorem: polynomial case 485
14.6.3 Two-projection for the PRR 486
14.6.4 Error analysis 486
14.7 Summary 487
14.8 Problems 488

15 Kernel methods for estimation, prediction, and system identification
15.1 Introduction 494
15.2 Kernel regressors for deterministic generation models 495
15.3 Kernel regressors for statistical generation models 500
15.3.1 The prior model and training data set 500
15.3.2 The Gauss–Markov theorem for statistical models 501
15.3.3 KRR regressors in empirical space 507
15.3.4 KRR regressors with Gaussian distribution 509
15.4 Kernel regressors for errors-in-variables (EiV) models 510
15.4.1 The Gauss–Markov theorem for EiV learning models 511
15.4.2 EiV regressors in empirical space 515
15.4.3 EiV regressors with Gaussian distribution 517
15.4.4 Finite-order EiV regressors 518
15.5 Recursive KRR learning algorithms 521
  15.5.1 The recursive KRR algorithm in intrinsic space 522
  15.5.2 The recursive KRR algorithm in empirical space 524
  15.5.3 The recursive KRR algorithm in intrinsic space with a forgetting factor 525
  15.5.4 The recursive KRR algorithm in empirical space with a forgetting factor and a finite window 527
15.6 Recursive EiV learning algorithms 529
  15.6.1 Recursive EiV learning models in intrinsic space 529
  15.6.2 The recursive EiV algorithm in empirical space 530
15.7 Summary 531
15.8 Problems 531

Part VIII Appendices 537

Appendix A Validation and testing of learning models 539
  A.1 Cross-validation techniques 539
  A.2 Hypothesis testing and significance testing 541
    A.2.1 Hypothesis testing based on the likelihood ratio 542
    A.2.2 Significance testing from the distribution of the null hypothesis 545
  A.3 Problems 547

Appendix B $k$NN, PNN, and Bayes classifiers 549
  B.1 Bayes classifiers 550
    B.1.1 The GMM-based-classifier 551
    B.1.2 The basic Bayes classifier 552
  B.2 Classifiers with no prior learning process 554
    B.2.1 $k$ nearest neighbors ($k$NN) 554
    B.2.2 Probabilistic neural networks (PNN) 555
    B.2.3 The log-likelihood classifier (LLC) 557
  B.3 Problems 559

References 561
Index 578