Learning with Kernels

Support Vector Machines, Regularization, Optimization, and Beyond

Bernhard Schölkopf
Alexander J. Smola

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Series Foreword xiii

Preface xv

1 A Tutorial Introduction 1
1.1 Data Representation and Similarity 1
1.2 A Simple Pattern Recognition Algorithm 4
1.3 Some Insights From Statistical Learning Theory 6
1.4 Hyperplane Classifiers 11
1.5 Support Vector Classification 15
1.6 Support Vector Regression 17
1.7 Kernel Principal Component Analysis 19
1.8 Empirical Results and Implementations 21

I CONCEPTS AND TOOLS 23

2 Kernels 25
2.1 Product Features 26
2.2 The Representation of Similarities in Linear Spaces 29
2.3 Examples and Properties of Kernels 45
2.4 The Representation of Dissimilarities in Linear Spaces 48
2.5 Summary 55
2.6 Problems 55

3 Risk and Loss Functions 61
3.1 Loss Functions 62
3.2 Test Error and Expected Risk 65
3.3 A Statistical Perspective 68
3.4 Robust Estimators 75
3.5 Summary 83
3.6 Problems 84

4 Regularization 87
4.1 The Regularized Risk Functional 88

4.2 The Representer Theorem 89
4.3 Regularization Operators 92
4.4 Translation Invariant Kernels 96
4.5 Translation Invariant Kernels in Higher Dimensions 105
4.6 Dot Product Kernels 110
4.7 Multi-Output Regularization 113
4.8 Semiparametric Regularization 115
4.9 Coefficient Based Regularization 118
4.10 Summary 121
4.11 Problems 122

5 Elements of Statistical Learning Theory 125
5.1 Introduction 125
5.2 The Law of Large Numbers 128
5.3 When Does Learning Work: the Question of Consistency 131
5.4 Uniform Convergence and Consistency 131
5.5 How to Derive a VC Bound 134
5.6 A Model Selection Example 144
5.7 Summary 146
5.8 Problems 146

6 Optimization 149
6.1 Convex Optimization 150
6.2 Unconstrained Problems 154
6.3 Constrained Problems 165
6.4 Interior Point Methods 175
6.5 Maximum Search Problems 179
6.6 Summary 183
6.7 Problems 184

II SUPPORT VECTOR MACHINES 187

7 Pattern Recognition 189
7.1 Separating Hyperplanes 189
7.2 The Role of the Margin 192
7.3 Optimal Margin Hyperplanes 196
7.4 Nonlinear Support Vector Classifiers 200
7.5 Soft Margin Hyperplanes 204
7.6 Multi-Class Classification 211
7.7 Variations on a Theme 214
7.8 Experiments 215
7.9 Summary 222
7.10 Problems 222

8 Single-Class Problems: Quantile Estimation and Novelty Detection 227
8.1 Introduction . . . 228
8.2 A Distribution's Support and Quantiles . . . 229
8.3 Algorithms . . . 230
8.4 Optimization . . . 234
8.5 Theory . . . 236
8.6 Discussion . . . 241
8.7 Experiments . . . 243
8.8 Summary . . . 247
8.9 Problems . . . 248

9 Regression Estimation 251
9.1 Linear Regression with Insensitive Loss Function . . . 251
9.2 Dual Problems . . . 254
9.3 ν-SV Regression . . . 260
9.4 Convex Combinations and ℓ_1-Norms . . . 266
9.5 Parametric Insensitivity Models . . . 269
9.6 Applications . . . 272
9.7 Summary . . . 273
9.8 Problems . . . 274

10 Implementation 279
10.1 Tricks of the Trade . . . 281
10.2 Sparse Greedy Matrix Approximation . . . 288
10.3 Interior Point Algorithms . . . 295
10.4 Subset Selection Methods . . . 300
10.5 Sequential Minimal Optimization . . . 305
10.6 Iterative Methods . . . 312
10.7 Summary . . . 327
10.8 Problems . . . 329

11 Incorporating Invariances 333
11.1 Prior Knowledge . . . 333
11.2 Transformation Invariance . . . 335
11.3 The Virtual SV Method . . . 337
11.4 Constructing Invariance Kernels . . . 343
11.5 The Jittered SV Method . . . 354
11.6 Summary . . . 356
11.7 Problems . . . 357

12 Learning Theory Revisited 359
12.1 Concentration of Measure Inequalities . . . 360
12.2 Leave-One-Out Estimates . . . 366
12.3 PAC-Bayesian Bounds . . . 381
12.4 Operator-Theoretic Methods in Learning Theory . . . 391

12.5 Summary 403
12.6 Problems 404

III KERNEL METHODS **405**

13 Designing Kernels **407**
13.1 Tricks for Constructing Kernels 408
13.2 String Kernels 412
13.3 Locality-Improved Kernels 414
13.4 Natural Kernels 418
13.5 Summary 423
13.6 Problems 423

14 Kernel Feature Extraction **427**
14.1 Introduction 427
14.2 Kernel PCA 429
14.3 Kernel PCA Experiments 437
14.4 A Framework for Feature Extraction 442
14.5 Algorithms for Sparse KFA 447
14.6 KFA Experiments 450
14.7 Summary 451
14.8 Problems 452

15 Kernel Fisher Discriminant **457**
15.1 Introduction 457
15.2 Fisher's Discriminant in Feature Space 458
15.3 Efficient Training of Kernel Fisher Discriminants 460
15.4 Probabilistic Outputs 464
15.5 Experiments 466
15.6 Summary 467
15.7 Problems 468

16 Bayesian Kernel Methods **469**
16.1 Bayesics 470
16.2 Inference Methods 475
16.3 Gaussian Processes 480
16.4 Implementation of Gaussian Processes 488
16.5 Laplacian Processes 499
16.6 Relevance Vector Machines 506
16.7 Summary 511
16.8 Problems 513

17 Regularized Principal Manifolds **517**
17.1 A Coding Framework 518

17.2 A Regularized Quantization Functional 522
17.3 An Algorithm for Minimizing $R_{\mathrm{reg}}[f]$ 526
17.4 Connections to Other Algorithms 529
17.5 Uniform Convergence Bounds 533
17.6 Experiments 537
17.7 Summary 539
17.8 Problems 540

18 **Pre-Images and Reduced Set Methods** **543**
18.1 The Pre-Image Problem 544
18.2 Finding Approximate Pre-Images 547
18.3 Reduced Set Methods 552
18.4 Reduced Set Selection Methods 554
18.5 Reduced Set Construction Methods 561
18.6 Sequential Evaluation of Reduced Set Expansions 564
18.7 Summary 566
18.8 Problems 567

A **Addenda** **569**
A.1 Data Sets 569
A.2 Proofs 572

B **Mathematical Prerequisites** **575**
B.1 Probability 575
B.2 Linear Algebra 580
B.3 Functional Analysis 586

References **591**

Index **617**

Notation and Symbols **625**