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

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>m</td>
</tr>
</tbody>
</table>

## Preface

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
</tr>
</tbody>
</table>

## A Tutorial Introduction

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Data Representation and Similarity.</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>A Simple Pattern Recognition Algorithm.</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Some Insights From Statistical Learning Theory.</td>
<td>6</td>
</tr>
<tr>
<td>1.4</td>
<td>Hyperplane Classifiers.</td>
<td>11</td>
</tr>
<tr>
<td>1.5</td>
<td>Support Vector Classification.</td>
<td>15</td>
</tr>
<tr>
<td>1.6</td>
<td>Support Vector Regression.</td>
<td>17</td>
</tr>
<tr>
<td>1.7</td>
<td>Kernel Principal Component Analysis.</td>
<td>19</td>
</tr>
<tr>
<td>1.8</td>
<td>Empirical Results and Implementations.</td>
<td>21</td>
</tr>
</tbody>
</table>

## Concepts and Tools

### Kernels

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Product Features.</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>The Representation of Similarities in Linear Spaces.</td>
<td>29</td>
</tr>
<tr>
<td>2.3</td>
<td>Examples and Properties of Kernels.</td>
<td>45</td>
</tr>
<tr>
<td>2.4</td>
<td>The Representation of Dissimilarities in Linear Spaces.</td>
<td>48</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary.</td>
<td>55</td>
</tr>
<tr>
<td>2.6</td>
<td>Problems.</td>
<td>55</td>
</tr>
</tbody>
</table>

### Risk and Loss Functions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Loss Functions.</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>Test Error and Expected Risk.</td>
<td>65</td>
</tr>
<tr>
<td>3.3</td>
<td>A Statistical Perspective.</td>
<td>68</td>
</tr>
<tr>
<td>3.4</td>
<td>Robust Estimators.</td>
<td>75</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary.</td>
<td>83</td>
</tr>
<tr>
<td>3.6</td>
<td>Problems.</td>
<td>84</td>
</tr>
</tbody>
</table>

### Regularization

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The Regularized Risk Functional</td>
<td>87</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.2</td>
<td>The Representer Theorem</td>
<td>89</td>
</tr>
<tr>
<td>4.3</td>
<td>Regularization Operators</td>
<td>92</td>
</tr>
<tr>
<td>4.4</td>
<td>Translation Invariant Kernels</td>
<td>96</td>
</tr>
<tr>
<td>4.5</td>
<td>Translation Invariant Kernels in Higher Dimensions</td>
<td>105</td>
</tr>
<tr>
<td>4.6</td>
<td>Dot Product Kernels</td>
<td>110</td>
</tr>
<tr>
<td>4.7</td>
<td>Multi-Output Regularization</td>
<td>113</td>
</tr>
<tr>
<td>4.8</td>
<td>Semiparametric Regularization</td>
<td>115</td>
</tr>
<tr>
<td>4.9</td>
<td>Coefficient Based Regularization</td>
<td>118</td>
</tr>
<tr>
<td>4.10</td>
<td>Summary</td>
<td>121</td>
</tr>
<tr>
<td>4.11</td>
<td>Problems</td>
<td>122</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>125</td>
</tr>
<tr>
<td>5.2</td>
<td>The Law of Large Numbers</td>
<td>128</td>
</tr>
<tr>
<td>5.3</td>
<td>When Does Learning Work: the Question of Consistency</td>
<td>131</td>
</tr>
<tr>
<td>5.4</td>
<td>Uniform Convergence and Consistency</td>
<td>131</td>
</tr>
<tr>
<td>5.5</td>
<td>How to Derive a VC Bound</td>
<td>134</td>
</tr>
<tr>
<td>5.6</td>
<td>A Model Selection Example</td>
<td>144</td>
</tr>
<tr>
<td>5.7</td>
<td>Summary</td>
<td>146</td>
</tr>
<tr>
<td>5.8</td>
<td>Problems</td>
<td>146</td>
</tr>
<tr>
<td>6.1</td>
<td>Convex Optimization</td>
<td>150</td>
</tr>
<tr>
<td>6.2</td>
<td>Unconstrained Problems</td>
<td>154</td>
</tr>
<tr>
<td>6.3</td>
<td>Constrained Problems</td>
<td>165</td>
</tr>
<tr>
<td>6.4</td>
<td>Interior Point Methods</td>
<td>175</td>
</tr>
<tr>
<td>6.5</td>
<td>Maximum Search Problems</td>
<td>179</td>
</tr>
<tr>
<td>6.6</td>
<td>Summary</td>
<td>183</td>
</tr>
<tr>
<td>6.7</td>
<td>Problems</td>
<td>184</td>
</tr>
<tr>
<td>7.1</td>
<td>Separating Hyperplanes</td>
<td>189</td>
</tr>
<tr>
<td>7.2</td>
<td>The Role of the Margin</td>
<td>192</td>
</tr>
<tr>
<td>7.3</td>
<td>Optimal Margin Hyperplanes</td>
<td>196</td>
</tr>
<tr>
<td>7.4</td>
<td>Nonlinear Support Vector Classifiers</td>
<td>200</td>
</tr>
<tr>
<td>7.5</td>
<td>Soft Margin Hyperplanes</td>
<td>204</td>
</tr>
<tr>
<td>7.6</td>
<td>Multi-Class Classification</td>
<td>211</td>
</tr>
<tr>
<td>7.7</td>
<td>Variations on a Theme</td>
<td>214</td>
</tr>
<tr>
<td>7.8</td>
<td>Experiments</td>
<td>215</td>
</tr>
<tr>
<td>7.9</td>
<td>Summary</td>
<td>222</td>
</tr>
<tr>
<td>7.10</td>
<td>Problems</td>
<td>222</td>
</tr>
</tbody>
</table>
Contents

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 I/-SV Regression ....................................... 260
  9.4 Convex Combinations and L1-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
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.2 A Regularized Quantization Functional</td>
<td>522</td>
</tr>
<tr>
<td>17.3 An Algorithm for Minimizing $R_{\text{reg}}[.]$</td>
<td>526</td>
</tr>
<tr>
<td>17.4 Connections to Other Algorithms</td>
<td>529</td>
</tr>
<tr>
<td>17.5 Uniform Convergence Bounds</td>
<td>533</td>
</tr>
<tr>
<td>17.6 Experiments</td>
<td>537</td>
</tr>
<tr>
<td>17.7 Summary</td>
<td>539</td>
</tr>
<tr>
<td>17.8 Problems</td>
<td>540</td>
</tr>
<tr>
<td>18 Pre-Images and Reduced Set Methods</td>
<td>543</td>
</tr>
<tr>
<td>18.1 The Pre-Image Problem</td>
<td>544</td>
</tr>
<tr>
<td>18.2 Finding Approximate Pre-Images</td>
<td>547</td>
</tr>
<tr>
<td>18.3 Reduced Set Methods</td>
<td>552</td>
</tr>
<tr>
<td>18.4 Reduced Set Selection Methods</td>
<td>554</td>
</tr>
<tr>
<td>18.5 Reduced Set Construction Methods</td>
<td>561</td>
</tr>
<tr>
<td>18.6 Sequential Evaluation of Reduced Set Expansions</td>
<td>564</td>
</tr>
<tr>
<td>18.7 Summary</td>
<td>566</td>
</tr>
<tr>
<td>18.8 Problems</td>
<td>567</td>
</tr>
<tr>
<td>A Addenda</td>
<td>569</td>
</tr>
<tr>
<td>A.I Data Sets</td>
<td>569</td>
</tr>
<tr>
<td>A.2 Proofs</td>
<td>572</td>
</tr>
<tr>
<td>B Mathematical Prerequisites</td>
<td>575</td>
</tr>
<tr>
<td>B.I Probability</td>
<td>575</td>
</tr>
<tr>
<td>B.2 Linear Algebra</td>
<td>580</td>
</tr>
<tr>
<td>B.3 Functional Analysis</td>
<td>586</td>
</tr>
<tr>
<td>References</td>
<td>591</td>
</tr>
<tr>
<td>Index</td>
<td>617</td>
</tr>
<tr>
<td>Notation and Symbols</td>
<td>625</td>
</tr>
</tbody>
</table>