Contents

1 Introduction 1
1.1 An Overview of Machine Learning 1
1.2 Challenges in Machine Learning 3
1.2.1 Solving Large-Scale SVMs 4
1.2.2 Feature Reduction with Support Vector Machines 5
1.2.3 Graph-Based Semi-supervised Learning Algorithms 6
1.2.4 Unsupervised Learning Based on Principle of Redundancy Reduction 7

2 Support Vector Machines in Classification and Regression – An Introduction 11
2.1 Basics of Learning from Data 12
2.2 Support Vector Machines in Classification and Regression 21
2.2.1 Linear Maximal Margin Classifier for Linearly Separable Data 21
2.2.2 Linear Soft Margin Classifier for Overlapping Classes 32
2.2.3 The Nonlinear SVMs Classifier 36
2.2.4 Regression by Support Vector Machines 48
2.3 Implementation Issues 57

3 Iterative Single Data Algorithm for Kernel Machines from Huge Data Sets: Theory and Performance 61
3.1 Introduction 61
3.2 Iterative Single Data Algorithm for Positive Definite Kernels without Bias Term b 63
3.2.1 Kernel AdaTron in Classification 64
3.2.2 SMO without Bias Term b in Classification 65
3.2.3 Kernel AdaTron in Regression 66
3.2.4 SMO without Bias Term b in Regression 67
3.2.5 The Coordinate Ascent Based Learning for Nonlinear Classification and Regression Tasks 68

3.2.6 Discussion on ISDA Without a Bias Term b 73
3.3 Iterative Single Data Algorithm with an Explicit Bias Term b . 73
3.3.1 Iterative Single Data Algorithm for SVMs Classification with a Bias Term b 74
3.4 Performance of the Iterative Single Data Algorithm and Comparisons 80
3.5 Implementation Issues 83
3.5.1 Working-set Selection and Shrinking of ISDA for Classification 83
3.5.2 Computation of the Kernel Matrix and Caching of ISDA for Classification 89
3.5.3 Implementation Details of ISDA for Regression 92
3.6 Conclusions 94

4 Feature Reduction with Support Vector Machines and Application in DNA Microarray Analysis 97
4.1 Introduction 97
4.2 Basics of Microarray Technology 99
4.3 Some Prior Work 101
4.3.1 Recursive Feature Elimination with Support Vector Machines 101
4.3.2 Selection Bias and How to Avoid It 102
4.4 Influence of the Penalty Parameter C in RFE-SVMs 103
4.5 Gene Selection for the Colon Cancer and the Lymphoma Data Sets 104
4.5.1 Results for Various C Parameters 104
4.5.2 Simulation Results with Different Preprocessing Procedures 107
4.6 Comparison between RFE-SVMs and the Nearest Shrunken Centroid Method 112
4.6.1 Basic Concept of Nearest Shrunken Centroid Method .. 112
4.6.2 Results on the Colon Cancer Data Set and the Lymphoma Data Set 115
4.7 Comparison of Genes' Ranking with Different Algorithms 120
4.8 Conclusions 122

5 Semi-supervised Learning and Applications 125
5.1 Introduction 125
5.2 Gaussian Random Fields Model and Consistency Method 127
5.2.1 Gaussian Random Fields Model 127
5.2.2 Global Consistency Model 130
5.2.3 Random Walks on Graph 133
5.3 An Investigation of the Effect of Unbalanced labeled Data on CM and GRFM Algorithms 136
5.3.1 Background and Test Settings 136

5.3.2 Results on the *Rec* Data Set 139
5.3.3 Possible Theoretical Explanations on the Effect of Unbalanced Labeled Data 139
5.4 Classifier Output Normalization: A Novel Decision Rule for Semi-supervised Learning Algorithm 142
5.5 Performance Comparison of Semi-supervised Learning Algorithms 145
5.5.1 Low Density Separation: Integration of Graph-Based Distances and ∇TSVM 146
5.5.2 Combining Graph-Based Distance with Manifold Approaches 149
5.5.3 Test Data Sets 150
5.5.4 Performance Comparison Between the LDS and the Manifold Approaches 152
5.5.5 Normalizatioin Steps and the Effect of σ 154
5.6 Implementation of the Manifold Approaches 154
5.6.1 Variants of the Manifold Approaches Implemented in the Software Package SemiL 155
5.6.2 Implementation Details of SemiL 157
5.6.3 Conjugate Gradient Method with Box Constraints 162
5.6.4 Simulation Results on the *MNIST* Data Set 166
5.7 An Overview of Text Classification 167
5.8 Conclusions 171

6 Unsupervised Learning by Principal and Independent Component Analysis 175
6.1 Principal Component Analysis 180
6.2 Independent Component Analysis 197
6.3 Concluding Remarks 208

A Support Vector Machines 209
A.1 L2 Soft Margin Classifier 210
A.2 L2 Soft Regressor 211
A.3 Geometry and the Margin 213

B Matlab Code for ISDA Classification 217

C Matlab Code for ISDA Regression 223

D Matlab Code for Conjugate Gradient Method with Box Constraints 229

E Uncorrelatedness and Independence 233

F Independent Component Analysis by Empirical Estimation of Score Functions i.e., Probability Density Functions 237

G SemiL User Guide 241
G.1 Installation 241
G.2 Input Data Format 243
G.2.1 Raw Data Format: 243
G.3 Getting Started 244
G.3.1 Design Stage 245

References 247

Index 257