Contents

Preface xiii

I MACHINE LEARNING FROM WEAK SUPERVISION

1 Introduction 3
   1.1 Machine Learning 3
      1.1.1 Supervised Learning 3
      1.1.2 Unsupervised Learning 5
      1.1.3 Reinforcement Learning 6
   1.2 Elements of Classification 7
      1.2.1 Classifiers 7
      1.2.2 Learning Criteria 8
      1.2.3 Optimization Algorithms 8
   1.3 Aspects of Machine Learning 9
      1.3.1 Logical Learning, Biologically Inspired Learning, and Statistical Learning 9
      1.3.2 Frequentist Learning and Bayesian Learning 11
      1.3.3 Generative Classification and Discriminative Classification 12
      1.3.4 Induction, Deduction, and Transduction 13
   1.4 Improving Data Collection and Weakly Supervised Learning 13
   1.5 Organization of This Book 15
      1.5.1 Weakly Supervised Learning for Binary Classification 15
      1.5.2 Weakly Supervised Learning for Multi-Class Classification 18
      1.5.3 Advanced Topics and Perspectives 18

2 Formulation and Notation 21
   2.1 Binary Classification 21
      2.1.1 Formulation 21
      2.1.2 Classification Models 22
         2.1.2.1 Linear-in-input model 22
2.1.2.2 Linear-in-parameter model 22
2.1.2.3 Kernel model 23
2.1.2.4 Neural network model 25
2.1.3 Surrogate Losses 26
2.1.4 Training Samples 28
2.1.5 Regularization 30
2.2 Multi-Class Classification 31
2.2.1 Formulation 31
2.2.2 Surrogate Losses 32
2.2.3 Training Samples 33

3 Supervised Classification 35
3.1 Positive-Negative (PN) Classification 35
3.1.1 Formulation 35
3.1.1.1 One-sample case 35
3.1.1.2 Two-sample case 36
3.1.1.3 Comparison 37
3.1.2 Theoretical Analysis 38
3.1.2.1 Targets of convergence 38
3.1.2.2 Measures of convergence 40
3.1.2.3 Rademacher complexity 43
3.1.2.4 Rademacher complexity bounds 46
3.1.2.5 Estimation error bounds 51
3.2 Multi-Class Classification 56
3.2.1 Formulation 56
3.2.2 Theoretical Analysis 58
3.2.2.1 Estimation error bounds 58
3.2.2.2 Classification calibration 61

II WEAKLY SUPERVISED LEARNING FOR BINARY CLASSIFICATION

4 Positive-Unlabeled (PU) Classification 67
4.1 Introduction 67
4.2 Formulation 68
4.3 Unbiased Risk Estimation from PU Data 69
4.3.1 General Approach 69
4.3.2 Cost-Sensitive Approach 71
4.3.3 Convex Approach 72
4.4 Theoretical Analysis 75
4.4.1 PU Classification  
4.4.2 NU Classification  
4.4.3 Comparisons with PN Classification  
  4.4.3.1 Finite-sample comparisons  
  4.4.3.2 Asymptotic comparisons

5 Positive-Negative-Unlabeled (PNU) Classification  
  5.1 Introduction  
  5.2 Formulation  
  5.3 Manifold-Based Semi-Supervised Classification  
    5.3.1 Laplacian Regularization  
    5.3.2 Implementation  
  5.4 Information-Theoretic Semi-Supervised Classification  
    5.4.1 Squared-Loss Mutual Information Regularization  
    5.4.2 Implementation  
  5.5 PU+PN Classification  
    5.5.1 PNU and PU+NU Risk Estimators  
    5.5.2 PNU vs. PU+NU Classification  
    5.5.3 Theoretical Analysis  
      5.5.3.1 Estimation error bounds  
      5.5.3.2 Variance reduction  
  5.6 Experiments  
    5.6.1 Datasets  
    5.6.2 PNU Risk for Validation  
    5.6.3 Comparison with Other Methods  
  5.7 Extensions  
    5.7.1 Multi-Class Extension  
    5.7.2 AUC Maximization  
    5.7.3 Matrix Imputation

6 Positive-Confidence (Pconf) Classification  
  6.1 Introduction  
  6.2 Related Works  
  6.3 Problem Formulation  
  6.4 Empirical Risk Minimization (ERM) Framework  
  6.5 Theoretical Analysis  
  6.6 Implementation  
  6.7 Experiments  
    6.7.1 Synthetic Experiments with Linear Models  
    6.7.2 Benchmark Experiments with Neural Network Models
7 Pairwise-Constraint Classification

7.1 Introduction

7.2 Formulation

7.2.1 One-Sample Case

7.2.2 Two-Sample Case

7.2.3 Comparison of Sampling Schemes

7.2.4 Pairwise Constraints as Pointwise Data

7.3 Similar-Unlabeled (SU) Classification

7.3.1 Classification Risk Estimation

7.3.2 Minimum-Variance Risk Estimation

7.3.3 Convex Formulation

7.3.4 Class-Priors in SU Classification

7.4 Similar-Dissimilar (SD) and Dissimilar-Unlabeled (DU) Classification

7.4.1 Classification Risk Estimation

7.4.2 Interpretation of SD Risk

7.5 Similar-Dissimilar-Unlabeled (SDU) Classification

7.6 Theoretical Analysis

7.6.1 Derivation of Estimation Error Bounds

7.6.2 Comparison of Estimation Error Bounds

7.7 Experiments

7.7.1 Setup

7.7.2 Illustration of SU Classification

7.7.3 Comparison of SU Classification with Other Methods

7.7.4 Comparison of SDU Classification with Other Methods

7.8 Ongoing Research

8 Unlabeled-Unlabeled (UU) Classification

8.1 Introduction

8.2 Problem Formulation

8.2.1 Data Generation Process

8.2.2 Performance Measures

8.2.3 Relation to Classification with Noisy Labels

8.3 Risk Estimation from UU Data

8.3.1 Risk Estimation from One Set of U Data

8.3.2 Risk Estimation from Two Sets of U Data

8.3.3 Theoretical Analysis

8.3.4 Experiments

8.3.4.1 Setup
8.3.4.2 Benchmark experiments with neural network models
8.3.4.3 Comparison with other methods

8.4 Generative Approach
  8.4.1 Analysis of Bayes-Optimal Classifier
  8.4.2 KDE-Based Algorithm
  8.4.3 LSDD-Based Algorithm
  8.4.4 DSDD-Based Algorithm
  8.4.5 Experiments

III WEAKLY SUPERVISED LEARNING FOR MULTI-CLASS CLASSIFICATION

9 Complementary-Label Classification
  9.1 Introduction
  9.2 Risk Estimation from CL Data
    9.2.1 Formulation
    9.2.2 Risk Estimation
    9.2.3 Case-Study for Symmetric Losses
    9.2.4 Relation to Classification with Noisy Labels
  9.3 Theoretical Analysis
  9.4 Incorporation of Ordinary-Labels
  9.5 Experiments
    9.5.1 Experiments with CL
    9.5.2 Experiments with CL and OL
  9.6 Incorporation of Multi-Complementary-Labels
    9.6.1 Formulation
    9.6.2 Comparison with Multiple Single CLs
    9.6.3 Unbiased Risk Estimator
    9.6.4 Estimation Error Bound

10 Partial-Label Classification
  10.1 Introduction
  10.2 Formulation and Assumptions
    10.2.1 Formulation
    10.2.2 Data Generation Assumption
  10.3 Risk Estimation
  10.4 Experiments
  10.5 Proper Partial-Label (PPL) Classification
    10.5.1 Data Generation Assumption
    10.5.2 Risk Estimation
    10.5.3 Theoretical Analysis
IV ADVANCED TOPICS AND PERSPECTIVES

11 Non-Negative Correction for Weakly Supervised Classification 207
  11.1 Introduction 207
  11.2 Overfitting of Unbiased Learning Objectives 208
    11.2.1 Binary Classification 208
    11.2.2 Multi-Class Classification 210
  11.3 Numerical Illustration 211
  11.4 Non-Negative Correction 213
    11.4.1 nnPU Classification 213
    11.4.2 nnPNU Classification 215
    11.4.3 nnUU Classification 216
    11.4.4 nnCL Classification 217
    11.4.5 ccUU Classification 217
  11.5 Theoretical Analyses 218
    11.5.1 Bias and Consistency 219
    11.5.2 Estimation Error 224
  11.6 Experiments 229
    11.6.1 Comparison of PN, uPU, and nnPU Classification 229
    11.6.2 Comparison of uCL and nnCL Classification 233
    11.6.3 Comparison of uUU and ccUU Classification 235

12 Class-Prior Estimation 239
  12.1 Introduction 239
  12.2 Full Distribution Matching 241
  12.3 Mixture Proportion Estimation 242
    12.3.1 Estimation Goal and Optimization Goal 243
    12.3.2 Redefinition of Optimization Goal 244
    12.3.3 Irreducibility Assumption 245
    12.3.4 Anchor Set/Point Assumption 246
    12.3.5 Remarks 248
  12.4 Partial Distribution Matching 248
    12.4.1 Formulation 248
    12.4.2 Differentiable Divergences 249
    12.4.3 Non-Differentiable Divergences 251
    12.4.4 Empirical f-Divergence Estimation 252
  12.5 Penalized $L_1$-Distance Minimization 254
    12.5.1 Penalized $L_1$-Distance 254
    12.5.2 Practical Implementation 256
12.5.3 Theoretical Analysis 258
   12.5.3.1 Realizability assumption 258
   12.5.3.2 Summary of main results 259
   12.5.3.3 Proofs of main results 259
   12.5.3.4 On the convergence rate of \( \hat{\pi}_p \) 261

12.6 Class-Prior Estimation with Regrouping 262
   12.6.1 Motivation 262
   12.6.2 Practical Implementation 263
   12.6.3 Theoretical Justification 265
      12.6.3.1 A formal definition of regrouping 265
      12.6.3.2 Bias reduction 267
      12.6.3.3 Convergence analysis 268
      12.6.3.4 Computationally efficient identification of \( A^* \) 270
      12.6.3.5 Approximation of \( p'_p \) with a surrogate 271

12.7 Class-Prior Estimation from Pairwise Data 273

13 Conclusions and Prospects 275

Notes 279
Bibliography 283
Index 293