MACHINE LEARNING IN NON-STATIONARY ENVIRONMENTS

Introduction to Covariate Shift Adaptation

Masashi Sugiyama and Motoaki Kawanabe

The MIT Press
Cambridge, Massachusetts
London, England
Contents

Foreword xi
Preface xiii

I INTRODUCTION

1 Introduction and Problem Formulation 3
  1.1 Machine Learning under Covariate Shift 3
  1.2 Quick Tour of Covariate Shift Adaptation 5
  1.3 Problem Formulation 7
    1.3.1 Function Learning from Examples 7
    1.3.2 Loss Functions 8
    1.3.3 Generalization Error 9
    1.3.4 Covariate Shift 9
    1.3.5 Models for Function Learning 10
    1.3.6 Specification of Models 13
  1.4 Structure of This Book 14
    1.4.1 Part II: Learning under Covariate Shift 14
    1.4.2 Part III: Learning Causing Covariate Shift 17

II LEARNING UNDER COVARIATE SHIFT

2 Function Approximation 21
  2.1 Importance-Weighting Techniques for Covariate Shift Adaptation 22
    2.1.1 Importance-Weighted ERM 22
    2.1.2 Adaptive IWERM 23
    2.1.3 Regularized IWERM 23
  2.2 Examples of Importance-Weighted Regression Methods 25
    2.2.1 Squared Loss: Least-Squares Regression 26
    2.2.2 Absolute Loss: Least-Absolute Regression 30
    2.2.3 Huber Loss: Huber Regression 31
    2.2.4 Deadzone-Linear Loss: Support Vector Regression 33
  2.3 Examples of Importance-Weighted Classification Methods 35
2.3.1 Squared Loss: Fisher Discriminant Analysis 36
2.3.2 Logistic Loss: Logistic Regression Classifier 38
2.3.3 Hinge Loss: Support Vector Machine 39
2.3.4 Exponential Loss: Boosting 40
2.4 Numerical Examples 40
  2.4.1 Regression 40
  2.4.2 Classification 41
2.5 Summary and Discussion 45

3  Model Selection 47
  3.1 Importance-Weighted Akaike Information Criterion 47
  3.2 Importance-Weighted Subspace Information Criterion 50
    3.2.1 Input Dependence vs. Input Independence in Generalization Error Analysis 51
    3.2.2 Approximately Correct Models 53
    3.2.3 Input-Dependent Analysis of Generalization Error 54
  3.3 Importance-Weighted Cross-Validation 64
  3.4 Numerical Examples 66
    3.4.1 Regression 66
    3.4.2 Classification 69
  3.5 Summary and Discussion 70

4  Importance Estimation 73
  4.1 Kernel Density Estimation 73
  4.2 Kernel Mean Matching 75
  4.3 Logistic Regression 76
  4.4 Kullback–Leibler Importance Estimation Procedure 78
    4.4.1 Algorithm 78
    4.4.2 Model Selection by Cross-Validation 81
    4.4.3 Basis Function Design 82
  4.5 Least-Squares Importance Fitting 83
    4.5.1 Algorithm 83
    4.5.2 Basis Function Design and Model Selection 84
    4.5.3 Regularization Path Tracking 85
  4.6 Unconstrained Least-Squares Importance Fitting 87
    4.6.1 Algorithm 87
    4.6.2 Analytic Computation of Leave-One-Out Cross-Validation 88
  4.7 Numerical Examples 88
    4.7.1 Setting 90
    4.7.2 Importance Estimation by KLIEP 90
    4.7.3 Covariate Shift Adaptation by IWLS and IWCV 92
  4.8 Experimental Comparison 94
  4.9 Summary 101

5  Direct Density-Ratio Estimation with Dimensionality Reduction 103
  5.1 Density Difference in Hetero-Distributional Subspace 103
  5.2 Characterization of Hetero-Distributional Subspace 104
5.3 Identifying Hetero-Distributional Subspace 106
  5.3.1 Basic Idea 106
  5.3.2 Fisher Discriminant Analysis 108
  5.3.3 Local Fisher Discriminant Analysis 109
5.4 Using LFDA for Finding Hetero-Distributional Subspace 112
5.5 Density-Ratio Estimation in the Hetero-Distributional Subspace 113
5.6 Numerical Examples 113
  5.6.1 Illustrative Example 113
  5.6.2 Performance Comparison Using Artificial Data Sets 117
5.7 Summary 121

6 Relation to Sample Selection Bias 125
  6.1 Heckman's Sample Selection Model 125
  6.2 Distributional Change and Sample Selection Bias 129
  6.3 The Two-Step Algorithm 131
  6.4 Relation to Covariate Shift Approach 134

7 Applications of Covariate Shift Adaptation 137
  7.1 Brain–Computer Interface 137
    7.1.1 Background 137
    7.1.2 Experimental Setup 138
    7.1.3 Experimental Results 140
  7.2 Speaker Identification 142
    7.2.1 Background 142
    7.2.2 Formulation 142
    7.2.3 Experimental Results 144
  7.3 Natural Language Processing 149
    7.3.1 Formulation 149
    7.3.2 Experimental Results 151
  7.4 Perceived Age Prediction from Face Images 152
    7.4.1 Background 152
    7.4.2 Formulation 153
    7.4.3 Incorporating Characteristics of Human Age Perception 153
    7.4.4 Experimental Results 155
  7.5 Human Activity Recognition from Accelerometric Data 157
    7.5.1 Background 157
    7.5.2 Importance-Weighted Least-Squares Probabilistic Classifier 157
    7.5.3 Experimental Results 160
  7.6 Sample Reuse in Reinforcement Learning 165
    7.6.1 Markov Decision Problems 165
    7.6.2 Policy Iteration 166
    7.6.3 Value Function Approximation 167
    7.6.4 Sample Reuse by Covariate Shift Adaptation 168
    7.6.5 On-Policy vs. Off-Policy 169
    7.6.6 Importance Weighting in Value Function Approximation 170
    7.6.7 Automatic Selection of the Flattening Parameter 174
### Contents

- 10.1.4 Estimating Generalization Error for Active Learning 229
- 10.1.5 Designing Sampling Policies 230
- 10.1.6 Active Learning in Policy Iteration 231
- 10.1.7 Robot Control Experiments 232
- 10.2 Wafer Alignment in Semiconductor Exposure Apparatus 234

## IV CONCLUSIONS

11 Conclusions and Future Prospects 241

11.1 Conclusions 241
11.2 Future Prospects 242

Appendix: List of Symbols and Abbreviations 243
Bibliography 247
Index 259