Contents

Acknowledgements **iv**

Abstract **v**

I Optimal Data Distributions in Machine Learning **1**

1 Introduction **2**

1.1 Overview 3

1.2 Literature overview 4

1.3 The learning setup 5

2 Is it better to have $P_R = P_S$? **7**

2.1 Empirical results in the classification setting 8

2.1.1 Fixing the training distribution 10

2.1.2 Fixing the test distribution 11

2.2 Empirical and analytic results in the regression setting 13

3 The dual distribution **20**

3.1 Discrete input spaces 20

3.2 The continuous case 23

3.2.1 Analytic condition for the dual distribution 24

3.2.2 Dual distribution examples 27

3.3 Variability of the dual distribution 29

3.3.1 Asymptotic behavior 30

3.3.2 Effect of noise and complexity 32

3.4 Using the dual distribution in a practical setting 33

3.5 Computational and implementation details 36

3.6 Differences with active learning 39

4 To weight or not to weight **41**

4.1 What makes weighting work sometimes only? . . . 41

4.1.1 The effective sample size . . . 42

4.1.2 Weighting vs sampling . . . 44

4.1.3 Decomposition of the weighting effect . . . 49

4.2 The algorithm: Targeted matching . . . 50

4.3 Experimental results . . . 53

4.3.1 Results on the Netflix dataset . . . 53

4.3.2 Results on further benchmark datasets . . . 55

5 A novel class of matching algorithms **57**

5.1 Previous algorithms . . . 57

5.1.1 Indirect ratio estimation via KDE . . . 57

5.1.2 Logistic regression methods . . . 58

5.1.3 Kernel mean matching (KMM) . . . 59

5.1.4 Parametric models for the ratio: KLIEP, LSIF, RuLSIF, etc. . . . 59

5.1.5 Discrepancy minimization . . . 61

5.2 A new class of algorithms . . . 63

5.2.1 Hard matching . . . 63

5.2.2 Soft Matching . . . 67

5.2.3 Hard matching with slack variables . . . 70

5.2.4 Statistical approach . . . 72

5.2.5 Probabilistic approach . . . 74

II Behavior Analysis with Machine Learning **77**

6 Supervised behavior classification **78**

6.1 Pre-processing stage . . . 79

6.2 Learning algorithm . . . 80

6.3 Post-processing stage . . . 80

7 CUBA: Caltech Unsupervised Behavior Analysis **82**

7.1 Problem statement . . . 83

7.2 The method . . . 84

7.2.1 Detecting movemes . . . 84

7.2.2 Detecting actions . . . 87

7.2.3 Finding stories 88
7.3 Results on real datasets 90
7.3.1 CUBA in the Fear in Flies dataset 90
7.3.2 CUBA in the Fly Bowl dataset 101

Conclusion **106**

A An analytic learning setup **108**

Bibliography **111**