Contents

Preface xiii

1 Introduction 1
  1.1 What is machine learning? 1
  1.2 What kind of problems can be tackled using machine learning? 2
  1.3 Some standard learning tasks 3
  1.4 Learning stages 4
  1.5 Learning scenarios 6
  1.6 Generalization 7

2 The PAC Learning Framework 9
  2.1 The PAC learning model 9
  2.2 Guarantees for finite hypothesis sets — consistent case 15
  2.3 Guarantees for finite hypothesis sets — inconsistent case 19
  2.4 Generalities 21
    2.4.1 Deterministic versus stochastic scenarios 21
    2.4.2 Bayes error and noise 22
  2.5 Chapter notes 23
  2.6 Exercises 23

3 Rademacher Complexity and VC-Dimension 29
  3.1 Rademacher complexity 30
  3.2 Growth function 34
  3.3 VC-dimension 36
  3.4 Lower bounds 43
  3.5 Chapter notes 48
  3.6 Exercises 50

4 Model Selection 61
  4.1 Estimation and approximation errors 61
  4.2 Empirical risk minimization (ERM) 62
  4.3 Structural risk minimization (SRM) 64
10.4 RankBoost 244
  10.4.1 Bound on the empirical error 246
  10.4.2 Relationship with coordinate descent 248
  10.4.3 Margin bound for ensemble methods in ranking 250
10.5 Bipartite ranking 251
  10.5.1 Boosting in bipartite ranking 252
  10.5.2 Area under the ROC curve 255
10.6 Preference-based setting 257
  10.6.1 Second-stage ranking problem 257
  10.6.2 Deterministic algorithm 259
  10.6.3 Randomized algorithm 260
  10.6.4 Extension to other loss functions 262
10.7 Other ranking criteria 262
10.8 Chapter notes 263
10.9 Exercises 264

11 Regression 267
  11.1 The problem of regression 267
  11.2 Generalization bounds 268
    11.2.1 Finite hypothesis sets 268
    11.2.2 Rademacher complexity bounds 269
    11.2.3 Pseudo-dimension bounds 271
  11.3 Regression algorithms 275
    11.3.1 Linear regression 275
    11.3.2 Kernel ridge regression 276
    11.3.3 Support vector regression 281
    11.3.4 Lasso 285
    11.3.5 Group norm regression algorithms 289
    11.3.6 On-line regression algorithms 289
  11.4 Chapter notes 290
  11.5 Exercises 292

12 Maximum Entropy Models 295
  12.1 Density estimation problem 295
    12.1.1 Maximum Likelihood (ML) solution 296
    12.1.2 Maximum a Posteriori (MAP) solution 297
  12.2 Density estimation problem augmented with features 297
  12.3 Maxent principle 298
  12.4 Maxent models 299
  12.5 Dual problem 299
  12.6 Generalization bound 303
  12.7 Coordinate descent algorithm 304
  12.8 Extensions 306
  12.9 $L_2$-regularization 308
## Contents

### B Convex Optimization
- B.1 Differentiation and unconstrained optimization 415
- B.2 Convexity 415
- B.3 Constrained optimization 419
- B.4 Fenchel duality 422
  - B.4.1 Subgradients 422
  - B.4.2 Core 423
  - B.4.3 Conjugate functions 423
- B.5 Chapter notes 426
- B.6 Exercises 427

### C Probability Review
- C.1 Probability 429
- C.2 Random variables 429
- C.3 Conditional probability and independence 431
- C.4 Expectation and Markov's inequality 431
- C.5 Variance and Chebyshev's inequality 432
- C.6 Moment-generating functions 434
- C.7 Exercises 435

### D Concentration Inequalities
- D.1 Hoeffding's inequality 437
- D.2 Sanov's theorem 438
- D.3 Multiplicative Chernoff bounds 439
- D.4 Binomial distribution tails: Upper bounds 440
- D.5 Binomial distribution tails: Lower bound 440
- D.6 Azuma's inequality 441
- D.7 McDiarmid's inequality 442
- D.8 Normal distribution tails: Lower bound 443
- D.9 Khintchine-Kahane inequality 443
- D.10 Maximal inequality 444
- D.11 Chapter notes 445
- D.12 Exercises 445

### E Notions of Information Theory
- E.1 Entropy 449
- E.2 Relative entropy 450
- E.3 Mutual information 453
- E.4 Bregman divergences 453
- E.5 Chapter notes 456
- E.6 Exercises 457