

# CONTENTS

---

## PREFACE

xvii

## **1** INTRODUCTION

---

1

- 1.1 Machine Perception, 1
- 1.2 An Example, 1
  - 1.2.1 Related Fields, 8
- 1.3 Pattern Recognition Systems, 9
  - 1.3.1 Sensing, 9
  - 1.3.2 Segmentation and Grouping, 9
  - 1.3.3 Feature Extraction, 11
  - 1.3.4 Classification, 12
  - 1.3.5 Post Processing, 13
- 1.4 The Design Cycle, 14
  - 1.4.1 Data Collection, 14
  - 1.4.2 Feature Choice, 14
  - 1.4.3 Model Choice, 15
  - 1.4.4 Training, 15
  - 1.4.5 Evaluation, 15
  - 1.4.6 Computational Complexity, 16
- 1.5 Learning and Adaptation, 16
  - 1.5.1 Supervised Learning, 16
  - 1.5.2 Unsupervised Learning, 17
  - 1.5.3 Reinforcement Learning, 17
- 1.6 Conclusion, 17
- Summary by Chapters, 17
- Bibliographical and Historical Remarks, 18
- Bibliography, 19

## **2** BAYESIAN DECISION THEORY

---

20

- 2.1 Introduction, 20
- 2.2 Bayesian Decision Theory—Continuous Features, 24
  - 2.2.1 Two-Category Classification, 25
- 2.3 Minimum-Error-Rate Classification, 26
  - \*2.3.1 Minimax Criterion, 27

vii

- 6.9.3 Quickprop, 320
- 6.9.4 Conjugate Gradient Descent, 321
  - Example 1 Conjugate Gradient Descent, 322
- \*6.10 Additional Networks and Training Methods, 324
  - 6.10.1 Radial Basis Function Networks (RBFs), 324
  - 6.10.2 Special Bases, 325
  - 6.10.3 Matched Filters, 325
  - 6.10.4 Convolutional Networks, 326
  - 6.10.5 Recurrent Networks, 328
  - 6.10.6 Cascade-Correlation, 329
- 6.11 Regularization, Complexity Adjustment and Pruning, 330
- Summary, 333
- Bibliographical and Historical Remarks, 333
- Problems, 335
- Computer exercises, 343
- Bibliography, 347

## 7

 STOCHASTIC METHODS

350

- 7.1 Introduction, 350
- 7.2 Stochastic Search, 351
  - 7.2.1 Simulated Annealing, 351
  - 7.2.2 The Boltzmann Factor, 352
  - 7.2.3 Deterministic Simulated Annealing, 357
- 7.3 Boltzmann Learning, 360
  - 7.3.1 Stochastic Boltzmann Learning of Visible States, 360
  - 7.3.2 Missing Features and Category Constraints, 365
  - 7.3.3 Deterministic Boltzmann Learning, 366
  - 7.3.4 Initialization and Setting Parameters, 367
- \*7.4 Boltzmann Networks and Graphical Models, 370
  - 7.4.1 Other Graphical Models, 372
- \*7.5 Evolutionary Methods, 373
  - 7.5.1 Genetic Algorithms, 373
  - 7.5.2 Further Heuristics, 377
  - 7.5.3 Why Do They Work?, 378
- \*7.6 Genetic Programming, 378
- Summary, 381
- Bibliographical and Historical Remarks, 381
- Problems, 383
- Computer exercises, 388
- Bibliography, 391

## 8

 NONMETRIC METHODS

394

- 8.1 Introduction, 394
- 8.2 Decision Trees, 395
- 8.3 CART, 396
  - 8.3.1 Number of Splits, 397
  - 8.3.2 Query Selection and Node Impurity, 398
  - 8.3.3 When to Stop Splitting, 402
  - 8.3.4 Pruning, 403

- 8.3.5 Assignment of Leaf Node Labels, 404
  - Example 1 A Simple Tree, 404
- 8.3.6 Computational Complexity, 406
- 8.3.7 Feature Choice, 407
- 8.3.8 Multivariate Decision Trees, 408
- 8.3.9 Priors and Costs, 409
- 8.3.10 Missing Attributes, 409
  - Example 2 Surrogate Splits and Missing Attributes, 410
- 8.4 Other Tree Methods, 411
  - 8.4.1 ID3, 411
  - 8.4.2 C4.5, 411
  - 8.4.3 Which Tree Classifier Is Best?, 412
- \*8.5 Recognition with Strings, 413
  - 8.5.1 String Matching, 415
  - 8.5.2 Edit Distance, 418
  - 8.5.3 Computational Complexity, 420
  - 8.5.4 String Matching with Errors, 420
  - 8.5.5 String Matching with the "Don't-Care" Symbol, 421
- 8.6 Grammatical Methods, 421
  - 8.6.1 Grammars, 422
  - 8.6.2 Types of String Grammars, 424
    - Example 3 A Grammar for Pronouncing Numbers, 425
  - 8.6.3 Recognition Using Grammars, 426
- 8.7 Grammatical Inference, 429
  - Example 4 Grammatical Inference, 431
- \*8.8 Rule-Based Methods, 431
  - 8.8.1 Learning Rules, 433
- Summary, 434
- Bibliographical and Historical Remarks, 435
- Problems, 437
- Computer exercises, 446
- Bibliography, 450

## 9

 ALGORITHM-INDEPENDENT MACHINE LEARNING

453

- 9.1 Introduction, 453
- 9.2 Lack of Inherent Superiority of Any Classifier, 454
  - 9.2.1 No Free Lunch Theorem, 454
    - Example 1 No Free Lunch for Binary Data, 457
  - \*9.2.2 Ugly Duckling Theorem, 458
  - 9.2.3 Minimum Description Length (MDL), 461
  - 9.2.4 Minimum Description Length Principle, 463
  - 9.2.5 Overfitting Avoidance and Occam's Razor, 464
- 9.3 Bias and Variance, 465
  - 9.3.1 Bias and Variance for Regression, 466
  - 9.3.2 Bias and Variance for Classification, 468
- 9.4 Resampling for Estimating Statistics, 471
  - 9.4.1 Jackknife, 472
    - Example 2 Jackknife Estimate of Bias and Variance of the Mode, 473
  - 9.4.2 Bootstrap, 474
- 9.5 Resampling for Classifier Design, 475

- 9.5.1 Bagging, 475
- 9.5.2 Boosting, 476
- 9.5.3 Learning with Queries, 480
- 9.5.4 Arcing, Learning with Queries, Bias and Variance, 482
- 9.6 Estimating and Comparing Classifiers, 482
  - 9.6.1 Parametric Models, 483
  - 9.6.2 Cross-Validation, 483
  - 9.6.3 Jackknife and Bootstrap Estimation of Classification Accuracy, 485
  - 9.6.4 Maximum-Likelihood Model Comparison, 486
  - 9.6.5 Bayesian Model Comparison, 487
  - 9.6.6 The Problem-Average Error Rate, 489
  - 9.6.7 Predicting Final Performance from Learning Curves, 492
  - 9.6.8 The Capacity of a Separating Plane, 494
- 9.7 Combining Classifiers, 495
  - 9.7.1 Component Classifiers with Discriminant Functions, 496
  - 9.7.2 Component Classifiers without Discriminant Functions, 498
- Summary, 499
- Bibliographical and Historical Remarks, 500
- Problems, 502
- Computer exercises, 508
- Bibliography, 513

## 10 UNSUPERVISED LEARNING AND CLUSTERING 517

- 10.1 Introduction, 517
- 10.2 Mixture Densities and Identifiability, 518
- 10.3 Maximum-Likelihood Estimates, 519
- 10.4 Application to Normal Mixtures, 521
  - 10.4.1 Case 1: Unknown Mean Vectors, 522
  - 10.4.2 Case 2: All Parameters Unknown, 524
  - 10.4.3  $k$ -Means Clustering, 526
  - \*10.4.4 Fuzzy  $k$ -Means Clustering, 528
- 10.5 Unsupervised Bayesian Learning, 530
  - 10.5.1 The Bayes Classifier, 530
  - 10.5.2 Learning the Parameter Vector, 531
    - Example 1 Unsupervised Learning of Gaussian Data, 534
  - 10.5.3 Decision-Directed Approximation, 536
- 10.6 Data Description and Clustering, 537
  - 10.6.1 Similarity Measures, 538
- 10.7 Criterion Functions for Clustering, 542
  - 10.7.1 The Sum-of-Squared-Error Criterion, 542
  - 10.7.2 Related Minimum Variance Criteria, 543
  - 10.7.3 Scatter Criteria, 544
    - Example 2 Clustering Criteria, 546
- \*10.8 Iterative Optimization, 548
- 10.9 Hierarchical Clustering, 550
  - 10.9.1 Definitions, 551
  - 10.9.2 Agglomerative Hierarchical Clustering, 552
  - 10.9.3 Stepwise-Optimal Hierarchical Clustering, 555
  - 10.9.4 Hierarchical Clustering and Induced Metrics, 556
- \*10.10 The Problem of Validity, 557

- \*10.11 On-line clustering, 559
  - 10.11.1 Unknown Number of Clusters, 561
  - 10.11.2 Adaptive Resonance, 563
  - 10.11.3 Learning with a Critic, 565
- \*10.12 Graph-Theoretic Methods, 566
- 10.13 Component Analysis, 568
  - 10.13.1 Principal Component Analysis (PCA), 568
  - 10.13.2 Nonlinear Component Analysis (NLCA), 569
  - \*10.13.3 Independent Component Analysis (ICA), 570
- 10.14 Low-Dimensional Representations and Multidimensional Scaling (MDS), 573
  - 10.14.1 Self-Organizing Feature Maps, 576
  - 10.14.2 Clustering and Dimensionality Reduction, 580
- Summary, 581
- Bibliographical and Historical Remarks, 582
- Problems, 583
- Computer exercises, 593
- Bibliography, 598

## A MATHEMATICAL FOUNDATIONS 601

- A.1 Notation, 601
- A.2 Linear Algebra, 604
  - A.2.1 Notation and Preliminaries, 604
  - A.2.2 Inner Product, 605
  - A.2.3 Outer Product, 606
  - A.2.4 Derivatives of Matrices, 606
  - A.2.5 Determinant and Trace, 608
  - A.2.6 Matrix Inversion, 609
  - A.2.7 Eigenvectors and Eigenvalues, 609
- A.3 Lagrange Optimization, 610
- A.4 Probability Theory, 611
  - A.4.1 Discrete Random Variables, 611
  - A.4.2 Expected Values, 611
  - A.4.3 Pairs of Discrete Random Variables, 612
  - A.4.4 Statistical Independence, 613
  - A.4.5 Expected Values of Functions of Two Variables, 613
  - A.4.6 Conditional Probability, 614
  - A.4.7 The Law of Total Probability and Bayes Rule, 615
  - A.4.8 Vector Random Variables, 616
  - A.4.9 Expectations, Mean Vectors and Covariance Matrices, 617
  - A.4.10 Continuous Random Variables, 618
  - A.4.11 Distributions of Sums of Independent Random Variables, 620
  - A.4.12 Normal Distributions, 621
- A.5 Gaussian Derivatives and Integrals, 623
  - A.5.1 Multivariate Normal Densities, 624
  - A.5.2 Bivariate Normal Densities, 626
- A.6 Hypothesis Testing, 628
  - A.6.1 Chi-Squared Test, 629
- A.7 Information Theory, 630
  - A.7.1 Entropy and Information, 630

A.7.2 Relative Entropy, 632  
A.7.3 Mutual Information, 632  
A.8 Computational Complexity, 633  
Bibliography, 635

**INDEX**