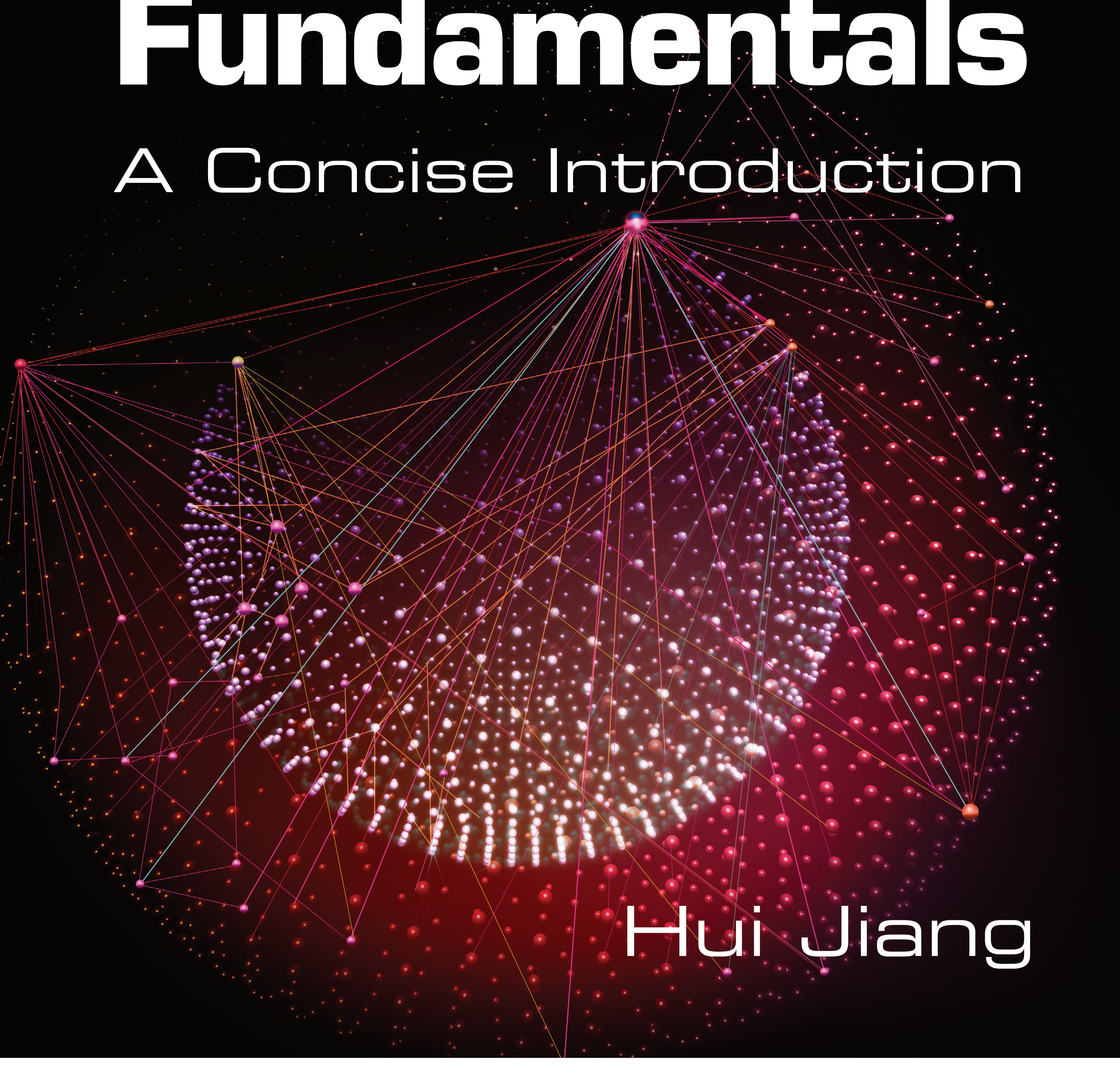
Machine Learning Fundamentals

A Concise Introduction



Hui Jiang

Machine Learning Fundamentals

A Concise Introduction

Hui Jiang

York University, Toronto

CAMBRIDGE
UNIVERSITY PRESS

Contents

Preface **xi**

Notation **xvii**

1 Introduction **1**

1.1 **What Is Machine Learning?** 1

1.2 **Basic Concepts in Machine Learning** 4

1.2.1 Classification versus Regression 4

1.2.2 Supervised versus Unsupervised Learning 5

1.2.3 Simple versus Complex Models 5

1.2.4 Parametric versus Nonparametric Models 7

1.2.5 Overfitting versus Underfitting 8

1.2.6 Bias–Variance Trade-Off 10

1.3 **General Principles in Machine Learning** 11

1.3.1 Occam's Razor 11

1.3.2 No-Free-Lunch Theorem 11

1.3.3 Law of the Smooth World 12

1.3.4 Curse of Dimensionality 14

1.4 **Advanced Topics in Machine Learning** 15

1.4.1 Reinforcement Learning 15

1.4.2 Meta-Learning 16

1.4.3 Causal Inference 16

1.4.4 Other Advanced Topics 16

Exercises 18

2 Mathematical Foundation **19**

2.1 **Linear Algebra** 19

2.1.1 Vectors and Matrices 19

2.1.2 Linear Transformation as Matrix Multiplication 20

2.1.3 Basic Matrix Operations 21

2.1.4 Eigenvalues and Eigenvectors . . . 23
2.1.5 Matrix Calculus . . . 25
2.2 **Probability and Statistics** . . . 27
2.2.1 Random Variables and Distributions . . . 27
2.2.2 Expectation: Mean, Variance, and Moments . . . 28
2.2.3 Joint, Marginal, and Conditional Distributions . . . 30
2.2.4 Common Probability Distributions . . . 33
2.2.5 Transformation of Random Variables . . . 40
2.3 **Information Theory** . . . 41
2.3.1 Information and Entropy . . . 41
2.3.2 Mutual Information . . . 43
2.3.3 KL Divergence . . . 46
2.4 **Mathematical Optimization** . . . 48
2.4.1 General Formulation . . . 49
2.4.2 Optimality Conditions . . . 50
2.4.3 Numerical Optimization Methods . . . 59
Exercises . . . 64

3 Supervised Machine Learning (in a Nutshell) 67
3.1 **Overview** . . . 67
3.2 **Case Studies** . . . 72

4 Feature Extraction 77
4.1 **Feature Extraction: Concepts** . . . 77
4.1.1 Feature Engineering . . . 77
4.1.2 Feature Selection . . . 78
4.1.3 Dimensionality Reduction . . . 79
4.2 **Linear Dimension Reduction** . . . 79
4.2.1 Principal Component Analysis . . . 80
4.2.2 Linear Discriminant Analysis . . . 84
4.3 **Nonlinear Dimension Reduction (I): Manifold Learning** . . . 86
4.3.1 Locally Linear Embedding . . . 87
4.3.2 Multidimensional Scaling . . . 88
4.3.3 Stochastic Neighborhood Embedding . . . 89
4.4 **Nonlinear Dimension Reduction (II): Neural Networks** . . . 90
4.4.1 Autoencoder . . . 90
4.4.2 Bottleneck Features . . . 91
Lab Project I . . . 92
Exercises . . . 93

DISCRIMINATIVE MODELS **95**

5 Statistical Learning Theory **97**
5.1 **Formulation of Discriminative Models** 97
5.2 **Learnability** 99
5.3 **Generalization Bounds** 100
5.3.1 Finite Model Space: $|\mathbb{H}|$ 100
5.3.2 Infinite Model Space: VC Dimension 102
Exercises 105

6 Linear Models **107**
6.1 **Perceptron** 108
6.2 **Linear Regression** 112
6.3 **Minimum Classification Error** 113
6.4 **Logistic Regression** 114
6.5 **Support Vector Machines** 116
6.5.1 Linear SVM 116
6.5.2 Soft SVM 121
6.5.3 Nonlinear SVM: The Kernel Trick 123
6.5.4 Solving Quadratic Programming 126
6.5.5 Multiclass SVM 127
Lab Project II 129
Exercises 130

7 Learning Discriminative Models in General **133**
7.1 **A General Framework to Learn Discriminative Models** 133
7.1.1 Common Loss Functions in Machine Learning 135
7.1.2 Regularization Based on L_p Norm 136
7.2 **Ridge Regression and LASSO** 139
7.3 **Matrix Factorization** 140
7.4 **Dictionary Learning** 145
Lab Project III 149
Exercises 150

8 Neural Networks **151**
8.1 **Artificial Neural Networks** 152
8.1.1 Basic Formulation of Artificial Neural Networks 152
8.1.2 Mathematical Justification: Universal Approximator . . . 154
8.2 **Neural Network Structures** 156
8.2.1 Basic Building Blocks to Connect Layers 156
8.2.2 Case Study I: Fully Connected Deep Neural Networks . . 165
8.2.3 Case Study II: Convolutional Neural Networks 166
8.2.4 Case Study III: Recurrent Neural Networks (RNNs) 170

8.2.5 Case Study IV: Transformer 172
8.3 **Learning Algorithms for Neural Networks** 174
8.3.1 Loss Function 175
8.3.2 Automatic Differentiation 176
8.3.3 Optimization Using Stochastic Gradient Descent 188
8.4 **Heuristics and Tricks for Optimization** 189
8.4.1 Other SGD Variant Optimization Methods: ADAM 192
8.4.2 Regularization 194
8.4.3 Fine-Tuning Tricks 196
8.5 **End-to-End Learning** 197
8.5.1 Sequence-to-Sequence Learning 198
Lab Project IV 200
Exercises 201

9 Ensemble Learning 203
9.1 **Formulation of Ensemble Learning** 203
9.1.1 Decision Trees 205
9.2 **Bagging** 208
9.2.1 Random Forests 208
9.3 **Boosting** 209
9.3.1 Gradient Boosting 210
9.3.2 AdaBoost 212
9.3.3 Gradient Tree Boosting 214
Lab Project V 216
Exercises 217

GENERATIVE MODELS 219

10 Overview of Generative Models 221
10.1 **Formulation of Generative Models** 221
10.2 **Bayesian Decision Theory** 222
10.2.1 Generative Models for Classification 223
10.2.2 Generative Models for Regression 227
10.3 **Statistical Data Modeling** 228
10.3.1 Plug-In MAP Decision Rule 229
10.4 **Density Estimation** 231
10.4.1 Maximum-Likelihood Estimation 231
10.4.2 Maximum-Likelihood Classifier 234
10.5 **Generative Models (in a Nutshell)** 234
10.5.1 Generative versus Discriminative Models 236
Exercises 237

11 Unimodal Models 239

11.1 **Gaussian Models** 240

11.2 **Multinomial Models** 243

11.3 **Markov Chain Models** 245

11.4 **Generalized Linear Models** 250

11.4.1 Probit Regression 252

11.4.2 Poisson Regression 252

11.4.3 Log-Linear Models 253

Exercises 256

12 Mixture Models 257

12.1 **Formulation of Mixture Models** 257

12.1.1 Exponential Family (e-Family) 259

12.1.2 Formal Definition of Mixture Models 261

12.2 **Expectation-Maximization Method** 261

12.2.1 Auxiliary Function: Eliminating Log-Sum 262

12.2.2 Expectation-Maximization Algorithm 265

12.3 **Gaussian Mixture Models** 268

12.3.1 K-Means Clustering for Initialization 270

12.4 **Hidden Markov Models** 271

12.4.1 HMMs: Mixture Models for Sequences 272

12.4.2 Evaluation Problem: Forward–Backward Algorithm 276

12.4.3 Decoding Problem: Viterbi Algorithm 279

12.4.4 Training Problem: Baum–Welch Algorithm 280

Lab Project VI 287

Exercises 288

13 Entangled Models 291

13.1 **Formulation of Entangled Models** 291

13.1.1 Framework of Entangled Models 292

13.1.2 Learning of Entangled Models in General 294

13.2 **Linear Gaussian Models** 296

13.2.1 Probabilistic PCA 296

13.2.2 Factor Analysis 298

13.3 **Non-Gaussian Models** 300

13.3.1 Independent Component Analysis (ICA) 300

13.3.2 Independent Factor Analysis (IFA) 301

13.3.3 Hybrid Orthogonal Projection and Estimation (HOPE) . . 302

13.4 **Deep Generative Models** 303

13.4.1 Variational Autoencoders (VAE) 304

13.4.2 Generative Adversarial Nets (GAN) 307

Exercises 309

14 Bayesian Learning 311
14.1 **Formulation of Bayesian Learning** . . . 311
14.1.1 Bayesian Inference . . . 313
14.1.2 Maximum a Posterior Estimation . . . 314
14.1.3 Sequential Bayesian Learning . . . 315
14.2 **Conjugate Priors** . . . 318
14.2.1 Maximum-Marginal-Likelihood Estimation . . . 323
14.3 **Approximate Inference** . . . 324
14.3.1 Laplace's Method . . . 324
14.3.2 Variational Bayesian (VB) Methods . . . 326
14.4 **Gaussian Processes** . . . 332
14.4.1 Gaussian Processes as Nonparametric Priors . . . 333
14.4.2 Gaussian Processes for Regression . . . 335
14.4.3 Gaussian Processes for Classification . . . 338
Exercises . . . 340

15 Graphical Models 343
15.1 **Concepts of Graphical Models** . . . 343
15.2 **Bayesian Networks** . . . 346
15.2.1 Conditional Independence . . . 346
15.2.2 Representing Generative Models as Bayesian Networks . 351
15.2.3 Learning Bayesian Networks . . . 353
15.2.4 Inference Algorithms . . . 355
15.2.5 Case Study I: Naive Bayes Classifier . . . 361
15.2.6 Case Study II: Latent Dirichlet Allocation . . . 362
15.3 **Markov Random Fields** . . . 366
15.3.1 Formulation: Potential and Partition Functions . . . 366
15.3.2 Case Study III: Conditional Random Fields . . . 368
15.3.3 Case Study IV: Restricted Boltzmann Machines . . . 370
Exercises . . . 372

APPENDIX 375

A Other Probability Distributions 377

Bibliography 381

Index 397