

# Contents

<b>1</b>	<b>Science of Information</b> . . . . .	1
1.1	Data Science . . . . .	1
1.1.1	Technological Dilemma . . . . .	2
1.1.2	Technological Advancement . . . . .	2
1.2	Big Data Paradigm . . . . .	3
1.2.1	Facts and Statistics of a System . . . . .	3
1.2.2	Big Data Versus Regular Data . . . . .	5
1.3	Machine Learning Paradigm . . . . .	7
1.3.1	Modeling and Algorithms . . . . .	7
1.3.2	Supervised and Unsupervised . . . . .	7
1.4	Collaborative Activities . . . . .	10
1.5	A Snapshot . . . . .	10
1.5.1	The Purpose and Interests . . . . .	10
1.5.2	The Goal and Objectives . . . . .	11
1.5.3	The Problems and Challenges . . . . .	11
	Problems . . . . .	11
	References . . . . .	12

## Part I Understanding Big Data

<b>2</b>	<b>Big Data Essentials</b> . . . . .	17
2.1	Big Data Analytics . . . . .	17
2.1.1	Big Data Controllers . . . . .	18
2.1.2	Big Data Problems . . . . .	19
2.1.3	Big Data Challenges . . . . .	19
2.1.4	Big Data Solutions . . . . .	20
2.2	Big Data Classification . . . . .	20
2.2.1	Representation Learning . . . . .	21
2.2.2	Distributed File Systems . . . . .	22
2.2.3	Classification Modeling . . . . .	23
2.2.4	Classification Algorithms . . . . .	25

2.3	Big Data Scalability	26
2.3.1	High-Dimensional Systems	27
2.3.2	Low-Dimensional Structures	27
	Problems	28
	References	28
<b>3</b>	<b>Big Data Analytics</b>	<b>31</b>
3.1	Analytics Fundamentals	31
3.1.1	Research Questions	32
3.1.2	Choices of Data Sets	33
3.2	Pattern Detectors	34
3.2.1	Statistical Measures	34
3.2.2	Graphical Measures	38
3.2.3	Coding Example	41
3.3	Patterns of Big Data	44
3.3.1	Standardization: A Coding Example	47
3.3.2	Evolution of Patterns	49
3.3.3	Data Expansion Modeling	51
3.3.4	Deformation of Patterns	62
3.3.5	Classification Errors	66
3.4	Low-Dimensional Structures	67
3.4.1	A Toy Example	67
3.4.2	A Real Example	69
	Problems	73
	References	74

## Part II Understanding Big Data Systems

<b>4</b>	<b>Distributed File System</b>	<b>79</b>
4.1	Hadoop Framework	79
4.1.1	Hadoop Distributed File System	80
4.1.2	MapReduce Programming Model	81
4.2	Hadoop System	81
4.2.1	Operating System	82
4.2.2	Distributed System	82
4.2.3	Programming Platform	83
4.3	Hadoop Environment	83
4.3.1	Essential Tools	84
4.3.2	Installation Guidance	85
4.3.3	RStudio Server	93
4.4	Testing the Hadoop Environment	94
4.4.1	Standard Example	94
4.4.2	Alternative Example	95

- 4.5 Multinode Hadoop ..... 95
  - 4.5.1 Virtual Network ..... 96
  - 4.5.2 Hadoop Setup ..... 96
- Problems ..... 97
- References ..... 97
- 5 MapReduce Programming Platform** ..... 99
  - 5.1 MapReduce Framework ..... 99
    - 5.1.1 Parametrization ..... 100
    - 5.1.2 Parallelization ..... 101
  - 5.2 MapReduce Essentials ..... 102
    - 5.2.1 Mapper Function ..... 102
    - 5.2.2 Reducer Function ..... 103
    - 5.2.3 MapReduce Function ..... 104
    - 5.2.4 A Coding Example ..... 104
  - 5.3 MapReduce Programming ..... 107
    - 5.3.1 Naming Convention ..... 107
    - 5.3.2 Coding Principles ..... 108
    - 5.3.3 Application of Coding Principles ..... 110
  - 5.4 File Handling in MapReduce ..... 113
    - 5.4.1 Pythagorean Numbers ..... 114
    - 5.4.2 File Split Example ..... 115
    - 5.4.3 File Split Improved ..... 116
- Problems ..... 118
- References ..... 118

**Part III Understanding Machine Learning**

- 6 Modeling and Algorithms** ..... 123
  - 6.1 Machine Learning ..... 123
    - 6.1.1 A Simple Example ..... 124
    - 6.1.2 Domain Division Perspective ..... 125
    - 6.1.3 Data Domain ..... 128
    - 6.1.4 Domain Division ..... 129
  - 6.2 Learning Models ..... 130
    - 6.2.1 Mathematical Models ..... 132
    - 6.2.2 Hierarchical Models ..... 134
    - 6.2.3 Layered Models ..... 135
    - 6.2.4 Comparison of the Models ..... 135
  - 6.3 Learning Algorithms ..... 140
    - 6.3.1 Supervised Learning ..... 140
    - 6.3.2 Types of Learning ..... 141
- Problems ..... 142
- References ..... 142

<b>7</b>	<b>Supervised Learning Models</b>	145
7.1	Supervised Learning Objectives	145
7.1.1	Parametrization Objectives	146
7.1.2	Optimization Objectives	148
7.2	Regression Models	150
7.2.1	Continuous Response	151
7.2.2	Theory of Regression Models	151
7.3	Classification Models	160
7.3.1	Discrete Response	160
7.3.2	Mathematical Models	162
7.4	Hierarchical Models	166
7.4.1	Decision Tree	167
7.4.2	Random Forest	167
7.5	Layered Models	170
7.5.1	Shallow Learning	171
7.5.2	Deep Learning	177
	Problems	179
	References	180
<b>8</b>	<b>Supervised Learning Algorithms</b>	183
8.1	Supervised Learning	183
8.1.1	Learning	185
8.1.2	Training	186
8.1.3	Testing	188
8.1.4	Validation	190
8.2	Cross-Validation	192
8.2.1	Tenfold Cross-Validation	193
8.2.2	Leave-One-Out	193
8.2.3	Leave-p-Out	194
8.2.4	Random Subsampling	195
8.2.5	Dividing Data Sets	195
8.3	Measures	196
8.3.1	Quantitative Measure	197
8.3.2	Qualitative Measure	198
8.4	A Simple 2D Example	202
	Problems	204
	References	205
<b>9</b>	<b>Support Vector Machine</b>	207
9.1	Linear Support Vector Machine	207
9.1.1	Linear Classifier: Separable Linearly	208
9.1.2	Linear Classifier: Nonseparable Linearly	218
9.2	Lagrangian Support Vector Machine	219
9.2.1	Modeling of LSVM	219
9.2.2	Conceptualized Example	219
9.2.3	Algorithm and Coding of LSVM	220

- 9.3 Nonlinear Support Vector Machine . . . . . 223
  - 9.3.1 Feature Space . . . . . 224
  - 9.3.2 Kernel Trick . . . . . 224
  - 9.3.3 SVM Algorithms on Hadoop . . . . . 227
  - 9.3.4 Real Application . . . . . 233
- Problems . . . . . 234
- References . . . . . 235

**10 Decision Tree Learning . . . . . 237**

- 10.1 The Decision Tree . . . . . 237
  - 10.1.1 A Coding Example—Classification Tree . . . . . 241
  - 10.1.2 A Coding Example—Regression Tree . . . . . 244
- 10.2 Types of Decision Trees . . . . . 245
  - 10.2.1 Classification Tree . . . . . 246
  - 10.2.2 Regression Tree . . . . . 247
- 10.3 Decision Tree Learning Model . . . . . 248
  - 10.3.1 Parametrization . . . . . 248
  - 10.3.2 Optimization . . . . . 249
- 10.4 Quantitative Measures . . . . . 250
  - 10.4.1 Entropy and Cross-Entropy . . . . . 250
  - 10.4.2 Gini Impurity . . . . . 252
  - 10.4.3 Information Gain . . . . . 255
- 10.5 Decision Tree Learning Algorithm . . . . . 256
  - 10.5.1 Training Algorithm . . . . . 257
  - 10.5.2 Validation Algorithm . . . . . 263
  - 10.5.3 Testing Algorithm . . . . . 263
- 10.6 Decision Tree and Big Data . . . . . 266
  - 10.6.1 Toy Example . . . . . 266
- Problems . . . . . 268
- References . . . . . 269

**Part IV Understanding Scaling-Up Machine Learning**

**11 Random Forest Learning . . . . . 273**

- 11.1 The Random Forest . . . . . 273
  - 11.1.1 Parallel Structure . . . . . 274
  - 11.1.2 Model Parameters . . . . . 275
  - 11.1.3 Gain/Loss Function . . . . . 276
  - 11.1.4 Bootstrapping and Bagging . . . . . 276
- 11.2 Random Forest Learning Model . . . . . 278
  - 11.2.1 Parametrization . . . . . 279
  - 11.2.2 Optimization . . . . . 279
- 11.3 Random Forest Learning Algorithm . . . . . 279
  - 11.3.1 Training Algorithm . . . . . 280
  - 11.3.2 Testing Algorithm . . . . . 283

- 11.4 Random Forest and Big Data . . . . . 284
  - 11.4.1 Random Forest Scalability . . . . . 284
  - 11.4.2 Big Data Classification . . . . . 284
- Problems . . . . . 287
- References . . . . . 288
  
- 12 Deep Learning Models . . . . . 289**
  - 12.1 Introduction . . . . . 289
  - 12.2 Deep Learning Techniques . . . . . 291
    - 12.2.1 No-Drop Deep Learning . . . . . 291
    - 12.2.2 Dropout Deep Learning . . . . . 291
    - 12.2.3 Dropconnect Deep Learning . . . . . 292
    - 12.2.4 Gradient Descent . . . . . 293
    - 12.2.5 A Simple Example . . . . . 297
    - 12.2.6 MapReduce Implementation . . . . . 298
  - 12.3 Proposed Framework . . . . . 301
    - 12.3.1 Motivation . . . . . 301
    - 12.3.2 Parameters Mapper . . . . . 301
  - 12.4 Implementation of Deep Learning . . . . . 303
    - 12.4.1 Analysis of Domain Divisions . . . . . 303
    - 12.4.2 Analysis of Classification Accuracies . . . . . 303
  - 12.5 Ensemble Approach . . . . . 305
  - Problems . . . . . 306
  - References . . . . . 306
  
- 13 Chandelier Decision Tree . . . . . 309**
  - 13.1 Unit Circle Algorithm . . . . . 309
    - 13.1.1 UCA Classification . . . . . 310
    - 13.1.2 Improved UCA Classification . . . . . 311
    - 13.1.3 A Coding Example . . . . . 312
    - 13.1.4 Drawbacks of UCA . . . . . 315
  - 13.2 Unit Circle Machine . . . . . 315
    - 13.2.1 UCM Classification . . . . . 315
    - 13.2.2 A Coding Example . . . . . 316
    - 13.2.3 Drawbacks of UCM . . . . . 318
  - 13.3 Unit Ring Algorithm . . . . . 318
    - 13.3.1 A Coding Example . . . . . 319
    - 13.3.2 Unit Ring Machine . . . . . 321
    - 13.3.3 A Coding Example . . . . . 321
    - 13.3.4 Drawbacks of URM . . . . . 323
  - 13.4 Chandelier Decision Tree . . . . . 323
    - 13.4.1 CDT-Based Classification . . . . . 324
    - 13.4.2 Extension to Random Chandelier . . . . . 328
  - Problems . . . . . 328
  - References . . . . . 328

- 14 Dimensionality Reduction** ..... 329
  - 14.1 Introduction ..... 329
  - 14.2 Feature Hashing Techniques ..... 330
    - 14.2.1 Standard Feature Hashing ..... 331
    - 14.2.2 Flagged Feature Hashing ..... 331
  - 14.3 Proposed Feature Hashing ..... 332
    - 14.3.1 Binning and Mitigation ..... 332
    - 14.3.2 Mitigation Justification ..... 333
    - 14.3.3 Toy Example ..... 333
  - 14.4 Simulation and Results ..... 334
    - 14.4.1 A Matlab Implementation ..... 334
    - 14.4.2 A MapReduce Implementation ..... 337
  - 14.5 Principal Component Analysis ..... 340
    - 14.5.1 Eigenvector ..... 341
    - 14.5.2 Principal Components ..... 343
    - 14.5.3 The Principal Directions ..... 346
    - 14.5.4 A 2D Implementation ..... 348
    - 14.5.5 A 3D Implementation ..... 350
    - 14.5.6 A Generalized Implementation ..... 352
  - Problems ..... 354
  - References ..... 354
  
- Index** ..... 357



<http://www.springer.com/978-1-4899-7640-6>

Machine Learning Models and Algorithms for Big Data  
Classification

Thinking with Examples for Effective Learning

Suthaharan, S.

2016, XIX, 359 p. 149 illus., 82 illus. in color., Hardcover

ISBN: 978-1-4899-7640-6