

# COMPUTATIONAL BUSINESS ANALYTICS

SUBRATA DAS

Machine Analytics, Inc.  
Belmont, Massachusetts, USA



CRC Press

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an Informa business

A CHAPMAN & HALL BOOK

---

# Contents

---

CHAPTER 1 ■ Analytics Background and Architectures	1
1.1 ANALYTICS DEFINED	1
1.2 ANALYTICS MODELING	5
1.3 ANALYTICS PROCESSES	8
1.3.1 Information Hierarchy	8
1.3.2 Information Processing Hierarchy	9
1.3.3 Human Information Processing Hierarchy	10
1.4 ANALYTICS AND DATA FUSION	11
1.4.1 JDL Fusion Model	12
1.4.2 OODA Loop	14
1.5 FURTHER READING	15
CHAPTER 2 ■ Mathematical and Statistical Preliminaries	17
2.1 STATISTICS AND PROBABILITY THEORY	17
2.2 LINEAR ALGEBRA FUNDAMENTALS	22
2.3 MATHEMATICAL LOGIC	25
2.4 GRAPHS AND TREES	32
2.5 MEASURES OF PERFORMANCE	36
2.6 ALGORITHMIC COMPLEXITY	37
2.7 FURTHER READING	41
CHAPTER 3 ■ Statistics for Descriptive Analytics	43
3.1 PROBABILITY DISTRIBUTIONS	43
3.2 DISCRETE PROBABILITY DISTRIBUTIONS	46
3.2.1 Binomial and Multinomial Distributions	47
3.2.2 Poisson Distribution and Process	48

3.3	CONTINUOUS PROBABILITY DISTRIBUTIONS	49
3.3.1	Gaussian or Normal Distribution	49
3.3.2	Lognormal	50
3.3.3	Exponential Distribution	51
3.3.4	Weibull Distribution	52
3.3.5	Beta and Dirichlet Distributions	53
3.3.6	Gamma Distribution	56
3.4	GOODNESS-OF-FIT TEST	57
3.4.1	Probability Plot	57
3.4.2	One-Way Chi-Square Goodness-of-Fit Test	59
3.4.3	Kolmogorov-Smirnov Test	61
3.5	FURTHER READING	64
<hr/> CHAPTER 4 ■ Bayesian Probability and Inference		65
4.1	BAYESIAN INFERENCE	65
4.2	PRIOR PROBABILITIES	68
4.2.1	Conjugate Priors	69
4.2.2	The Jeffreys Prior	70
4.3	FURTHER READING	73
<hr/> CHAPTER 5 ■ Inferential Statistics and Predictive Analytics		75
5.1	CHI-SQUARE TEST OF INDEPENDENCE	76
5.2	REGRESSION ANALYSES	77
5.2.1	Simple Linear Regression	77
5.2.2	Multiple Linear Regression	78
5.2.3	Logistic Regression	79
5.2.4	Polynomial Regression	81
5.3	BAYESIAN LINEAR REGRESSION	82
5.3.1	Gaussian Processes	84
5.4	PRINCIPAL COMPONENT AND FACTOR ANALYSES	87
5.5	SURVIVAL ANALYSIS	92
5.6	AUTOREGRESSION MODELS	97
5.7	FURTHER READING	98

**CHAPTER 6 ■ Artificial Intelligence for Symbolic Analytics 99**

---

6.1	ANALYTICS AND UNCERTAINTIES	99
6.1.1	Ignorance to Uncertainties	99
6.1.2	Approaches to Handling Uncertainties	103
6.2	NEO-LOGICIST APPROACH	105
6.2.1	Evolution of Rules	106
6.2.2	Inferencing in Rule-based Systems	110
6.2.3	Advantages and Disadvantages of Rule-Based Systems	111
6.3	NEO-PROBABILIST	112
6.4	NEO-CALCULIST APPROACH	114
6.4.1	Certainty Factors	114
6.4.2	Dempster-Shafer Theory of Belief Function	117
6.5	NEO-GRANULARIST	123
6.5.1	Probabilistic Logic	123
6.5.2	Fuzzy Logic	126
6.5.3	Fuzzy Logic for Customer Segmentation	132
6.6	FURTHER READING	134

**CHAPTER 7 ■ Probabilistic Graphical Modeling 135**

---

7.1	NAIVE BAYESIAN CLASSIFIER (NBC)	136
7.2	K-DEPENDENCE NAIVE BAYESIAN CLASSIFIER (KNBC)	138
7.3	BAYESIAN BELIEF NETWORKS	140
7.3.1	Conditional Independence in Belief Networks	145
7.3.2	Evidence, Belief, and Likelihood	152
7.3.3	Prior Probabilities in Networks without Evidence	154
7.3.4	Belief Revision	156
7.3.5	Evidence Propagation in Polytrees	161
	7.3.5.1 <i>Upward Propagation in a Linear Fragment</i>	161
	7.3.5.2 <i>Downward Propagation in a Linear Fragment</i>	164
	7.3.5.3 <i>Upward Propagation in a Tree Fragment</i>	167

7.3.5.4	<i>Downward Propagation in a Tree Fragment</i>	169
7.3.5.5	<i>Upward Propagation in a Polytree Fragment</i>	169
7.3.5.6	<i>Downward Propagation in a Polytree Fragment</i>	171
7.3.6	Propagation Algorithm	175
7.3.7	Evidence Propagation in Directed Acyclic Graphs	178
7.3.7.1	<i>Graphical Transformation</i>	181
7.3.7.2	<i>Join Tree Initialization</i>	187
7.3.7.3	<i>Propagation in Join Tree and Marginalization</i>	189
7.3.7.4	<i>Handling Evidence</i>	191
7.3.8	Complexity of Inference Algorithms	194
7.3.9	Acquisition of Probabilities	195
7.3.10	Advantages and Disadvantages of Belief Networks	198
7.3.11	Belief Network Tools	199
7.4	FURTHER READING	199
<b>CHAPTER 8 ■ Decision Support and Prescriptive Analytics</b>		<b>201</b>
8.1	EXPECTED UTILITY THEORY AND DECISION TREES	202
8.2	INFLUENCE DIAGRAMS FOR DECISION SUPPORT	204
8.2.1	Inferencing in Influence Diagrams	206
8.2.2	Compilation of Influence Diagrams	211
8.3	SYMBOLIC ARGUMENTATION FOR DECISION SUPPORT	219
8.3.1	Measuring Consensus	221
8.3.2	Combining Sources of Varying Confidence	226
8.4	FURTHER READING	226
<b>CHAPTER 9 ■ Time Series Modeling and Forecasting</b>		<b>229</b>
9.1	PROBLEM MODELING	229
9.1.1	State Transition and Observation Models	230
9.1.2	Estimation Problem	231
9.2	KALMAN FILTER (KF)	233

9.2.1	Extended Kalman Filter (EKF)	240
9.3	MARKOV MODELS	242
9.3.1	Hidden Markov Models (HMM)	244
9.3.2	The Forward Algorithm	248
9.3.3	The Viterbi Algorithm	252
9.3.4	Baum-Welch Algorithm for Learning HMM	253
9.4	DYNAMIC BAYESIAN NETWORKS (DBNS)	257
9.4.1	Inference Algorithms for DBNs	260
9.5	FURTHER READING	265
<hr/>		
CHAPTER 10	Monte Carlo Simulation	267
10.1	MONTE CARLO APPROXIMATION	267
10.2	GIBBS SAMPLING	270
10.3	METROPOLIS-HASTINGS ALGORITHM	272
10.4	PARTICLE FILTER (PF)	273
10.4.1	Particle Filter for Dynamical Systems	274
10.4.2	Particle Filter for DBN	277
10.4.3	Particle Filter Issues	279
10.5	FURTHER READING	280
<hr/>		
CHAPTER 11	Cluster Analysis and Segmentation	281
11.1	HIERARCHICAL CLUSTERING	282
11.2	K-MEANS CLUSTERING	284
11.3	K-NEAREST NEIGHBORS	286
11.4	SUPPORT VECTOR MACHINES	288
11.4.1	Linearly Separable Data	288
11.4.2	Preparation of Data and Packages	291
11.4.3	Non-Separable Data	291
11.4.4	Non-Linear Classifier	293
11.4.5	VC Dimension and Maximum Margin Classifier	296
11.5	NEURAL NETWORKS	298
11.5.1	Model Building and Data Preparation	300
11.5.2	Gradient Descent for Updating Weights	301
11.6	FURTHER READING	302

CHAPTER 12 ■ Machine Learning for Analytics Models	303
12.1 DECISION TREES	304
12.1.1 Algorithms for Constructing Decision Trees	305
12.1.2 Overfitting in Decision Trees	311
12.1.3 Handling Continuous Attributes	313
12.1.4 Advantages and Disadvantages of Decision Tree Techniques	315
12.2 LEARNING NAIVE BAYESIAN CLASSIFIERS	315
12.2.1 Semi-Supervised Learning of NBC via EM	318
12.3 LEARNING OF KNBC	322
12.4 LEARNING OF BAYESIAN BELIEF NETWORKS	323
12.4.1 Cases for Learning Bayesian Networks	324
12.4.2 Learning Probabilities	325
12.4.2.1 <i>Brief Survey</i>	325
12.4.2.2 <i>Learning Probabilities from Fully Observable Variables</i>	325
12.4.2.3 <i>Learning Probabilities from Partially Observable Variables</i>	327
12.4.2.4 <i>Online Adjustment of Parameters</i>	331
12.4.3 Structure Learning	332
12.4.3.1 <i>Brief Survey</i>	332
12.4.3.2 <i>Learning Structure from Fully Observable Variables</i>	333
12.4.3.3 <i>Learning Structure from Partially Observable Variables</i>	338
12.4.4 Use of Prior Knowledge from Experts	339
12.5 INDUCTIVE LOGIC PROGRAMMING	339
12.6 FURTHER READING	343
CHAPTER 13 ■ Unstructured Data and Text Analytics	345
13.1 INFORMATION STRUCTURING AND EXTRACTION	346
13.2 BRIEF INTRODUCTION TO NLP	348
13.2.1 Syntactic Analysis	349
13.2.1.1 <i>Tokenization</i>	349
13.2.1.2 <i>Morphological Analysis</i>	349

13.2.1.3	<i>Part-of-Speech (POS) Tagging</i>	350
13.2.1.4	<i>Syntactic Parsing</i>	351
13.2.2	Semantic Analysis	354
13.2.2.1	<i>Named Entity Recognition</i>	354
13.2.2.2	<i>Co-reference Resolution</i>	354
13.2.2.3	<i>Relation Extraction</i>	355
13.3	TEXT CLASSIFICATION AND TOPIC EXTRACTION	355
13.3.1	Naïve Bayesian Classifiers (NBC)	356
13.3.2	k-Dependence Naïve Bayesian Classifier (kNBC)	359
13.3.3	Latent Semantic Analysis	361
13.3.4	Probabilistic Latent Semantic Analysis (PLSA)	368
13.3.5	Latent Dirichlet Allocation (LDA)	369
13.4	FURTHER READING	372
<b>CHAPTER 14 ■ Semantic Web</b>		<b>373</b>
<hr/>		
14.1	RESOURCE DESCRIPTION FRAMEWORK (RDF)	373
14.1.1	RDF Schema (RDFS)	377
14.1.2	Ontology Web Language (OWL)	379
14.2	DESCRIPTION LOGICS	381
14.2.1	Description Logic Syntax	382
14.2.2	Description Logic Axioms	384
14.2.3	Description Logic Constructs and Subsystems	384
14.2.4	Description Logic and OWL Constructs in Relational Database	386
14.2.5	Description Logic as First-Order Logic	387
14.3	FURTHER READING	388
<b>CHAPTER 15 ■ Analytics Tools</b>		<b>389</b>
<hr/>		
15.1	INTELLIGENT DECISION AIDING SYSTEM (IDAS)	390
15.2	ENVIRONMENT FOR 5TH GENERATION APPLICATIONS (E5)	400
15.2.1	Rule-based Expert System Shell	401
15.2.2	Prolog Interpreter	404
15.2.3	Lisp Interpreter	405



15.3	ANALYSIS OF TEXT (ATEXT)	406
15.4	R AND MATLAB	419
15.5	SAS AND WEKA	421
<hr/>		
CHAPTER 16 ■ Analytics Case Studies		425
<hr/>		
16.1	RISK ASSESSMENT MODEL I3	425
16.2	RISK ASSESSMENT IN INDIVIDUAL LENDING USING IDAS	427
16.3	RISK ASSESSMENT IN COMMERCIAL LENDING USING E5 AND IDAS	430
16.4	FRAUD DETECTION	441
16.5	SENTIMENT ANALYSIS USING ATEXT	444
16.5.1	Text Corpus Classification	444
16.5.2	Evaluation Results	446
16.6	LIFE STATUS ESTIMATION USING DYNAMIC BAYESIAN NETWORKS	449
<hr/>		
APPENDIX A ■ Usage of Symbols		453
<hr/>		
A.1	SYMBOLS USED IN THE BOOK	453
<hr/>		
APPENDIX B ■ Examples and Sample Data		455
<hr/>		
B.1	PLAY-TENNIS EXAMPLE	455
B.2	UNITED STATES ELECTORAL COLLEGE DATA	456
<hr/>		
APPENDIX C ■ MATLAB and R Code Examples		457
<hr/>		
C.1	MATLAB CODE FOR STOCK PREDICTION USING KALMAN FILTER	457
C.2	R CODE FOR STOCK PREDICTION USING KALMAN FILTER	460
<hr/>		
Index		479
<hr/>		