SUPERVISED MACHINE LEARNING

OPTIMIZATION FRAMEWORK AND APPLICATIONS WITH SAS AND R

Tanya Kolosova and Samuel Berestizhevsky

CRC Press
Taylor & Francis Group

p65
6,800-

สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

616586682
012538590
122568475

Supervised Machine Learning

Optimization Framework and Applications with SAS and R

สำนักหอสมุดมหาวิทยาลัยเชียงใหม่
CHIANG MAI UNIVERSITY LIBRARY

Tanya Kolosova PhD
Associates in Analytics Inc., Boca Raton, Florida

Samuel Berestizhevsky MSc
Associates in Analytics Inc., Boca Raton, Florida

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

Acknowledgments xiii
Authors xv
Introduction: Challenges in the Application of Machine Learning Classification Methods xvii

Part I

1. Introduction to the AI Framework 3
1.1 Components of the AI Framework and Their Interaction 3
1.2 AI Framework in Detail 5
1.2.1 Creating Training and Test Datasets 5
1.2.2 Design of Experiments for a Classifier 6
1.2.3 Firth Logistic Regression 6
1.2.4 Data Contamination 6
1.2.5 Best Classifiers 7
1.3 SAS Procedures for the AI Framework Components 7
1.4 R Libraries for the AI Framework Components 7
References 8

2. Supervised Machine Learning and Its Deployment in SAS and R 9
2.1 Introduction 9
2.2 Principles of Supervised Machine Learning 10
2.3 Neural Network 12
2.3.1 Introduction 12
2.3.2 Neural Network Components 13
2.3.2.1 Activation Function 13
2.3.2.2 Neurons 14
2.3.2.3 Networks 15
2.3.3 R for Neural Networks 16
2.4 Support Vector Machine 16
2.4.1 Introduction 16
2.4.2 Kernel 18
2.4.3 Margin 19
2.4.4 Optimization 20
2.4.5 Bias–Variance Trade-off and SVM Hyperparameters 20
2.4.6 R for SVM 21
2.5 SVM Modification Using Firth's Regression 22
2.5.1 Introduction 22
2.5.2 Logistic Regression 23
2.5.3 Problem of Separation 23

2.5.4 R for Firth's Regression 25
2.5.5 SAS for Firth's Regression 25
2.6 Summary 27
References 27

3. Bootstrap Methods and Their Deployment in SAS and R 29
3.1 Introduction 29
3.2 Overview of Bootstrap Methods 30
3.2.1 The Basic Bootstrap 31
3.2.2 Hypothesis Tests, Estimates, and Confidence Intervals 32
3.2.3 Bias Reduction 34
3.2.4 The Parametric Bootstrap 35
3.2.5 m-out-of-n Bootstrap 36
3.2.6 Bootstrap Samples Similarity 36
3.3 Implementation of Bootstrap in SAS and R 37
3.3.1 m-out-of-n in SAS 37
3.3.2 m-out-of-n in R 38
3.4 Summary 39
References 40

4. Outliers Detection and Its Deployment in SAS and R 41
4.1 Introduction 41
4.2 Outliers Detection and Treatment 42
4.2.1 Minimum Covariance Determinant Method 42
4.2.2 MCD in SAS 43
4.3 Bias Reduction 44
4.4 Summary 45
References 45

5. Design of Experiments and Its Deployment in SAS and R 47
5.1 Introduction 47
5.2 Application of DoE in AI Framework 48
5.2.1 Terminology of DoE 49
5.2.1.1 Experiment 49
5.2.1.2 Experimental Unit 49
5.2.1.3 Factor 49
5.2.1.4 Treatment 49
5.2.2 Principles of DoE 49
5.2.2.1 Randomization 49
5.2.2.2 Statistical Replication 50
5.2.2.3 Blocking 50
5.2.2.4 Orthogonality 50
5.2.3 Full-Factorial Experiment 50
5.2.4 Fractional Factorial Experiment 57
5.2.5 Linear Mixed Models 58

5.2.6 Factors and Response Variables in the AI Framework.....59
5.2.7 Example.....60
5.2.8 Analysis of Linear Mixed Model Using SAS.....62
5.2.9 Analysis of Linear Mixed Model Using R.....65
5.3 Summary.....66
References.....67

Part II

6. Introduction to the SAS- and R-Based Table-Driven Environment....71
6.1 Principles of Code-Free Design.....71
6.2 The Data Dictionary Components for the AI Framework.....72
6.2.1 Relational Model.....72
6.2.2 Table.....73
6.2.3 Data Aspects.....73
6.2.4 Relational Data Structure.....73
6.2.5 Domains.....74
6.2.6 Relations and Tables.....74
6.2.7 Functions.....74
6.2.8 One-to-one Relationship.....75
6.2.9 One-to-many Relationship.....75
6.2.10 Primary Key.....75
6.2.11 Foreign Key.....75
6.2.12 Missing Values.....75
6.2.13 Data Dictionary.....76
6.3 Properties of the Data Dictionary.....76
6.3.1 The Library Table.....76
6.3.2 The Object Table.....77
6.3.3 The Location Table.....77
6.3.4 The Message Table.....77
6.3.5 The Property Table.....78
6.3.6 Meaning.....79
6.3.7 The Link Table.....79
6.3.8 Process of Application Data Model Definition.....79
6.3.9 Features of the Data Dictionary.....80
6.3.10 The Components of the Optimization Framework and Their Definitions in the Data Dictionary.....81
6.4 Deployment of Code-Free Design with SAS and R.....81
6.4.1 How to Generate Application Objects.....81
6.4.2 Generating R Datasets from the Data Dictionary Metadata.....84
6.4.3 SAS and R Interoperability.....86
6.5 Summary.....88
Reference.....88

7. Input Data Component 89
7.1 Overview of Data Management 89
7.1.1 Data Dictionary 89
7.1.1.1 The Input Data Dictionary 89
7.1.1.2 Input and Structure Tables 90
7.1.1.3 Outlier_Detection and Bias_Correction Tables... 91
7.1.1.4 Bootstrap Table 92
7.1.1.5 Output Table 93
7.1.2 SAS Macro Program 95
7.1.3 R Program 98
7.2 Summary 98

8. Design of Experiment for Machine Learning Component 99
8.1 Data Dictionary 99
8.1.1 Experiment Table 100
8.1.2 Features Table 100
8.1.3 Metrics Table 101
8.1.4 ML_Method Table 102
8.1.5 Hyperparameters_Domain Table 102
8.1.6 Results Table 102
8.1.7 Results_Metrics Table 103
8.2 SAS Macro Program 105
8.3 R Programs 105
8.4 Summary 106
Reference 106

9. "Contaminated" Training Datasets Component 107
9.1 Data Dictionary 107
9.1.1 Contamination Table 108
9.1.2 Cont_Experiment Table 109
9.1.3 Cont_Results Table 109
9.1.4 Cont_Metric Table 110
9.2 SAS Macro Program 110
9.3 R Programs 110
9.4 Summary 111
Reference 111

Part III

10. Insurance Industry: Underwriters' Decision-Making Process 115
10.1 Introduction 115
10.2 Review of Underwriters' Performance 116
10.2.1 Metrics of Underwriters' Performance 116
10.2.1.1 Hit Ratio 116

10.2.1.2 Conversion Rate 116
10.2.1.3 Dynamic Conversion Rate 117
10.2.1.4 Time-to-Deal 118
10.2.2 Analysis of Underwriters' Performance 119
10.2.2.1 Data Description 119
10.2.2.2 Application Flow 119
10.2.2.3 Dynamic Conversion Rate per Underwriter 121
10.2.2.4 Time-to-Deal per Underwriter 122
10.3 Traditional Approach to Knowledge Delivery 123
10.4 Anatomy of Artificial Intelligence Solution 124
10.4.1 Data Structure 124
10.4.2 Classification Approach 125
10.4.3 Bias–Variance Trade-Off and SVM Hyperparameters 125
10.4.4 Building the Classifier 127
10.4.5 "Contamination" of Training Datasets 130
10.4.6 Experimental Results 130
10.5 Summary 132
References 132

11. Insurance Industry: Claims Modeling and Prediction 135
11.1 Introduction 135
11.2 Data 136
11.3 The Cox Model for Claims Event Analysis 136
11.4 Application of the Cox Model for Claims Analysis 138
11.4.1 Data Transformation 139
11.4.2 Cox Model Assumption Validation 141
11.4.3 Bayesian Machine Learning Approach 144
11.4.4 Deployment with SAS 144
11.4.5 Interpretation of Results 146
11.5 Summary 152
References 153

Index 155