Contents at a Glance

CONTENTS AT A GLANCE 5

CHAPTER 1 1

INTRODUCTION TO MACHINE LEARNING 1

1.1 OVERVIEW OF MACHINE LEARNING 1

1.2 IMPORTANCE OF MACHINE LEARNING IN MODERN APPLICATIONS 3

1.3 ESSENTIAL MACHINE LEARNING TERMINOLOGIES 5

1.4 STAGES IN THE MACHINE LEARNING PIPELINE 6

1.5 SCOPE AND REAL-WORLD APPLICATIONS 8

1.6 LIMITATIONS AND CHALLENGES IN MACHINE LEARNING 10

1.7 REGRESSION METHODS IN MACHINE LEARNING 11

1.7.1 Concept of Regression 11

1.7.2 Classification of Regression Techniques 12

1.7.3 Linear Regression Model 13

1.7.4 Logistic Regression Model 13

1.8 MODEL SELECTION CRITERIA FOR LINEAR REGRESSION 14

1.8.1 Practical Uses of Linear Regression 15

1.8.2 Algorithmic Implementation 16

1.9 MATHEMATICAL DERIVATION OF LINEAR REGRESSION 22

1.10 MEAN SQUARED ERROR ANALYSIS 26

1.10.1 R-Squared Metric 28

EXERCISE 32

REFERENCE 34

CHAPTER 2 35

PYTHON FOR DATA SCIENCE 35

2.1 INTRODUCTION TO PYTHON FOR DATA SCIENCE 35

2.2. SETTING UP THE PYTHON ENVIRONMENT 35

2.2.1 Installing Python 35

2.2.2 Executing Python via Command Line *35*

2.2.3 Working with Integrated Development Environments (IDEs) *36*

2.3 CONTROL FLOW AND PROGRAMMING CONSTRUCTS 38

2.4 FUNDAMENTALS OF STATISTICAL ANALYSIS 41

2.4.1 Data Import Techniques *44*

2.4.2 Data Visualization Methods *45*

2.4.3 Correlation Analysis *46*

2.4.4 Simple Linear Regression *46*

2.4.5 Multiple Linear Regression Models *47*

2.5 INTRODUCTION TO MACHINE LEARNING CONCEPTS 48

2.5.1 Classification Techniques (Supervised Learning) *50*

2.5.2 Clustering Methods (Unsupervised Learning) *50*

2.5.3 Density Estimation Approaches (Unsupervised Learning) *51*

EXERCISE 52

REFERENCE 54

CHAPTER 3 ..... **55**

DATA PREPROCESSING **55**

3.1 PURPOSE OF DATA ANALYTICS IN DATA CLEANING 55

3.2 TOOLS FOR DATA ANALYTICS IN DATA CLEANING 55

3.3 LEVELS OF DATA CLEANING 56

3.4 MAPPING THE PURPOSES AND TOOLS OF ANALYTICS TO THE LEVELS OF DATA CLEANING 58

3.5 DATA CLEANING LEVEL I – CLEANING UP THE TABLE 58

3.5.1 Example 3.1 - Unwise Data Collection *59*

3.6 DATA CLEANING LEVEL II 63

3.6.1 Example 3.2 - Unpacking Columns and Reformulating the Table *63*

3.6.2 Example 3.3 – Restructuring the Table *74*

3.6.3 Example 3.4 - Level I And II Data Cleaning *78*

3.7 DATA CLEANING LEVEL III 91

3.8 MISSING VALUES 92

3.8.1 Detecting Missing Values *94*

3.8.2 Causes of Missing Values *96*

3.8.3 Types of Missing Values *97*

3.8.4 Diagnosis of Missing Values *98*

3.8.5 Dealing with Missing Values *116*

3.9 OUTLIERS 118

3.9.1 Detecting Outliers *119*

3.9.2 Dealing with Outliers *120*

EXERCISE 122

REFERENCE 124

CHAPTER 4 ..... **125**

EXPLORATORY DATA ANALYSIS (EDA) **125**

4.1 INTRODUCTION TO EXPLORATORY DATA ANALYSIS 125

4.2 TECHNIQUES FOR DATA VISUALIZATION 127

4.3 DESCRIPTIVE AND SUMMARY STATISTICS 133

4.3.1 Key Summary Measures *134*

4.4 DETECTION OF PATTERNS AND OUTLIERS IN DATA 135

4.4.1 Pattern Recognition Methods *135*

4.4.2 Outher Detection Approaches *137*

4.4.3 Analytical Tools for Pattern and Outlier Identification *137*

EXERCISE 138

REFERENCES 140

CHAPTER 5 ..... **141**

SUPERVISED LEARNING – REGRESSION **141**

5.1 INTRODUCTION 141

5.2 SUPERVISED LEARNING - REAL-LIFE SCENARIO 142

5.3 TYPES OF SUPERVISED LEARNING 143

5.3.1 SUPERVISED LEARNING - CLASSIFICATION *143*

5.3.2 SUPERVISED LEARNING – REGRESSION *145*

5.3.3 CLASSIFICATION VS REGRESSION *146*

5.3.4 Conversion Between Classification and Regression Problems *146*

5.4 LINEAR REGRESSION 147

5.4.1 Types of Linear Regression *149*

5.4.2 Geometric Intuition *154*

5.4.3 Mathematical Formulation *156*

5.4.4 Solving Optimization Problem *157*

5.4.5 Real-World Applications *164*

5.4.6 LINEAR REGRESSION IN PRACTICE USING PYTHON *169*

EXERCISE 179

REFERENCES 181

CHAPTER 6 182

SUPERVISED LEARNING – CLASSIFICATION 182

6.1 INTRODUCTION 182

6.2 TYPES OF SUPERVISED LEARNING PROBLEMS 188

6.3 CLASSIFICATION METHODS 189

6.4 REGRESSION METHODS 191

6.5 CLASSIFICATION MODEL 194

6.6 CLASSIFICATION LEARNING STEPS 196

EXERCISE 199

REFERENCE 201

CHAPTER 7 202

UNSUPERVISED LEARNING 202

7.1 INTRODUCTION 202

7.1.1 Density Estimation *202*

7.2 CLUSTERING 205

7.2.1 K-Means *205*

7.2.2 DBSCAN and HDBSCAN *207*

7.3 DETERMINING THE NUMBER OF CLUSTERS 208

7.3.1 Other Clustering Algorithms *210*

7.4 DIMENSIONALITY REDUCTION 214

7.5 PRINCIPAL COMPONENT ANALYSIS 215

7.5.1 UMAP *216*

7.6 OTHER FORMS OF LEARNING 219

7.6.1 Learning to Rank *221*

7.6.2 Learning to Recommend *224*

7.6.3 Factorization Machines *225*

7.6.4 Denoising Autoencoders *227*

7.6.5 Self-Supervised Learning: Word Embeddings *228*

7.7 PROBABILISTIC GRAPHICAL MODELS 230

7.7.1 Markov Chain Monte Carlo *231*

7.7.2 Genetic Algorithms *232*

7.8 REINFORCEMENT LEARNING 232

EXERCISE 234

REFERENCES 236

CHAPTER 8 ..... **237**

MODEL EVALUATION AND OPTIMIZATION **237**

8.1 EVALUATING REGRESSION MODELS 237

8.2 RMSE METRIC 240

8.3 R^2 METRIC 241

8.4 EVALUATING CLASSIFICATION MODELS 242

8.4.1 Binary Classification Models *242*

8.5 MULTI-CLASS CLASSIFICATION MODELS 254

8.6 CHOOSING HYPERPARAMETER VALUES 260

EXERCISE 263

REFERENCES 265