Contents

Preface *page* xiii
Notation xvii

1 **Overview of Adversarial Learning** 1
- 1.1 Machine Learning and Its Attack Vectors 2
- 1.2 Attacker/Defender Goals and Assumptions 2
- 1.3 Test-Time Evasion Attacks (TTEs) or Adversarial Inputs 4
- 1.4 Data Poisoning (DP) Attacks 8
- 1.5 Reverse-Engineering Attacks (REAs) Targeting the Deep Neural Network (DNN) 13
- 1.6 Attacks on Privacy of Training Data 14
- 1.7 Chapter Summary 17
- 1.8 References for Further Reading 18

2 **Deep Learning Background** 19
- 2.1 Deep Learning for Classification, Regression, or Prediction 20
- 2.2 Motivating Deep Neural Network (DNN) Classifiers 23
- 2.3 Linearly Separable Data 23
- 2.4 From Binary to K-ary Classification 24
- 2.5 Deep Neural Network (DNN) Architectures 24
- 2.6 Background on Gradient-Based Optimization 31
- 2.7 Heuristic Optimization Methods for Deep Learning 36
- 2.8 Overfitting and DNN Regularization 41
- 2.9 "Certified" Training and Classification Confidence 44
- 2.10 Neural Network Inversion 46
- 2.11 Identification and Visualization of Salient Features 47
- 2.12 Handling Label-Deficient Data: Transfer and Contrastive Learning 47
- 2.13 Other Methods of Extracting Salient Features 50
- 2.14 Other Types of Classifiers: Naive Bayes (NB) and Logistic Regression (LR) 50
- 2.15 Discussion: Statistical Confidence 52
- 2.16 Chapter Summary 53
- 2.17 References for Further Reading 54
- 2.18 Project: Classification of UC Irvine Datasets 54

2.19 Project: Membership-Inference Attack 55
2.20 Project: Classification for the CIFAR-10 Image Domain 55

3 **Basics of Detection and Mixture Models** 56
3.1 Mixture Densities 58
3.2 Estimating the Parameters: Maximum Likelihood Estimation (MLE) and Expectation-Maximization (EM) 59
3.3 K-Means Clustering as a Special Case 63
3.4 Model Order Selection 63
3.5 Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) 66
3.6 Some Detection Basics 70
3.7 Performance Measures for Detection 72
3.8 Chapter Summary 72
3.9 References for Further Reading 73
3.10 Projects: Receiver Operating Characteristic (ROC), Principal Component Analysis (PCA), and Gaussian Mixture Model (GMM) 73

4 **Test-Time Evasion Attacks (Adversarial Inputs)** 76
4.1 Previously Proposed Test-Time Evasion (TTE) Attacks 76
4.2 "Robust" and "Certified" Defenses for Test-Time Evasion (TTE) Attacks 80
4.3 Anomaly Detection (AD) of Test-Time Evasion (TTE) Attacks 85
4.4 Background on Generative Modeling and Generative Adversarial Networks (GANs) 90
4.5 Generative Adversarial Network (GAN) Based Test-Time Evasion (TTE) Attack Detection Methodology 93
4.6 Experiments 96
4.7 Deeper Consideration of Test-Time Evasion (TTE) Attack Scenarios 105
4.8 Discussion: Out-of-Distribution Detection (OODD) 110
4.9 Chapter Summary 111
4.10 Project: White Region Counting Defense 112
4.11 Project: Nearest Neighbor (NN) Classification Defense 114
4.12 Project: Test-Time Evasion (TTE) Attacks and Dropout 115

5 **Backdoors and Before/During Training Defenses** 116
5.1 Backdoor Attacks 117
5.2 Before/During Training Defender's Goals 120
5.3 Before/During Training Defenses 121
5.4 Training Set Cleansing Reverse-Engineering Defense (TSC-RED) 122
5.5 Experiments 128
5.6 Defense Variations and Additional Experiments 138
5.7 Chapter Summary 139
5.8 Project: Principal Component Analysis (PCA) Based Cluster Impurity (CI) Defense 140

6 **Post-Training Reverse-Engineering Defense (PT-RED) Against Imperceptible Backdoors** 141
6.1 The Post-Training (PT) Scenario 141
6.2 Some Post-Training (PT) Defenses 143
6.3 Imperceptible-Backdoor Post-Training Reverse-Engineering Defense (I-PT-RED) 146
6.4 Experiments 154
6.5 Lagrangian Post-Training Reverse-Engineering Defense (L-PT-RED) and Experiments 173
6.6 Discussion: Robust and Explainable AI 178
6.7 Chapter Summary 179
6.8 Project: Imperceptible Backdoor Post-Training Reverse-Engineering Defense (I-PT-RED) on Images 180
6.9 Project: Consensus Post-Training Reverse-Engineering Defense (C-PT-RED) 182
6.10 Project: Noisy Backdoor Incorporation 184

7 **Post-Training Reverse-Engineering Defense (PT-RED) Against Patch-Incorporated Backdoors** 185
7.1 Perceptible Backdoor Patterns 186
7.2 Choice of Source Class(es) and Target Class 188
7.3 Perceptible Backdoor Post-Training Reverse-Engineering Defense (P-PT-RED) 188
7.4 Experiments 195
7.5 Chapter Summary 207
7.6 Project: Exploring What the Deep Neural Network (DNN) is Learning 209
7.7 Project: Perceptible Backdoor Post-Training Reverse-Engineering Defense (P-PT-RED) and Variations on Images 209

8 **Transfer Post-Training Reverse-Engineering Defense (T-PT-RED) Against Backdoors** 211
8.1 Transferability of Sample-wise Minimal Perturbations 211
8.2 Transfer Post-Training Reverse-Engineering Defense (T-PT-RED) Detection Procedure 213
8.3 Experiments 216
8.4 Chapter Summary 222
8.5 Project: Targeted Transfer Post-Training Reverse-Engineering Defense (T-PT-RED) for Multiple Classes 223
8.6 Project: Transfer Post-Training Reverse-Engineering Defense (T-PT-RED) for Backdoor Patches 224

9 **Universal Post-Training (PT) Backdoor Defenses** 226
9.1 Universal Backdoor Detection (UnivBD) Without Clean Labeled Data 227
9.2 Universal Mitigation of Backdoor Attack (UnivBM) 234
9.3 Some Additional Experiments 238

9.4 Chapter Summary 241
9.5 Project: Universal Backdoor Detection (UnivBD) Versus Lagrangian Post-Training Reverse-Engineering Defense (L-PT-RED) 243
9.6 Project: Reverse-Engineering Backdoor Patterns (BPs) 243
9.7 Project: Universal Backdoor Detection (UnivBD) Leveraging Clean Labeled Data 243
9.8 Project: Universal Detector Based on Deep Neural Network (DNN) Weight Outliers 244
9.9 Project: Mitigation Using Surrogates for Correct Decision Rate 244
9.10 Project: Testing the Hypothesis that a Backdoor Preserves Clean Logits 245
9.11 Project: Modified UnivBM Objective to Reduce Margin 245
9.12 Project: Defense Against Error-Generic Data Poisoning 245

10 Test-Time Detection of Backdoor Triggers 246
10.1 Some Test-Time Backdoor Detection Methods 246
10.2 In-Flight Reverse-Engineering Defense (IF-RED) 247
10.3 Experiments 249
10.4 Chapter Summary 252
10.5 Project: Trigger Detection via Test-Time Evasion (TTE) Attack Detection Strategy 254
10.6 Project: In-Flight Reverse-Engineering Defense (IF-RED) Using Imperceptible Backdoor Post-Training Reverse-Engineering Defense (I-PT-RED) Applied to Embedded Features 255
10.7 Project: Margin as an In-Flight Detection Statistic 255

11 Backdoors for 3D Point Cloud (PC) Classifiers 256
11.1 3D Point Cloud (PC) Classification 258
11.2 Backdoor Attacks against 3D Point Cloud (PC) Classifiers 258
11.3 A Small Cluster of Backdoor Points 260
11.4 Attack Experiments 265
11.5 Point Cloud (PC) Anomaly Detectors (ADs) against Backdoor Attacks (BAs) 271
11.6 Point Cloud Post-Training Reverse-Engineering Defense (PC-PT-RED) 272
11.7 Attack/Defense Experiments 276
11.8 Chapter Summary 278
11.9 Project: During-Training Defense or Robustification of Point Cloud (PC) Classifiers 280

12 Robust Deep Regression and Active Learning 281
12.1 Background on Active Learning 283
12.2 Robust Deep Regression by Active Learning (RDR-AL) 284
12.3 A Localized Region of Regression Error 286
12.4 Experimental Results for Valuation of a Financial Option 287

12.5 Discussion Topics: Query by Committee, Reinforcement Learning (RL), Test-Time Evasion (TTE), Classification 290
12.6 Chapter Summary 292
12.7 Project: Clean Label Backdoor Attack 292

13 Error Generic Data Poisoning Defense 294
13.1 Threat Model 295
13.2 Data Poisoning Defenses 296
13.3 Bayesian Information Criterion Based Mixture Model Training Set Cleansing (BIC-MM-TSC) 298
13.4 Experiments on Binary, Discrete Feature Classification Tasks 304
13.5 Discussion: Experiments with $K > 2$ Classes 312
13.6 Chapter Summary 312
13.7 Project: K-Nearest Neighbor (KNN) Defense 313
13.8 Project: White Box Data Poisoning Attack 314

14 Reverse-Engineering Attacks (REAs) on Classifiers 315
14.1 Reverse-Engineering Attacks (REAs) Given Domain Samples 316
14.2 Overview of Defense Against Reverse-Engineering Attacks (REAs) 316
14.3 Anomaly Detection of Attacks (ADA) Based Defense Against Reverse-Engineering Attacks (REA-ADA) 317
14.4 Experiments 317
14.5 Chapter Summary 319
14.6 Project: Defense Against Random Querying 321

Appendix Support Vector Machines (SVMs) 322

References 333
Index 351