

Trustworthy Machine Learning

Kush R. Varshney

Part 1 Introduction and Preliminaries

1 Establishing Trust

- 1.1 Introduction
- 1.2 Defining trust
- 1.3 Organization of the book
- 1.4 Limitations
- 1.5 Positionality statement
- 1.6 Summary

2 Machine Learning Lifecycle

- 2.1 Introduction
- 2.2 Problem specification
- 2.3 Data understanding
- 2.4 Data preparation
- 2.5 Modeling
- 2.6 Evaluation
- 2.7 Deployment and monitoring
- 2.8 Summary

3 Uncertainty

- 3.1 Introduction
- 3.2 Aleatoric uncertainty and probability
- 3.3 Conditional probability and independence
- 3.4 Bayesian networks
- 3.5 Epistemic uncertainty, possibility, and imprecise probability
- 3.6 Summary

4 Detection Theory

- 4.1 Introduction
- 4.2 Confusion matrix and costs
- 4.3 Bayesian detection
- 4.4 Receiver operating characteristic and calibration
- 4.6 Min-max and Neyman-Pearson detection
- 4.7 Information-theoretic concepts
- 4.8 Summary

5 Causality

- 5.1 Introduction
- 5.2 Interventions
- 5.3 Counterfactuals
- 5.4 Causal graphs
- 5.5 Summary

Part 2 Data

6 Modalities and Sources

- 6.1 Introduction

6.2	Modalities
6.3	Administrative data
6.4	Social data
6.5	Crowdsourcing
6.6	Data augmentation
6.7	Summary
7	Biases
7.1	Introduction
7.2	Temporal biases
7.3	Sampling bias
7.4	Cognitive biases
7.5	Poisoning
7.6	Data preparation biases
7.7	Summary
8	Privacy and Consent
8.1	Introduction
8.2	Statistical foundations of privacy
8.3	Causal foundations of privacy
8.4	Consent
8.5	Summary
Part 3 Basic Modeling	
9	Risk Minimization
9.1	Introduction
9.2	Empirical risk minimization
9.3	Structural risk minimization
9.4	Summary
10	Decision Stumps and Their Generalizations
10.1	Introduction
10.2	Trees and forests
10.3	Perceptrons
10.4	Margin-based methods
10.5	Neural networks
10.6	Summary
11	Adversarial and Game-Theoretic Learning
11.1	Introduction
11.2	Game-theoretic interpretation of boosting
11.3	Generative adversarial networks
11.4	Summary
12	Causal Modeling
12.1	Introduction
12.2	Causal inference basics
12.3	Treatment effect estimation
12.4	Causal discovery
12.5	Summary

Part 4 Safety and Reliability

13 Epistemic Uncertainty in Machine Learning

- 13.1 Introduction
- 13.2 Definition of safety
- 13.3 Manifestations of epistemic uncertainty in machine learning
- 13.4 Summary

14 Distribution Shift

- 14.1 Introduction
- 14.2 Statistical foundations of distribution shift
- 14.3 Causal foundations of distribution shift
- 14.4 Domain adaptation
- 14.5 Invariant risk minimization
- 14.6 Performative prediction
- 14.7 Summary

15 Fairness

- 15.1 Introduction
- 15.2 Statistical foundations of fairness
- 15.3 Causal foundations of fairness
- 15.4 Bias mitigation algorithms
- 15.5 Summary

16 Adversarial Robustness

- 16.1 Introduction
- 16.2 Statistical foundations of adversarial robustness
- 16.3 Causal foundations of adversarial robustness
- 16.4 Attacks
- 16.5 Defenses
- 16.6 Summary

17 Testing

- 17.1 Introduction
- 17.2 Testing workflow
- 17.3 Testing components
- 17.4 Testing properties
- 17.5 Summary

Part 5 Interaction

18 Interpretability and Explainability

- 18.1 Introduction
- 18.2 Directly interpretable models
- 18.3 Post hoc local explanations
- 18.4 Post hoc global explanations
- 18.5 Explaining quantities other than predictions
- 18.6 Summary

19 Provenance and Transparency

- 19.1 Introduction

- 19.2 Documentation
- 19.3 Blockchain
- 19.4 Open platforms
- 19.5 Summary
- 20 Value Alignment
 - 20.1 Introduction
 - 20.2 Unified theory of trust
 - 20.3 Preference elicitation
 - 20.4 Specification gaming
 - 20.5 Summary
- Part 6 Purpose
- 21 Disinformation and Filter Bubbles
 - 21.1 Introduction
 - 21.2 Deepfakes
 - 21.3 Filter bubbles and echo chambers
 - 21.4 Summary
- 22 Professional Codes and Ethics Guidelines
 - 22.1 Introduction
 - 22.2 Landscape of codes and guidelines
 - 22.3 Ethicswashing
 - 22.4 From principles to practice
 - 22.5 Summary
- 23 Lived Experience
 - 23.1 Introduction
 - 23.2 Diversity and problem specification
 - 23.3 Diversity and solution development
 - 23.4 Summary
- 24 Social Good
 - 24.1 Introduction
 - 24.2 Examples of machine learning for social good
 - 24.3 Common patterns in machine learning for social good
 - 24.4 Open platforms for greater impact
 - 24.5 Summary