Table of Contents

Preface. xi

1. Overview of Machine Learning Systems. 1
When to Use Machine Learning 3
Machine Learning Use Cases 9
Understanding Machine Learning Systems 12
Machine Learning in Research Versus in Production 12
Machine Learning Systems Versus Traditional Software 22
Summary 23

2. Introduction to Machine Learning Systems Design. 25
Business and ML Objectives 26
Requirements for ML Systems 29
Reliability 29
Scalability 30
Maintainability 31
Adaptability 31
Iterative Process 32
Framing ML Problems 35
Types of ML Tasks 36
Objective Functions 40
Mind Versus Data 43
Summary 46

3. Data Engineering Fundamentals. 49
Data Sources 50
Data Formats 53
JSON 54

Row-Major Versus Column-Major Format 54
Text Versus Binary Format 57
Data Models 58
Relational Model 59
NoSQL 63
Structured Versus Unstructured Data 66
Data Storage Engines and Processing 67
Transactional and Analytical Processing 67
ETL: Extract, Transform, and Load 70
Modes of Dataflow 72
Data Passing Through Databases 72
Data Passing Through Services 73
Data Passing Through Real-Time Transport 74
Batch Processing Versus Stream Processing 78
Summary 79

4. Training Data. 81
Sampling 82
Nonprobability Sampling 83
Simple Random Sampling 84
Stratified Sampling 84
Weighted Sampling 85
Reservoir Sampling 86
Importance Sampling 87
Labeling 88
Hand Labels 88
Natural Labels 91
Handling the Lack of Labels 94
Class Imbalance 102
Challenges of Class Imbalance 103
Handling Class Imbalance 105
Data Augmentation 113
Simple Label-Preserving Transformations 114
Perturbation 114
Data Synthesis 116
Summary 118

5. Feature Engineering. 119
Learned Features Versus Engineered Features 119
Common Feature Engineering Operations 123
Handling Missing Values 123
Scaling 126

Discretization 128
Encoding Categorical Features 129
Feature Crossing 132
Discrete and Continuous Positional Embeddings 133
Data Leakage 135
Common Causes for Data Leakage 137
Detecting Data Leakage 140
Engineering Good Features 140
Feature Importance 141
Feature Generalization 144
Summary 146

6. Model Development and Offline Evaluation. 147
Model Development and Training 148
Evaluating ML Models 148
Ensembles 154
Experiment Tracking and Versioning 160
Distributed Training 165
AutoML 170
Model Offline Evaluation 176
Baselines 177
Evaluation Methods 179
Summary 186

7. Model Deployment and Prediction Service. 189
Machine Learning Deployment Myths 192
Myth 1: You Only Deploy One or Two ML Models at a Time 192
Myth 2: If We Don't Do Anything, Model Performance Remains the Same 193
Myth 3: You Won't Need to Update Your Models as Much 194
Myth 4: Most ML Engineers Don't Need to Worry About Scale 194
Batch Prediction Versus Online Prediction 195
From Batch Prediction to Online Prediction 199
Unifying Batch Pipeline and Streaming Pipeline 201
Model Compression 204
Low-Rank Factorization 204
Knowledge Distillation 206
Pruning 206
Quantization 207
ML on the Cloud and on the Edge 210
Compiling and Optimizing Models for Edge Devices 212
ML in Browsers 220
Summary 221

8. Data Distribution Shifts and Monitoring. 223
Causes of ML System Failures 224
Software System Failures 225
ML-Specific Failures 226
Data Distribution Shifts 234
Types of Data Distribution Shifts 234
General Data Distribution Shifts 238
Detecting Data Distribution Shifts 239
Addressing Data Distribution Shifts 245
Monitoring and Observability 247
ML-Specific Metrics 248
Monitoring Toolbox 253
Observability 256
Summary 258

9. Continual Learning and Test in Production. 261
Continual Learning 262
Stateless Retraining Versus Stateful Training 263
Why Continual Learning? 266
Continual Learning Challenges 268
Four Stages of Continual Learning 272
How Often to Update Your Models 277
Test in Production 279
Shadow Deployment 280
A/B Testing 281
Canary Release 282
Interleaving Experiments 283
Bandits 285
Summary 289

10. Infrastructure and Tooling for MLOps. 291
Storage and Compute 295
Public Cloud Versus Private Data Centers 297
Development Environment 300
Dev Environment Setup 301
Standardizing Dev Environments 304
From Dev to Prod: Containers 306
Resource Management 308
Cron, Schedulers, and Orchestrators 309
Data Science Workflow Management 312
ML Platform 317
Model Deployment 318

Model Store 319
Feature Store 323
Build Versus Buy 325
Summary 327

11. The Human Side of Machine Learning. 329
User Experience 329
Ensuring User Experience Consistency 330
Combatting “Mostly Correct” Predictions 330
Smooth Failing 332
Team Structure 332
Cross-functional Teams Collaboration 333
End-to-End Data Scientists 333
Responsible AI 337
Irresponsible AI: Case Studies 339
A Framework for Responsible AI 345
Summary 351

Epilogue. 353

Index. 355