

contents

<i>foreword</i>	xiii
<i>preface</i>	xv
<i>acknowledgments</i>	xvii
<i>about this book</i>	xviii
<i>about the authors</i>	xxi
<i>about the cover illustration</i>	xxii

What is machine learning? 3

- 1.1 Understanding how machines learn 4
- 1.2 Using data to make decisions 7
 - Traditional approaches* 8 ■ *The machine-learning approach* 11
 - Five advantages to machine learning* 16 ■ *Challenges* 16
- 1.3 Following the ML workflow: from data to deployment 17
 - Data collection and preparation* 18 ■ *Learning a model from data* 19 ■ *Evaluating model performance* 20
 - Optimizing model performance* 21

CONTENTS

- 1.1 Boosting model performance with advanced techniques 22
 - Data preprocessing and feature engineering* 22
 - *Improving models continually with online methods* 24
 - *Scaling models with data volume and velocity* 25
- 1.5 Summary 25
- 1.6 Terms from this chapter 25

Real-world data 27

- 2.1 Getting started: data collection 28
 - Which features should be included?* 30
 - *How can we obtain ground truth for the target variable?* 32
 - *How much training data is required?* 33
 - *Is the training set representative enough?* 35
- 2.2 Preprocessing the data for modeling 36
 - Categorical features* 36
 - *Dealing with missing data* 38
 - Simple feature engineering* 40
 - *Data normalization* 42
- 2.3 Using data visualization 43
 - Mosaic plots* 44
 - *Box plots* 46
 - *Density plots* 48
 - Scatter plots* 50
- 2.4 Summary 50
- 2.5 Terms from this chapter 51

Modeling and prediction 52

- 3.1 Basic machine-learning modeling 53
 - Finding the relationship between input and target* 53
 - The purpose of finding a good model* 55
 - *Types of modeling methods* 56
 - *Supervised versus unsupervised learning* 58
- 3.2 Classification: predicting into buckets 59
 - Building a classifier and making predictions* 61
 - Classifying complex, nonlinear data* 64
 - Classifying with multiple classes* 66
- 3.3 Regression: predicting numerical values 68
 - Building a regressor and making predictions* 69
 - Performing regression on complex, nonlinear data* 73
- 3.4 Summary 74
- 3.5 Terms from this chapter 75

Model evaluation and optimization 77

- 4.1 Model generalization: assessing predictive accuracy for new data 78
The problem: overfitting and model optimism 79 ■ *The solution: cross-validation 82* ■ *Some things to look out for when using cross-validation 86*
- 4.2 Evaluation of classification models 87
Class-wise accuracy and the confusion matrix 89
Accuracy trade-offs and ROC curves 90 ■ *Multiclass classification 93*
- 4.3 Evaluation of regression models 96
Using simple regression performance metrics 97
Examining residuals 99
- 4.4 Model optimization through parameter tuning 100
ML algorithms and their tuning parameters 100
Grid search 101
- 4.5 Summary 104
- 4.6 Terms from this chapter 105

Basic feature engineering 106

- 5.1 Motivation: why is feature engineering useful? 107
What is feature engineering? 107 ■ *Five reasons to use feature engineering 107* ■ *Feature engineering and domain expertise 109*
- 5.2 Basic feature-engineering processes 110
Example: event recommendation 110 ■ *Handling date and time features 112* ■ *Working with simple text features 114*
- 5.3 Feature selection 116
Forward selection and backward elimination 119 ■ *Feature selection for data exploration 121* ■ *Real-world feature selection example 123*
- 5.4 Summary 125
- 5.5 Terms from this chapter 126

Example: NYC taxi data 129

- 6.1 Data: NYC taxi trip and fare information 130
 - Visualizing the data 130* ■ *Defining the problem and preparing the data 134*
- 6.2 Modeling 137
 - Basic linear model 137* ■ *Nonlinear classifier 138*
 - Including categorical features 140* ■ *Including date-time features 142* ■ *Model insights 143*
- 6.3 Summary 144
- 6.4 Terms from this chapter 145

Advanced feature engineering 146

- 7.1 Advanced text features 146
 - Bag-of-words model 147* ■ *Topic modeling 149*
 - Content expansion 152*
- 7.2 Image features 154
 - Simple image features 154* ■ *Extracting objects and shapes 156*
- 7.3 Time-series features 160
 - Types of time-series data 160* ■ *Prediction on time-series data 163* ■ *Classical time-series features 163*
 - Feature engineering for event streams 168*
- 7.4 Summary 168
- 7.5 Terms from this chapter 170

Advanced NLP example: movie review sentiment 172

- 8.1 Exploring the data and use case 173
 - A first glance at the dataset 173* ■ *Inspecting the dataset 174*
 - So what's the use case? 175*
- 8.2 Extracting basic NLP features and building the initial model 178
 - Bag-of-words features 178* ■ *Building the model with the naïve Bayes algorithm 180* ■ *Normalizing bag-of-words features with the tf-idf algorithm 184* ■ *Optimizing model parameters 185*

- 8.3 Advanced algorithms and model deployment considerations 190
 - Word2vec features* 190 ■ *Random forest model* 192
- 8.4 Summary 195
- 8.5 Terms from this chapter 195

***Scaling machine-learning workflows* 196**

- 9.1 Before scaling up 197
 - Identifying important dimensions* 197 ■ *Subsampling training data in lieu of scaling?* 199 ■ *Scalable data management systems* 201
- 9.2 Scaling ML modeling pipelines 203
 - Scaling learning algorithms* 204
- 9.3 Scaling predictions 207
 - Scaling prediction volume* 208 ■ *Scaling prediction velocity* 209
- 9.4 Summary 211
- 9.5 Terms from this chapter 212

***Example: digital display advertising* 214**

- 10.1 Display advertising 215
- 10.2 Digital advertising data 216
- 10.3 Feature engineering and modeling strategy 216
- 10.4 Size and shape of the data 218
- 10.5 Singular value decomposition 220
- 10.6 Resource estimation and optimization 222
- 10.7 Modeling 224
- 10.8 K-nearest neighbors 224
- 10.9 Random forests 226
- 10.10 Other real-world considerations 227
- 10.11 Summary 228
- 10.12 Terms from this chapter 229
- 10.13 Recap and conclusion 229

appendix Popular machine-learning algorithms 232

index 236