

Contents

Chapter 1

Introduction to machine learning — 1

- 1.1 What is machine learning — 1
- 1.2 Types of machine learning algorithms — 2
- 1.3 Core concepts in machine learning — 3
 - 1.3.1 Classification — 3
 - 1.3.2 Numerical prediction — 5
 - 1.3.3 Dimension reduction — 8
- 1.4 Fundamentals of probability, statistics, and linear algebra — 9
 - 1.4.1 Statistical concepts — 9
- 1.5 Linear algebra: matrices — 25

Chapter 2

Preliminary steps for machine learning — 27

- 2.1 Introduction — 27
 - 2.1.1 Getting the software tools: MATLAB and R — 27
- 2.2 Data exploration — 31
 - 2.2.1 Data Exploration with R — 32
 - 2.2.2 Filtering and counting with R — 34
 - 2.2.3 Sorting with R — 35
- 2.3 Data cleaning — 36
 - 2.3.1 Working with missing values in R — 37
 - 2.3.2 Dealing with outliers in R — 39
 - 2.3.3 Creating subsets in R — 40
- 2.4 Data transformations — 42
 - 2.4.1 Assigning numerical values to categorical data using R — 42
 - 2.4.2 Assigning bins to numeric data with R — 44
 - 2.4.3 Data standardization with R — 46
- 2.5 Data visualization with R — 46
- 2.6 Basic statistical analysis with MATLAB — 56

Chapter 3

Predictive performance evaluation — 73

- 3.1 Accuracy for the prediction of numerical values — 73
 - 3.1.1 Mean squared error (MSE) — 73
 - 3.1.2 R^2 (coefficient of determination) — 73
- 3.2 Accuracy for classification — 75
 - 3.2.1 Confusion matrix — 75
 - 3.2.2 Accuracy — 76
 - 3.2.3 Precision — 76

3.2.4 Recall (sensitivity or true positive rate) — **76**
3.2.5 *F1* score — **76**

Chapter 4

Supervised learning models — 80

4.1 K-nearest neighbors — **80**
4.1.1 Classification with KNN in R — **80**
4.1.2 Numerical prediction with KNN in R — **86**
4.2 Analysis with MATLAB — **90**
4.2.1 KNN model — **90**
4.2.2 Naïve Bayes’ model — **91**
4.2.3 Support vector machine (SVM) — **94**
4.2.4 Decision tree — **94**
4.2.5 Linear regression — **100**
4.2.6 Neural network — **101**
4.2.7 Gradient descent algorithm — **103**

Chapter 5

Unsupervised learning model — 106

5.1 Dimension reduction using principal component analysis (PCA) — **106**
5.2 Cluster analysis — **110**
5.2.1 K-means cluster analysis — **110**
5.2.2 K-Means clustering on numerical data with R — **111**
5.2.3 K-medoids method (Partitioning Around Medoids) on categorical data with R — **115**
5.2.4 Hierarchical Clustering — **120**
5.2.5 Hierarchical clustering on numerical data with R — **121**
5.2.6 Hierarchical clustering on categorical data with R — **124**
5.3 Clustering with MATLAB — **126**

Chapter 6

Deep learning — 129

6.1 Transfer learning — **129**
6.2 Long short-term memory (LSTM) — **136**

Chapter 7

Text mining — 141

- 7.1 Introduction — 141
- 7.2 Text extraction — 141
 - 7.2.1 Extracting data from a pdf file — 141
 - 7.2.2 Creating bag of words, term document matrix, and word frequencies — 143
 - 7.2.3 Creating word cloud — 146
 - 7.2.4 Sentiment analysis — 148
 - 7.2.5 Topic modeling and clustering — 150

References — 159

Index — 161