Contents

Acknowledgements *xiii*

Introduction *xv*

1 From Big Data to Deep Learning *1*

1.1 Introduction *1*

1.2 Examples of the Use of Big Data and Deep Learning *6*

1.3 Big Data and Deep Learning for Companies and Organizations *9*

1.3.1 Big Data in Finance *10*

1.3.1.1 Google Trends *10*

1.3.1.2 Google Trends and Stock Prices *11*

1.3.1.3 The `quantmod` Package for Financial Analysis *11*

1.3.1.4 Google Trends in R *13*

1.3.1.5 Matching Data from `quantmod` and Google Trends *14*

1.3.2 Big Data and Deep Learning in Insurance *18*

1.3.3 Big Data and Deep Learning in Industry *18*

1.3.4 Big Data and Deep Learning in Scientific Research and Education *20*

1.3.4.1 Big Data in Physics and Astrophysics *20*

1.3.4.2 Big Data in Climatology and Earth Sciences *21*

1.3.4.3 Big Data in Education *21*

1.4 Big Data and Deep Learning for Individuals *21*

1.4.1 Big Data and Deep Learning in Healthcare *21*

1.4.1.1 Connected Health and Telemedicine *21*

1.4.1.2 Geolocation and Health *22*

1.4.1.3 The Google Flu Trends *23*

1.4.1.4 Research in Health and Medicine *26*

1.4.2 Big Data and Deep Learning for Drivers *28*

1.4.3 Big Data and Deep Learning for Citizens *29*

1.4.4 Big Data and Deep Learning in the Police *30*

1.5 Risks in Data Processing *32*

1.5.1 Insufficient Quantity of Training Data *32*

1.5.2 Poor Data Quality *32*

1.5.3 Non-Representative Samples *33*

1.5.4 Missing Values in the Data *33*

COPYRIGHTED MATERIAL

1.5.5 Spurious Correlations *34*
1.5.6 Overfitting *35*
1.5.7 Lack of Explainability of Models *35*
1.6 Protection of Personal Data *36*
1.6.1 The Need for Data Protection *36*
1.6.2 Data Anonymization *38*
1.6.3 The General Data Protection Regulation *41*
1.7 Open Data *43*
Notes *44*

2 Processing of Large Volumes of Data *49*
2.1 Issues *49*
2.2 The Search for a Parsimonious Model *50*
2.3 Algorithmic Complexity *51*
2.4 Parallel Computing *51*
2.5 Distributed Computing *52*
2.5.1 MapReduce *53*
2.5.2 Hadoop *54*
2.5.3 Computing Tools for Distributed Computing *55*
2.5.4 Column-Oriented Databases *56*
2.5.5 Distributed Architecture and “Analytics” *57*
2.5.6 Spark *58*
2.6 Computer Resources *60*
2.6.1 Minimum Resources *60*
2.6.2 Graphics Processing Units (GPU) and Tensor Processing Units (TPU) *61*
2.6.3 Solutions in the Cloud *62*
2.7 R and Python Software *62*
2.8 Quantum Computing *67*
Notes *68*

3 Reminders of Machine Learning *71*
3.1 General *71*
3.2 The Optimization Algorithms *74*
3.3 Complexity Reduction and Penalized Regression *85*
3.4 Ensemble Methods *89*
3.4.1 Bagging *89*
3.4.2 Random Forests *89*
3.4.3 Extra-Trees *91*
3.4.4 Boosting *92*
3.4.5 Gradient Boosting Methods *97*
3.4.6 Synthesis of the Ensemble Methods *100*
3.5 Support Vector Machines *100*
3.6 Recommendation Systems *105*
Notes *108*

4 Natural Language Processing *111*
4.1 From Lexical Statistics to Natural Language Processing *111*
4.2 Uses of Text Mining and Natural Language Processing *113*
4.3 The Operations of Textual Analysis *114*
4.3.1 Textual Data Collection *115*
4.3.2 Identification of the Language *115*
4.3.3 Tokenization *116*
4.3.4 Part-of-Speech Tagging *117*
4.3.5 Named Entity Recognition *119*
4.3.6 Coreference Resolution *124*
4.3.7 Lemmatization *124*
4.3.8 Stemming *129*
4.3.9 Simplifications *129*
4.3.10 Removal of Stop Words *130*
4.4 Vector Representation and Word Embedding *132*
4.4.1 Vector Representation *132*
4.4.2 Analysis on the Document-Term Matrix *133*
4.4.3 TF-IDF Weighting *142*
4.4.4 Latent Semantic Analysis *144*
4.4.5 Latent Dirichlet Allocation *152*
4.4.6 Word Frequency Analysis *160*
4.4.7 Word2Vec Embedding *162*
4.4.8 GloVe Embedding *174*
4.4.9 FastText Embedding *176*
4.5 Sentiment Analysis *180*
Notes *184*

5 Social Network Analysis *187*
5.1 Social Networks *187*
5.2 Characteristics of Graphs *188*
5.3 Characterization of Social Networks *189*
5.4 Measures of Influence in a Graph *190*
5.5 Graphs with R *191*
5.6 Community Detection *200*
5.6.1 The Modularity of a Graph *201*
5.6.2 Community Detection by Divisive Hierarchical Clustering *202*
5.6.3 Community Detection by Agglomerative Hierarchical Clustering *203*
5.6.4 Other Methods *204*
5.6.5 Community Detection with R *205*
5.7 Research and Analysis on Social Networks *208*
5.8 The Business Model of Social Networks *209*
5.9 Digital Advertising *211*
5.10 Social Network Analysis with R *212*
5.10.1 Collecting Tweets *213*
5.10.2 Formatting the Corpus *215*

5.10.3 Stemming and Lemmatization *216*
5.10.4 Example *217*
5.10.5 Clustering of Terms and Documents *225*
5.10.6 Opinion Scoring *230*
5.10.7 Graph of Terms with Their Connotation *231*
Notes *234*

6 Handwriting Recognition *237*
6.1 Data *237*
6.2 Issues *238*
6.3 Data Processing *238*
6.4 Linear and Quadratic Discriminant Analysis *243*
6.5 Multinomial Logistic Regression *245*
6.6 Random Forests *246*
6.7 Extra-Trees *247*
6.8 Gradient Boosting *249*
6.9 Support Vector Machines *253*
6.10 Single Hidden Layer Perceptron *258*
6.11 H2O Neural Network *262*
6.12 Synthesis of “Classical” Methods *267*
Notes *268*

7 Deep Learning *269*
7.1 The Principles of Deep Learning *269*
7.2 Overview of Deep Neural Networks *272*
7.3 Recall on Neural Networks and Their Training *274*
7.4 Difficulties of Gradient Backpropagation *284*
7.5 The Structure of a Convolutional Neural Network *286*
7.6 The Convolution Mechanism *288*
7.7 The Convolution Parameters *290*
7.8 Batch Normalization *292*
7.9 Pooling *293*
7.10 Dilated Convolution *295*
7.11 Dropout and DropConnect *295*
7.12 The Architecture of a Convolutional Neural Network *297*
7.13 Principles of Deep Network Learning for Computer Vision *299*
7.14 Adaptive Learning Algorithms *301*
7.15 Progress in Image Recognition *304*
7.16 Recurrent Neural Networks *312*
7.17 Capsule Networks *317*
7.18 Autoencoders *318*
7.19 Generative Models *322*
7.19.1 Generative Adversarial Networks *323*
7.19.2 Variational Autoencoders *324*

7.20 Other Applications of Deep Learning *326*
7.20.1 Object Detection *326*
7.20.2 Autonomous Vehicles *333*
7.20.3 Analysis of Brain Activity *334*
7.20.4 Analysis of the Style of a Pictorial Work *336*
7.20.5 Go and Chess Games *338*
7.20.6 Other Games *340*
Notes *341*

8 Deep Learning for Computer Vision *347*
8.1 Deep Learning Libraries *347*
8.2 MXNet *349*
8.2.1 General Information about MXNet *349*
8.2.2 Creating a Convolutional Network with MXNet *350*
8.2.3 Model Management with MXNet *361*
8.2.4 CIFAR-10 Image Recognition with MXNet *362*
8.3 Keras and TensorFlow *367*
8.3.1 General Information about Keras *370*
8.3.2 Application of Keras to the MNIST Database *371*
8.3.3 Application of Pre-Trained Models *375*
8.3.4 Explain the Prediction of a Computer Vision Model *379*
8.3.5 Application of Keras to CIFAR-10 Images *382*
8.3.6 Classifying Cats and Dogs *393*
8.4 Configuring a Machine's GPU for Deep Learning *408*
8.4.1 Checking the Compatibility of the Graphics Card *408*
8.4.2 NVIDIA Driver Installation *409*
8.4.3 Installation of Microsoft Visual Studio *409*
8.4.4 NVIDIA CUDA Toolkit Installation *409*
8.4.5 Installation of cuDNN *410*
8.5 Computing in the Cloud *411*
8.6 PyTorch *417*
8.6.1 The Python `PyTorch` Package *417*
8.6.2 The R `torch` Package *424*
Notes *428*

9 Deep Learning for Natural Language Processing *431*
9.1 Neural Network Methods for Text Analysis *431*
9.2 Text Generation Using a Recurrent Neural Network LSTM *432*
9.3 Text Classification Using a LSTM or GRU Recurrent Neural Network *438*
9.4 Text Classification Using a H2O Model *450*
9.5 Application of Convolutional Neural Networks *454*
9.6 Spam Detection Using a Recurrent Neural Network LSTM *458*
9.7 Transformer Models, BERT, and Its Successors *459*
Notes *477*

10 Artificial Intelligence *479*
10.1 The Beginnings of Artificial Intelligence *479*
10.2 Human Intelligence and Artificial Intelligence *484*
10.3 The Different Forms of Artificial Intelligence *486*
10.4 Ethical and Societal Issues of Artificial Intelligence *491*
10.5 Fears and Hopes of Artificial Intelligence *495*
10.6 Some Dates of Artificial Intelligence *498*
Notes *501*

Conclusion *505*
Note *506*

Annotated Bibliography *507*
On Big Data and High Dimensional Statistics *507*
On Deep Learning *509*
On Artificial Intelligence *511*
On the Use of R and Python in Data Science and on Big Data *512*

Index *515*