

Pramod Gupta · Naresh Kumar Sehgal ·
John M. Acken

Introduction to Machine Learning with Security

Theory and Practice Using Python in the Cloud

Second Edition

Synthesis Lectures on Engineering, Science, and Technology

The focus of this series is general topics, and applications about, and for, engineers and scientists on a wide array of applications, methods and advances. Most titles cover subjects such as professional development, education, and study skills, as well as basic introductory undergraduate material and other topics appropriate for a broader and less technical audience.

Pramod Gupta · Naresh Kumar Sehgal ·
John M. Acken

Introduction to Machine Learning with Security

Theory and Practice Using Python
in the Cloud

Second Edition

 Springer

Pramod Gupta
Brentwood, CA, USA

Naresh Kumar Sehgal
NovaSignal Corp
Santa Clara, CA, USA

John M. Acken
Portland State University
Vancouver, WA, USA

ISSN 2690-0300 ISSN 2690-0327 (electronic)
Synthesis Lectures on Engineering, Science, and Technology
ISBN 978-3-031-59169-3 ISBN 978-3-031-59170-9 (eBook)
<https://doi.org/10.1007/978-3-031-59170-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer
Nature Switzerland AG 2021, 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Pramod dedicates this book to his late parents who always encouraged him to pursue his goals.

Naresh dedicates this book to his father, late Shri Pawan Kumar Sehgal, whose inspiration encouraged Naresh to continue his lifelong education.

John dedicates this book to his wife, Dr. Charlotte Acken, and his daughter, McKinsey Schenker.

Foreword to the Second Edition

In the recent past, artificial intelligence and machine learning has deeply impacted every industry: it is now used in virtually every industry, namely, web search, smart phone, speech recognition, medical and life sciences, self-driving cars to name a few.

Pramod et al. have written an excellent and comprehensive textbook enabling readers to understand and comprehend backgrounds to each of these domains. The book takes the students to comprehend and learn the concepts of data science and statistics, thereby setting up their own machine learning platform with open-source tools. The intention behind the book is to concentrate more on the usage and application of machine learning. The book covers a wide base of techniques, from the simplest and most commonly used algorithms to complex machine learning algorithms.

The authors of this book have leveraged their hands-on experience with solving real-world problems using Python and the Machine Learning ecosystem to help the readers gain solid knowledge needed to apply essential concepts, methodologies, tools, and techniques for solving their own real-world problems.

The book aims to cater to readers with varying skill levels ranging from beginners to experts and enable them in structuring and building practical Machine Learning and AI solutions.

This book is appropriate for both advanced undergraduate or master's students who want to work in this domain, or for individuals working in the area of machine learning. It is an excellent resource for those who wish to start learning data science and machine learning so as to understand and use these powerful techniques in their work area.

By the end of the book, readers will have the knowledge on the tools needed to begin their journey in the domain of machine learning and artificial intelligence.

The book will also support students to treat information securely as both AI and ML are intertwined due to data confidentiality, model integrity, and system availability.

Solutions such as encryption, hashing, and secure computations have been used for both AI and ML.

Readers will find their knowledge significantly enhanced after going through the book.

Mumbai, India

Dr. Praveer Sinha
CEO and MD
The Tata Power Company Limited

Foreword to the First Edition

The invention of the cloud brought massive advancement in computer architecture. Seemingly overnight the delivery and consumption of services were transformed. Cloud computing has been an accelerant to the pace of technology adoption. In fact, it is cloud computing that has enabled artificial intelligence to move from the world of sci-fi to mainstream consumption. The cloud delivers the massive data storage and computes capacity needed for cost-effective analytics and insights.

As the promises of reduced time to deployment, lower cost of operation, and increased accessibility have been realized, adoption of the cloud has moved from novel to normal. And yet the theoretical understanding of the underlying technology remains locked in the brains of a relatively small group. The availability of cloud and AI curriculum in all educational settings needs to match the availability of cloud and AI services if the current pace of innovation is to continue.

For those looking to gain insight into the complex algorithms of artificial intelligence and the cloud architecture that fuels them, this book fills the void with highly valuable instruction.

Los Angeles, California, USA
January 2021

Diane Bryant
CEO and Chairman
<https://www.NovaSignal.com>

Preface to the Second Edition

Since the publication of our book's first edition in 2021, the AI world has changed in many ways. It is akin to the Gold Rush of California in 1849, where prospectors came from all over the world to become rich. Some were lucky and found gold, but most of the riches went to the companies that were providing supplies to the prospectors. One of them, Levi Strauss, started by an eponymous immigrant from Bavaria still survives. He recognized a need among the hardworking miners for clothing to endure anything. This need was met with blue jeans. Similarly, in the modern-day gold rush of AI, a few companies are doing very well supplying GPUs, Cloud, and other tools to the folks currently doing datamining. It remains to be seen how many of these will last as long as Levi Strauss has endured.

We felt a need to update our book to reflect this new development in the AI world, including a not-so-broadly recognized gap with security of data and code. Hence, we included a new co-author, Dr. John M. Acken. He is the professor of hardware security at Portland State University in Oregon, before finishing his Ph.D. at Stanford and then working for over half-a-century at various organizations such as US Army, Sandia National Labs, and Intel. The new Security chapters and sections that John has added to our book fill an important gap in the current practices of AI in the real world. Pramod and Naresh have expanded their existing chapters to include new material related to AI algorithms, practices to remove noisy data, optimizations in the Cloud, and an introduction to the new products related to Large Language Models (LLMs). Special thanks to Shiva Kintali for introducing us to jailbreaking LLMs. A tutorial for implementing Transformers Architecture has been added in the Appendix.

We hope that our readers will find this updated edition useful in their continuing learning and professional practices. As always, please reach out to us for your suggestions for our future editions.

Brentwood, CA, USA
Santa Clara, CA, USA
Portland, OR, USA

Pramod Gupta
Naresh Kumar Sehgal
John M. Acken

Preface to the First Edition

The idea for this book came from our lead co-author, Pramod Gupta, who has been teaching Data Sciences and related classes at University of California, Santa Cruz Extension, and most recently at University of California, Berkeley for several years. Prior to that his hands-on experience in the industry uniquely qualifies him to write the AI and ML parts of this book. Pramod had met with Naresh K. Sehgal decades ago during their undergraduate studies at Punjab Engineering College, Chandigarh in India. Since then, Naresh's career path took him through Chip design and later on a journey through Cloud Computing. The amalgamation of their respective work experiences has resulted in this book in your hands. It would not have been possible without the guidance and inspiration of Prof. PCP Bhatt, to whom the authors would like to dedicate this book.

It starts with an introduction to Machine Learning in Chap. 1 and laying down a deeper foundation of ML algorithms in the Chap. 2. Then Chap. 3 serves as a bridge to the Cloud, making a case for using the abundant compute and storage availability for training and inference purposes in the context of Deep Learning. Chapter 4 further explains some basic concepts of Cloud Computing with emphasis on the key characteristics that differentiate it from the enterprise computing, or even running AI algorithms on a laptop. Chapter 5 expands the usage of Cloud for Machine Learning by enumerating its data pipeline stages. Chapter 6 touches on a very important aspect of security in the Cloud for AI and ML algorithms as well as datasets. Chapter 7 delves into some practical aspects of running ML in Amazon's Cloud setup. Chapter 8 gives an example of using Cloud for health care-based AI and ML solutions. Lastly, in the Chap. 9, we look at efforts underway to speed up AI and ML using various hardware-based solutions. No engineering book can be complete without some practical problems and solutions. To meet that expectation, we present three real-life projects that Pramod's students had implemented using Python in Appendices A through C. For each of these, migration of these projects' code to a commercial Public Cloud is illustrated for the reader to practice. Appendix D has solutions to various Points to Ponder, which were posed at the end of each chapter. The motivation here is for the reader to think and then compare one's own answers with our proposed solutions. It can also be the basis of discussion in a classroom setting. The book wraps up with additional questions in Appendix E, the answers to which we leave for the readers to complete.

As with any major project, writing this book took months of planning and over a year to execute it. Even though only two co-authors are listed, there was a major contribution by Prof. PCP Bhatt, who met and reviewed our progress every week for over a year. In addition, we are thankful to our colleague, Aditya Srinivasan, who wrote two sections in Chap. 8 on Multi-Cloud Solutions and UCSD Antibioqram case study. We also used coding examples of several students from Pramod's classes in the Appendices. Naresh picked some ideas and sections from his earlier books with Dr. John M. Acken and Prof. Bhatt on Cloud Computing and Security. NovaGuide View application developed by Shiv Shankar has been used to illustrate NovaSignal's growing presence in the Cloud. Needless to say, several other resources and sites were used to learn and leverage educational material that has been duly acknowledged in our reference sections.

We sincerely hope that readers will have as much fun reading it as we had in writing the varied material in this book. We accept ownership for all the mistakes in this book, but please don't miss sending these to us for corrections, in addition to any suggestions for a future edition. If you include at least part of the sentence the error appears in, that makes it easy for us to search. Page and section numbers are fine, too. Thanks!

San Jose, CA, USA
Santa Clara, CA, USA

Pramod Gupta
Naresh Kumar Sehgal

About This Book

Objective

The purpose of this book is to introduce Machine Learning and Cloud Computing, both from a conceptual level and its Usages with underlying infrastructure. The focus areas of this book include Best Practices for using AI and ML in a Dynamic Infrastructure with Cloud Computing and high Security.

Target audiences are as follows:

1. Senior UG students who have studied programming languages, operating systems.
2. Senior UG and PG Students in software engineering or Information Technology Disciplines.
3. SW developers engaged in migrating in-house ML applications to Public Cloud with Security.
4. Information Technology managers for improving AI/ML performance in Cloud with Security.
5. Professionals who want to learn about the ML, Cloud, Security and technologies behind them.

Level of the book: Mostly at the senior UG or first semester of PG in software engineering, data science, Machine Learning or IT systems.

Contents

Part I Theory

1	Machine Learning Concepts	3
1.1	Introduction	3
1.2	Terminology	4
1.3	What is Machine Learning?	5
1.3.1	Mitchell's Notion of Machine Learning	6
1.4	Basic Differences Between ML and Traditional Programming	9
1.5	How Do Machines Learn?	10
1.6	Steps To Apply ML	12
1.7	A Brief History of Machine Learning	16
1.8	Paradigms of Learning	17
1.8.1	Supervised Machine Learning	17
1.8.2	Unsupervised Machine Learning	19
1.8.3	Reinforcement Machine Learning	20
1.9	Type of Problems in Machine Learning	21
1.9.1	Classification Problem	22
1.9.2	Regression Problem	22
1.9.3	Clustering	23
1.10	Machine Learning in Practice	23
1.11	Why Use Machine Learning?	25
1.12	How to Choose the Right Algorithm?	27
1.13	Why Now?	29
1.14	Applications of Machine Learning	29
1.14.1	Applications from Day-to-Day Life	31
1.14.2	Usage of ML Algorithms	31
1.15	Computing Requirements	33
1.16	What Tools Are Used in Machine Learning?	34
1.17	Machine Learning is Not Perfect	35
1.18	Best Practices	36

1.19	Challenges and Limitations of Machine Learning	37
1.20	What is the Future of Machine Learning?	41
1.21	Summary	42
1.22	Points to Ponder	42
1.23	Answers	42
	References	43
2	Machine Learning Algorithms	45
2.1	Introduction	45
2.2	Supervised Machine Learning Algorithms	46
2.2.1	Regression	46
2.2.2	Classification	47
2.3	Machine Learning Algorithms that Use Supervised Learning	47
2.3.1	Unsupervised Machine Learning Algorithms	47
2.3.2	Clustering	48
2.3.3	Dimension Reduction	48
2.3.4	Anomaly Detection	48
2.4	Machine Learning Algorithms that Use Unsupervised Learning	48
2.5	Considerations When Choosing an Algorithm	49
2.6	Scikit-Learn	51
2.6.1	Consistency	52
2.6.2	Inspection	52
2.6.3	Limited Object Hierarchy	52
2.6.4	Composition	52
2.6.5	Sensible Defaults	52
2.6.6	What Are the Features?	53
2.6.7	Why Use Scikit-Learn for Machine Learning?	53
2.6.8	How Data is Represented in Scikit-Learn?	54
2.6.9	Basics of the Scikit-Learn API	56
2.7	Performance Metrics of ML Algorithms	57
2.7.1	Performance Metrics for Classification Models	58
2.7.2	Regression Metrics	62
2.8	What are the Most Common and Popular Machine Learning Algorithms?	64
2.8.1	Linear Regression (Supervised Learning/Regression)	64
2.8.2	K-Nearest Neighbors (KNN) (Supervised Learning)	78
2.8.3	Logistic Regression (Supervised Learning—Classification)	96
2.8.4	Naïve Bayes Classifier Algorithm (Supervised Learning—Classification)	100
2.8.5	Support Vector Machine Algorithm	119

2.8.6	Decision Trees (Supervised Learning—Classification/ Regression)	122
2.8.7	Ensemble Learning	136
2.8.8	Random Forests (Supervised Learning—Classification/ Regression)	137
2.8.9	K-Means Clustering Algorithm (Unsupervised Learning—Clustering)	160
2.8.10	Artificial Neural Networks (Supervised Learning)	172
2.9	Summary	174
2.10	Points to Ponder	175
2.11	Answers	175
	References	176
3	Deep Learning and Cloud Computing	177
3.1	Introduction	177
3.2	Deep Learning (DL)	178
3.3	Historical Trends	180
3.4	How Do Deep Learning Algorithm Learn?	180
3.5	Architectures	182
3.5.1	Deep Neural Network (DNN)	183
3.5.2	Recurrent Neural Network (RNN)	185
3.5.3	Convolution Neural Networks (CNN)	186
3.6	Choosing a Network	188
3.7	Deep Learning Development Flow	188
3.8	What is Deep About Deep Learning?	188
3.9	Data Used for Deep Learning	189
3.10	Difference Between Machine Learning and Deep Learning	190
3.11	Why Deep Learning Became Popular Now?	190
3.12	Should You Always Use Deep Learning Instead of Machine Learning?	192
3.13	Why is Deep Learning Important?	193
3.14	What Are the Drawbacks of Deep Learning?	193
3.15	Which Deep Learning Software Frameworks Are Available?	194
3.16	Classical Problems Deep Learning Solves	195
3.16.1	Image Classification	195
3.16.2	Natural Language Processing	196
3.17	Summary	197
3.18	Points to Ponder	197
3.19	Answers	197
	References	198

4	Cloud Computing Concepts	201
4.1	Roots of Cloud Computing	201
4.2	Key Characteristics of Cloud Computing	202
4.3	Various Cloud Stakeholders	204
4.4	Pain Points in Cloud Computing	205
4.5	AI and ML in Cloud	206
4.6	Expanding Cloud Reach	209
4.7	Future Trends	210
4.8	Summary	212
4.9	Points to Ponder	212
4.10	Answers	213
	References	213
5	Information Security and Cloud Computing	215
5.1	Information Security Background and Context	215
5.1.1	Privacy Issues	217
5.1.2	Security Concerns of Cloud Operating Models	218
5.1.3	Secure Transmissions, Storage and Computation	220
5.1.4	A Few Key Challenges Related to Cloud Computing and Virtualization	221
5.1.5	Security Practices for Cloud Computing	221
5.1.6	Role of ML for Cybersecurity	223
5.2	Summary	226
5.3	Points to Ponder	226
5.4	Answers	226
	References	227
6	Trust and Security in a Cloud Environment	229
6.1	General Information Security and Trust Concepts	229
6.2	Key Characteristics of Information Security Attack Models	229
6.3	Trust Model Characteristics	232
6.4	Attack Models Addressed by Trust Evaluation	237
6.5	Attack Models on the Trust Evaluation System	237
6.6	Special Issues for Trust Relative to ML	238
6.7	Some Tools for Security Assessment and Tracking	239
6.8	Future Trends	240
6.9	Summary	240
6.10	Related Material	241
6.11	Points to Ponder	242
6.12	Answers	242
	References	245

7	Hardware Based AI and ML	247
7.1	Revisiting History of AI	247
7.2	Current Limitations of AI and ML	248
7.3	Emergence of AI Hardware Accelerators	249
7.3.1	Use of GPUs	250
7.3.2	Use of FPGAs	251
7.3.3	Dedicated AI Accelerators Using ASICs	252
7.3.4	Cerebras’s Wafer Scale AI Engine	256
7.3.5	Google Cloud TPUs	259
7.3.6	Workload Mapping to Different Types of Hardware	260
7.3.7	Amazon’s Inference Engine	261
7.3.8	Intel’s Habana Products	262
7.3.9	Intel’s Movidius VPU	264
7.3.10	Apple’s Image Processing	265
7.4	Platform Based AI	267
7.5	Summary	267
7.6	Points to Ponder	268
7.7	Answers	268
	References	269
8	Hardware Based Security	271
8.1	Introduction	271
8.2	Supply Chain Security in the Cloud	271
8.3	Hardware Elements that Support Security	272
8.4	Characterizing Hardware to Support Security	275
8.5	Future Work and Research Opportunities	275
8.6	Summary	276
8.7	Points to Ponder	276
8.8	Answers	276
	References	277
 Part II Practices		
9	Practical Aspects in Machine Learning	281
9.1	Introduction	281
9.2	Preprocessing Data	281
9.3	Challenges in Data Preparation	284
9.4	When to Use Data Preprocessing?	285
9.4.1	Advantages and Benefits	286
9.5	Framework for Data Preparation Techniques	286
9.5.1	Data Preparation	288
9.5.2	Data Selection (Aka Feature Selection)	288

9.5.3	Data Preprocessing	288
9.5.4	Data Cleaning	289
9.5.5	Insufficient Data	289
9.5.6	Non-representative Data	289
9.5.7	Substandard Data	289
9.5.8	Data Transformation	290
9.5.9	Handling Missing Values	290
9.6	Modification of Categorical or Text Values to Numerical Values	294
9.6.1	One Hot Encode with Scikit-Learn	295
9.6.2	Label Encoding	296
9.6.3	Frequency Encoding	297
9.7	Feature Scaling/Normalizing Values	297
9.7.1	Techniques of Feature Scaling	298
9.8	Inconsistent Values	301
9.9	Duplicated Values	301
9.10	Low Variation Data	303
9.11	Irrelevant Data	303
9.12	Standardized Capitalization	303
9.13	Outliers	303
9.14	Date and Time Features	304
9.14.1	Extracting Date Components	304
9.14.2	Time Since a Reference Point	304
9.14.3	Periodicity and Cyclical Encoding	305
9.14.4	Time-Based Aggregations	306
9.15	Feature Aggregation	306
9.16	Feature Sampling	307
9.16.1	Sampling Without Replacement	307
9.16.2	Sampling with Replacement	307
9.17	Multicollinearity and Its Impact	308
9.18	Feature Selection	308
9.18.1	Importance of Feature Selection	309
9.18.2	How Many Features to Have in the Model?	310
9.18.3	Types of Feature Selection	310
9.19	Dimensionality Reduction	317
9.19.1	Principal Component Analysis (PCA)	318
9.19.2	Linear Discriminant Analysis	320
9.19.3	T-Distributed Stochastic Neighbor Embedding (t-SNE)	321
9.20	Dealing with Imbalanced Data	321
9.20.1	Use the Right Evaluation Metrics	322
9.20.2	Sampling Based Approaches	323

9.21	Evaluating the Impact of Feature Engineering	326
9.22	How is Data Preprocessing Used?	327
9.23	Summary	327
9.24	Points to Ponder	327
9.25	Answers	328
	References	329
10	Analytics in the Cloud	331
10.1	Background	331
10.2	Analytics Services in the Cloud	332
10.3	Introduction to MapReduce	335
10.4	Introduction to Hadoop	336
10.5	Examples of Cloud Based ML	337
10.5.1	Cloud Security Monitoring Using AWS	337
10.5.2	Greener Energy Future with ML in GCP	339
10.5.3	Monorail Monitoring in Azure	341
10.5.4	Detecting Online Hate Speech Using NLP	343
10.6	Future Possibilities	346
10.7	Summary	348
10.8	Points to Ponder	348
10.9	Answers	349
	References	350
11	Healthcare in the Cloud: A Few Case Studies	351
11.1	Introduction	351
11.2	Existing TCD Solution	352
11.3	Trail of Bubbles	354
11.4	Moving Data to the Cloud	355
11.5	A Reader in the Cloud	356
11.6	Heart Care Data in Cloud	357
11.7	Cloud Based Collaborative Tools	359
11.8	Multi Cloud Solutions	362
11.9	UCSD Antibigram: Using Unclassifiable Data	363
11.10	Next Steps	366
11.11	Summary	367
11.12	Points to Ponder	368
11.13	Answers	368
	References	369
12	Evolution and Risks of LLMs	371
12.1	Introduction	371
12.2	NLP Data Preprocessing	372
12.3	NLP Tasks	372