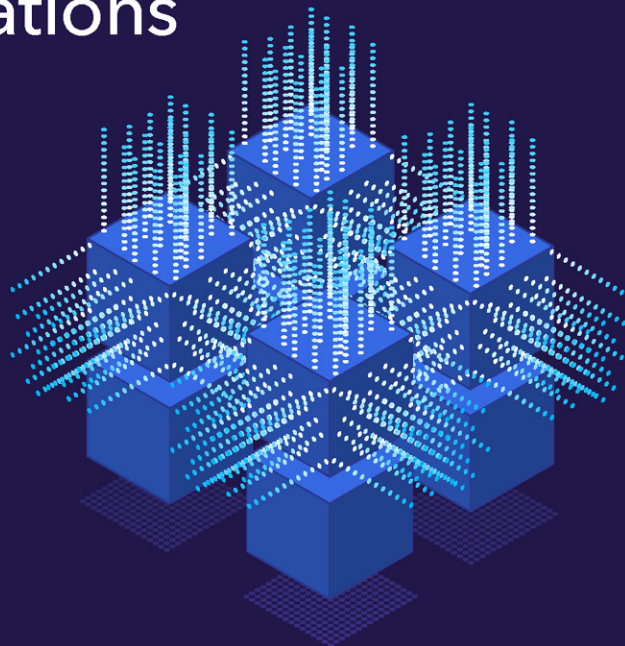


ADDISON WESLEY DATA & ANALYTICS SERIES



MACHINE LEARNING IN PRODUCTION

Developing and Optimizing
Data Science Workflows and
Applications



ANDREW KELLEHER | ADAM KELLEHER

Machine Learning in Production

Machine Learning in Production: Developing and Optimizing Data Science Workflows and Applications

Table of Contents

Cover

Half Title

Title Page

Copyright Page

Dedication

Contents

Foreword

Preface

About the Authors

I: Principles of Framing

1 The Role of the Data Scientist

1.1 Introduction

1.2 The Role of the Data Scientist

1.2.1 Company Size

1.2.2 Team Context

1.2.3 Ladders and Career Development

1.2.4 Importance

1.2.5 The Work Breakdown

1.3 Conclusion

2 Project Workflow

2.1 Introduction

Table of Contents

2.2 The Data Team Context

2.2.1 Embedding vs. Pooling Resources

2.2.2 Research

2.2.3 Prototyping

2.2.4 A Combined Work ow

2.3 Agile Development and the Product Focus

2.3.1 The 12 Principles

2.4 Conclusion

3 Quantifying Error

3.1 Introduction

3.2 Quantifying Error in Measured Values

3.3 Sampling Error

3.4 Error Propagation

3.5 Conclusion

4 Data Encoding and Preprocessing

4.1 Introduction

4.2 Simple Text Preprocessing

4.2.1 Tokenization

4.2.2 N-grams

4.2.3 Sparsity

4.2.4 Feature Selection

4.2.5 Representation Learning

4.3 Information Loss

4.4 Conclusion

5 Hypothesis Testing

5.1 Introduction

5.2 What Is a Hypothesis?

5.3 Types of Errors

5.4 P-values and Confidence Intervals

5.5 Multiple Testing and P-hacking

Table of Contents

5.6 An Example

5.7 Planning and Context

5.8 Conclusion

6 Data Visualization

6.1 Introduction

6.2 Distributions and Summary Statistics

6.2.1 Distributions and Histograms

6.2.2 Scatter Plots and Heat Maps

6.2.3 Box Plots and Error Bars

6.3 Time-Series Plots

6.3.1 Rolling Statistics

6.3.2 Auto-Correlation

6.4 Graph Visualization

6.4.1 Layout Algorithms

6.4.2 Time Complexity

6.5 Conclusion

II: Algorithms and Architectures

7 Introduction to Algorithms and Architectures

7.1 Introduction

7.2 Architectures

7.2.1 Services

7.2.2 Data Sources

7.2.3 Batch and Online Computing

7.2.4 Scaling

7.3 Models

7.3.1 Training

7.3.2 Prediction

7.3.3 Validation

7.4 Conclusion

8 Comparison

8.1 Introduction

Table of Contents

8.2 Jaccard Distance

- 8.2.1 The Algorithm
- 8.2.2 Time Complexity
- 8.2.3 Memory Considerations
- 8.2.4 A Distributed Approach

8.3 MinHash

- 8.3.1 Assumptions
- 8.3.2 Time and Space Complexity
- 8.3.3 Tools
- 8.3.4 A Distributed Approach

8.4 Cosine Similarity

- 8.4.1 Complexity
- 8.4.2 Memory Considerations
- 8.4.3 A Distributed Approach

8.5 Mahalanobis Distance

- 8.5.1 Complexity
- 8.5.2 Memory Considerations
- 8.5.3 A Distributed Approach

8.6 Conclusion

9 Regression

9.1 Introduction

- 9.1.1 Choosing the Model
- 9.1.2 Choosing the Objective Function
- 9.1.3 Fitting
- 9.1.4 Validation

9.2 Linear Least Squares

- 9.2.1 Assumptions
- 9.2.2 Complexity
- 9.2.3 Memory Considerations
- 9.2.4 Tools
- 9.2.5 A Distributed Approach
- 9.2.6 A Worked Example

Table of Contents

9.3 Nonlinear Regression with Linear Regression

9.3.1 Uncertainty

9.4 Random Forest

9.4.1 Decision Trees

9.4.2 Random Forests

9.5 Conclusion

10 Classification and Clustering

10.1 Introduction

10.2 Logistic Regression

10.2.1 Assumptions

10.2.2 Time Complexity

10.2.3 Memory Considerations

10.2.4 Tools

10.3 Bayesian Inference, Naive Bayes

10.3.1 Assumptions

10.3.2 Complexity

10.3.3 Memory Considerations

10.3.4 Tools

10.4 K-Means

10.4.1 Assumptions

10.4.2 Complexity

10.4.3 Memory Considerations

10.4.4 Tools

10.5 Leading Eigenvalue

10.5.1 Complexity

10.5.2 Memory Considerations

10.5.3 Tools

10.6 Greedy Louvain

10.6.1 Assumptions

10.6.2 Complexity

10.6.3 Memory Considerations

10.6.4 Tools

Table of Contents

10.7 Nearest Neighbors

10.7.1 Assumptions

10.7.2 Complexity

10.7.3 Memory Considerations

10.7.4 Tools

10.8 Conclusion

11 Bayesian Networks

11.1 Introduction

11.2 Causal Graphs, Conditional Independence, and Markovity

11.2.1 Causal Graphs and Conditional Independence

11.2.2 Stability and Dependence

11.3 D-separation and the Markov Property

11.3.1 Markovity and Factorization

11.3.2 D-separation

11.4 Causal Graphs as Bayesian Networks

11.4.1 Linear Regression

11.5 Fitting Models

11.6 Conclusion

12 Dimensional Reduction and Latent Variable Models

12.1 Introduction

12.2 Priors

12.3 Factor Analysis

12.4 Principal Components Analysis

12.4.1 Complexity

12.4.2 Memory Considerations

12.4.3 Tools

12.5 Independent Component Analysis

12.5.1 Assumptions

12.5.2 Complexity

12.5.3 Memory Considerations

12.5.4 Tools

Table of Contents

12.6 Latent Dirichlet Allocation

12.7 Conclusion

13 Causal Inference

13.1 Introduction

13.2 Experiments

13.3 Observation: An Example

13.4 Controlling to Block Non-causal Paths

13.4.1 The G-formula

13.5 Machine-Learning Estimators

13.5.1 The G-formula Revisited

13.5.2 An Example

13.6 Conclusion

14 Advanced Machine Learning

14.1 Introduction

14.2 Optimization

14.3 Neural Networks

14.3.1 Layers

14.3.2 Capacity

14.3.3 Overfitting

14.3.4 Batch Fitting

14.3.5 Loss Functions

14.4 Conclusion

III: Bottlenecks and Optimizations

15 Hardware Fundamentals

15.1 Introduction

15.2 Random Access Memory

15.2.1 Access

15.2.2 Volatility

15.3 Nonvolatile/Persistent Storage

15.3.1 Hard Disk Drives or Spinning Disks

Table of Contents

15.3.2 SSDs

15.3.3 Latency

15.3.4 Paging

15.3.5 Thrashing

15.4 Throughput

15.4.1 Locality

15.4.2 Execution-Level Locality

15.4.3 Network Locality

15.5 Processors

15.5.1 Clock Rate

15.5.2 Cores

15.5.3 Threading

15.5.4 Branch Prediction

15.6 Conclusion

16 Software Fundamentals

16.1 Introduction

16.2 Paging

16.3 Indexing

16.4 Granularity

16.5 Robustness

16.6 Extract, Transfer/Transform, Load

16.7 Conclusion

17 Software Architecture

17.1 Introduction

17.2 Client-Server Architecture

17.3 N-tier/Service-Oriented Architecture

17.4 Microservices

17.5 Monolith

17.6 Practical Cases (Mix-and-Match Architectures)

17.7 Conclusion

18 The CAP Theorem

Table of Contents

18.1 Introduction

18.2 Consistency/Concurrency

18.2.1 Conflict-Free Replicated Data Types

18.3 Availability

18.3.1 Redundancy

18.3.2 Front Ends and Load Balancers

18.3.3 Client-Side Load Balancing

18.3.4 Data Layer

18.3.5 Jobs and Taskworkers

18.3.6 Failover

18.4 Partition Tolerance

18.4.1 Split Brains

18.5 Conclusion

19 Logical Network Topological Nodes

19.1 Introduction

19.2 Network Diagrams

19.3 Load Balancing

19.4 Caches

19.4.1 Application-Level Caching

19.4.2 Cache Services

19.4.3 Write-Through Caches

19.5 Databases

19.5.1 Primary and Replica

19.5.2 Multimaster

19.5.3 A/B Replication

19.6 Queues

19.6.1 Task Scheduling and Parallelization

19.6.2 Asynchronous Process Execution

19.6.3 API Buffering

19.7 Conclusion

Bibliography

Table of Contents

Index