

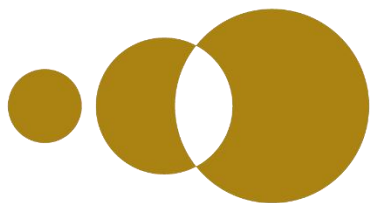
Data Analytics

Interactive Visualization, Probabilistic Modeling,
and Machine Learning with Mathematica



Poomjai Nacaskul, Ph.D. | ดร.พุมใจ นาคสกุล

Applied Digital Intelligence – Chulalongkorn School of Integrated Innovation
Chulalongkorn University



CSII

**Chulalongkorn School of
Integrated Innovation**

BAScii

**Bachelor of Arts and Science in
Integrated Innovation**

Data Analytics: Interactive Visualization, Probabilistic Modeling, and Machine Learning with Mathematica

Poomjai Nacaskul, Ph.D. | ดร.พุมใจ นาคสกุล

Faculty, Applied Digital Intelligence
Chulalongkorn School of Integrated Innovation
Chulalongkorn University



© 2024 Poomjai Nacaskul

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

ISBN: 978-616-616-147-2

Published by
Chulalongkorn University Press
Chulalongkorn University
Chamchuri 6 Building
254 Phyathai Road, Pathumwan
Bangkok 10330, Thailand

Printed in Thailand
First Edition

Chapter 1 - Intro to Data Analytics	5
1. <i>What, Why, When, Where, How</i>	5
1.1. <i>What</i> Data Analytics?	5
1.1.1. But what <i>is</i> Data?	5
1.1.2. And what <i>about</i> Analytics?	6
1.2. <i>Why</i> Data Analytics?	7
1.2.1. Why Now?	7
1.2.2. Why this Book?	7
1.3. <i>When/Where/How</i> Data Analytics?	8
1.3.1. When to use <i>any</i> Data Analytics?	8
1.3.2. When to use <i>which</i> Data Analytics?	9
1.3.3. How to use <i>whichever</i> Data Analytics?	9
2. Data Analytics in Context	9
2.1. Data, Datasets vs. Data Points	9
2.1.1. Data Set	9
2.1.2. Data Points	10
2.2. Data vs. Numbers	11
2.2.1. What are Numbers?	11
2.2.2. Numbers in Context of Data Analytics?	12
2.2.3. Dimensionality of Numbers	14
2.2.4. Generalising the Notion of Data	16
2.3. Tabular vs. Graph Data	17
2.3.1. Tabular Data	17
2.3.2. Graph Data	17
2.4. Data Analytics vs. Machine Learning vs. Artificial Intelligence	18
2.4.1. Data Analytics vs. Machine Learning	18
2.4.2. Machine Learning vs. Artificial Intelligence	18
2.4.3. Artificial vs. Business vs. Computational vs. Digital Intelligence	18
2.4.4. <i>Engineered</i> Intelligence, <i>Intelligence Engine</i> , and <i>Intelligence Engineering</i>	19
3. Intro to Machine Learning	21
3.1. What is Machine Learning?	21
3.1.1. What do we mean by <i>Machine</i> ?	21
3.1.2. What do we mean by <i>Learning</i> ?	22
3.1.3. What constitutes a <i>Learning Machine</i> ?	22
3.1.4. What thus defines Machine Learning?	22
3.2. What are Machine Learning Paradigms?	22
3.2.1. <i>Unsupervised</i> Learning	22
3.2.2. <i>Supervised</i> Learning	23
3.2.3. <i>Reinforcement</i> Learning	23
3.2.4. <i>Representation</i> Learning	23

Chapter 2 - Intro to Wolfram Mathematica.....	24
1. Intro to Mathematica (the Scientific Computation <i>Platform</i>).....	24
1.1. Evolution of Mathematica.....	24
1.1.1. Mathematica 1.0.....	24
1.1.2. Mathematica Notebook.....	26
1.1.3. Mathematica 14.....	28
1.2. Mathematica in Context.....	28
1.2.1. Programming Paradigm.....	28
1.2.2. Wolfram's "Computational Intelligence" Platform.....	29
1.2.3. Science, Engineering & Industry.....	31
1.2.4. Software Ecosystem.....	32
2. Intro to Wolfram (the Mathematica Programming <i>Language</i>).....	36
2.1. Language Syntax/Programming Primitives.....	36
2.1.1. Assignment.....	36
2.1.2. List.....	37
2.1.3. Function.....	38
2.1.4. Univariate Function.....	39
2.1.5. Multivariate Function.....	44
2.1.6. List/Array Construction.....	50
2.1.7. Association.....	53
2.1.8. Useful Operations on/using List/Association.....	56
2.1.9. Pure Functions.....	59
2.1.10. User-defined Functions.....	61
2.1.11. Defining Scope of Variable Definitions with Module[]	62
2.1.12. Function Shorthand.....	64
2.1.13. Handling Tabular Data with Dataset[]	66
2.1.14. Files & Storage.....	68
2.2. Analytical/Computational/Algorithmic Functionalities.....	68
2.2.1. Symbolic Computation Engine.....	68
2.2.2. Analytical Computation Engine.....	70
2.2.3. Numerical Computation Engine.....	73
2.2.4. Semi-Analytical Computation Engine.....	74

Chapter 3 - Interactive Visualisation.....76

1. Static Visualisation.....76

1.1. Visualising Parametric Objects.....76

1.1.1. Visualising Mathematical Functions.....76

1.1.2. Visualising Geometric Objects.....76

1.2. Visualising Data Points.....77

1.2.1. Visualising 1D Data.....77

1.2.2. Visualising 2D Data.....77

1.2.3. Visualising 3D Data.....79

1.3. Visualising Histogram.....79

1.3.1. Visualising 1D Data on 2D Histogram.....79

1.3.2. Visualising 2D Data on 3D Histogram.....81

2. Interactive Visualisation.....83

2.1. Interactive User Interface.....83

2.1.1. Single Control.....83

2.1.2. Multiple Controls.....84

2.2. Encapsulating Interactive Design.....89

2.2.1. Static Visualisation of Prototype with Initial Design Parameters.....89

2.2.2. Encapsulate Parametric Design as a Parameterised Function.....90

2.2.3. Enable Interactive Control over the Design Parameters.....92

Chapter 4 - Probabilistic Modelling.....95

1. Descriptive Statistics.....95

1.1. Statistical Moment.....95

1.1.1. Moments vs. Central Moments.....95

1.1.2. Measure of Central Tendency.....96

1.1.3. Measure of Dispersion.....98

1.1.4. Measure of Asymmetry.....100

1.1.5. Measure of Heavy-Tailedness.....101

1.2. Scaling/Normalisation.....101

1.2.1. Standardise by Subtraction and Division.....101

1.2.2. Dividing by Vector Norm.....102

1.2.3. Distributional Rescaling.....103

2. Probability Theory	104
2.1. Foundation of Probability Theory	104
2.1.1. Probability Axioms	104
2.1.2. Random Variable ...	105
2.1.3. Support of Probability Distribution	106
2.2. Probability Distribution Function	107
2.2.1. Probability Mass/Density Function & Statistical Moments	107
2.2.2. Probability & Quantiles	109
2.2.3. Cumulative Distribution Function (CDF) & Inverse CDF	109
2.2.4. Generating Random Variates	111
2.3. Univariate Probability Distribution	112
2.3.1. Finite (Discrete) Families	112
2.3.2. Countable (Discrete) Families	113
2.3.3. Bounded (Continuous) Families	114
2.3.4. Semi-Bounded (Continuous) Families	115
2.3.5. Unbounded (Continuous) Families	118
2.4. Multivariate Probability Distributions	121
2.4.1. Families of Multivariate Probability Distributions	121
2.4.2. Copula	127
3. Inferential Statistics	133
3.1. Parametric Estimation	133
3.1.1. Point Estimation	133
3.1.2. Confidence Interval	134
3.2. Hypothesis Testing	135
3.2.1. Null Hypothesis & Test Statistics	135
3.2.2. Normality & Distribution Fit Tests	136
4. Nonparametric Statistics	138
4.1. Definition	138
4.1.1. What makes Parametric Statistics Parametric?	138
4.1.2. What makes Nonparametric Statistics Nonparametric?	138
4.1.3. What makes Semiparametric Statistics Semiparametric?	138
4.2. Nonparametric/Semiparametric Distribution Fit	139
4.2.1. (Nonparametric) Empirical Distribution	139
4.2.2. (Semiparametric) Kernel Method	139
4.3. Data Mining	141
4.3.1. Frequent Itemset Mining	141
4.3.2. Association Rule Mining	146
4.3.3. Frequent Sequence Mining	146
4.3.4. Sequential Rule Mining	146

Chapter 5 - Unsupervised Machine Learning

	147
1. Unsupervised Machine Learning - Cluster Analysis	147
1.1. Modes of Cluster Analysis	147
1.1.1. "Once-and-Done"	147
1.1.2. "Learn-the-Partition"	147
1.2. Cluster Analysis Methods	148
1.2.1. Hierarchical Branching	148
1.2.2. Distance-based Cluster Analysis	148
1.2.3. Density-based Cluster Analysis	149
1.2.4. Hybrid/Nearest-Neighbour Clustering	150
1.3. Demonstration on Example Data	151
1.3.1. on Randomly Generated Hues	151
1.3.2. on "Old Faithful" 2D Data	153
1.3.3. on Generated "Quad-Mode" 2D Data	155
1.4. Cluster Analysis - Data Analytics Pipeline	157
1.4.1. Components of Cluster Analysis Pipeline	157
1.4.2. Encapsulate the Cluster Analysis Pipeline	163
2. Unsupervised Machine Learning - Dimensionality Reduction	168
2.1. Modes of Analysis	168
2.1.1. "Once-and-Done"	168
2.1.2. "Learn-the-Mapping"	168
2.2. Dimensionality Reduction Methods	168
2.2.1. Linear Dimensionality Reduction Method	168
2.2.2. Nonlinear Dimensionality Reduction Method	168
2.2.3. Dimensionality Reduction on Generated 3D Data w/ Hidden Labels	168
2.3. Cluster Analysis together with Dimensionality Reduction	170
2.3.1. (<i>Don't Do This!</i>) Dimensionality Reduction on Cluster-Partitioned Datasets	170
2.3.2. (OK, But Still <i>Not</i> Recommended) Cluster Analysis on Dimensionality-Reduced Dataset	170
2.3.3. (Generally Recommended Practice) Dimensionality-Reduced Visualisation of Cluster-Partitioned Datasets	171
2.4. Cluster Analysis together with Mixture Distribution Fit	172
2.4.1. Mixture Distribution Fit on 1D Data Partitioned into Clusters	172
2.4.2. Encapsulate/Visualise Mixture Distribution Fit on 1D Data Partitioned into Clusters	173

Chapter 6 - Supervised Machine Learning.....175

1. Supervised Machine Learning - Problems.....	175
1.1. Prediction (Regression/Forecasting) Problem.....	175
1.1.1. What is a Prediction Problem (in the Context of Supervised Machine Learning)?.....	175
1.1.2. Regression as Numerical Prediction Problem.....	175
1.2. Classification (Identification/Discriminant) Problem.....	178
1.2.1. What is a Classification Problem (in the Context of Supervised Machine Learning)?.....	178
1.2.2. Logistic Regression as Binary Classification Problem.....	178
2. Supervised Machine Learning – Algorithms.....	186
2.1. Unitary Supervised Learning Machines.....	186
2.1.1. Supervised Learning Trees.....	186
2.1.2. Artificial Neural Network (ANN).....	187
2.2. Composite Supervised Machine Learning.....	190
2.2.1. Ensemble Supervised Learning Trees - Boosting and Bagging.....	190
2.2.2. Ensemble Neural Machine Learning Architectures.....	190
3. Supervised Machine Learning - Procedure.....	191
3.1. Training Machine Learners.....	191
3.1.1. Partition Dataset.....	191
3.1.2. Input-Output Format.....	192
3.1.3. Running the Machine Learning Algorithm on Training Data.....	194
3.2. Assessing Machine Learning Performance.....	195
3.2.1. Predictor Performance Measurements.....	195
3.2.2. Classifier Performance Measurements.....	197
3.3. Attempting to Understand Learned Machines.....	200
3.3.1. Model-Specific.....	200
3.3.2. Model-Agnostic.....	200

Chapter 7 - Analytics on Graph Data	203
1. Graph as Mathematical Representation of Network	203
1.1. Definition	203
1.1.1. Graph, Vertex (Node), Edge (Arc)	203
1.1.2. Graph vs. Network	204
1.1.3. Graph Data	204
1.1.4. Graph Plot, Graph Layout, Graph Embedding	205
1.1.5. Unweighted (Boolean) vs. Weighted (Ordinal/Interval/Ratio-Scale) Edges/Graphs	207
1.1.6. Undirected vs. Directed Edges/Graphs	207
1.1.7. Monoplex vs. Multiplex Graphs	208
1.1.8. Adjacency Matrix, Adjacency Tensor	209
1.2. Graph Properties	212
1.2.1. Connected vs. Unconnected Graph	212
1.2.2. Acyclic vs. Cyclic Graph	213
1.2.3. Directedness	214
1.2.4. Sparsity	215
1.3. Graph Visualisation	215
1.3.1. Graph Plot with Arbitrary Graph Layout	215
1.3.2. Graph Plot with Specified Vertex Coordinates	216
1.3.3. Graph Communities & Community Graph Plot	220
2. Graph-Theoretic (Network-Centric)	
Centrality Analysis	224
2.1. Definition	224
2.1.1. Centrality Analysis = Vector-Valued Function	224
2.1.2. Centrality Analysis = Matrix Computation	224
2.2. Methods	224
2.2.1. "Local" Method	224
2.2.2. "Recursive" Method	224
2.2.3. "Global" Method	224
2.2.4. Demonstration	224
3. Algebraic Computation on Adjacency Matrices	228
3.1. Matrix Power	228
3.1.1. Operation and Limits	228
3.1.2. Application	230
3.2. Inner Operation	232
3.2.1. Dot Product (and by extension Matrix Multiplication) as Special Case of Inner Operation	232
3.2.2. Application	234
Glossary	237
Biography	240



Data Analytics: Interactive Visualization, Probabilistic Modeling, and Machine Learning with Mathematica

© 2024 Poomjai Nacaskul

ISBN: 978-616-616-147-2

Chulalongkorn Univ. Press

Bangkok, Thailand



฿ 480.