
Laura Igual · Santi Seguí

Introduction to Data Science

A Python Approach to Concepts,
Techniques and Applications

With contributions from Jordi Vitrià, Eloi Puertas
Petia Radeva, Oriol Pujol, Sergio Escalera, Francesc Dantí
and Lluís Garrido

Contents

1	Introduction to Data Science	1
1.1	What is Data Science?	1
1.2	About This Book	3
2	Toolboxes for Data Scientists	5
2.1	Introduction	5
2.2	Why Python?	6
2.3	Fundamental Python Libraries for Data Scientists	6
2.3.1	Numeric and Scientific Computation: NumPy and SciPy	7
2.3.2	SCIKIT-Learn: Machine Learning in Python	7
2.3.3	PANDAS: Python Data Analysis Library	7
2.4	Data Science Ecosystem Installation	7
2.5	Integrated Development Environments (IDE)	8
2.5.1	Web Integrated Development Environment (WIDE): Jupyter	9
2.6	Get Started with Python for Data Scientists	10
2.6.1	Reading	14
2.6.2	Selecting Data	16
2.6.3	Filtering Data	17
2.6.4	Filtering Missing Values	17
2.6.5	Manipulating Data	18
2.6.6	Sorting	22
2.6.7	Grouping Data	23
2.6.8	Rearranging Data	24
2.6.9	Ranking Data	25
2.6.10	Plotting	26
2.7	Conclusions	28
3	Descriptive Statistics	29
3.1	Introduction	29
3.2	Data Preparation	30
3.2.1	The Adult Example	30

3.3	Exploratory Data Analysis	32
3.3.1	Summarizing the Data	32
3.3.2	Data Distributions	36
3.3.3	Outlier Treatment	38
3.3.4	Measuring Asymmetry: Skewness and Pearson's Median Skewness Coefficient	41
3.3.5	Continuous Distribution	42
3.3.6	Kernel Density	44
3.4	Estimation	46
3.4.1	Sample and Estimated Mean, Variance and Standard Scores	46
3.4.2	Covariance, and Pearson's and Spearman's Rank Correlation.	47
3.5	Conclusions	50
	References	50
4	Statistical Inference	51
4.1	Introduction	51
4.2	Statistical Inference: The Frequentist Approach	52
4.3	Measuring the Variability in Estimates.	52
4.3.1	Point Estimates	53
4.3.2	Confidence Intervals	56
4.4	Hypothesis Testing.	59
4.4.1	Testing Hypotheses Using Confidence Intervals	60
4.4.2	Testing Hypotheses Using p -Values	61
4.5	But Is the Effect E Real?	64
4.6	Conclusions	64
	References	65
5	Supervised Learning.	67
5.1	Introduction	67
5.2	The Problem	68
5.3	First Steps	69
5.4	What Is Learning?	78
5.5	Learning Curves.	79
5.6	Training, Validation and Test.	82
5.7	Two Learning Models	86
5.7.1	Generalities Concerning Learning Models	86
5.7.2	Support Vector Machines	87
5.7.3	Random Forest	90
5.8	Ending the Learning Process	91
5.9	A Toy Business Case.	92
5.10	Conclusion	95
	Reference	96

6	Regression Analysis	97
6.1	Introduction	97
6.2	Linear Regression	98
6.2.1	Simple Linear Regression	98
6.2.2	Multiple Linear Regression and Polynomial Regression	103
6.2.3	Sparse Model	104
6.3	Logistic Regression	110
6.4	Conclusions	113
	References	114
7	Unsupervised Learning	115
7.1	Introduction	115
7.2	Clustering.	116
7.2.1	Similarity and Distances	117
7.2.2	What Constitutes a Good Clustering? Defining Metrics to Measure Clustering Quality	117
7.2.3	Taxonomies of Clustering Techniques	120
7.3	Case Study.	132
7.4	Conclusions	138
	References	139
8	Network Analysis	141
8.1	Introduction	141
8.2	Basic Definitions in Graphs	142
8.3	Social Network Analysis	144
8.3.1	Basics in NetworkX	144
8.3.2	Practical Case: Facebook Dataset	145
8.4	Centrality	147
8.4.1	Drawing Centrality in Graphs	152
8.4.2	PageRank	154
8.5	Ego-Networks	157
8.6	Community Detection	162
8.7	Conclusions	163
	References	164
9	Recommender Systems	165
9.1	Introduction	165
9.2	How Do Recommender Systems Work?	166
9.2.1	Content-Based Filtering	166
9.2.2	Collaborative Filtering	167
9.2.3	Hybrid Recommenders	167
9.3	Modeling User Preferences	167
9.4	Evaluating Recommenders	168

9.5	Practical Case	169
9.5.1	MovieLens Dataset	169
9.5.2	User-Based Collaborative Filtering	171
9.6	Conclusions	179
	References	179
10	Statistical Natural Language Processing for Sentiment	
	Analysis	181
10.1	Introduction	181
10.2	Data Cleaning	182
10.3	Text Representation	185
10.3.1	Bi-Grams and n-Grams	190
10.4	Practical Cases	191
10.5	Conclusions	196
	References	196
11	Parallel Computing	199
11.1	Introduction	199
11.2	Architecture	200
11.2.1	Getting Started	201
11.2.2	Connecting to the Cluster (The Engines)	202
11.3	Multicore Programming	203
11.3.1	Direct View of Engines	203
11.3.2	Load-Balanced View of Engines	206
11.4	Distributed Computing	207
11.5	A Real Application: New York Taxi Trips	208
11.5.1	A Direct View Non-Blocking Proposal	209
11.5.2	Results	212
11.6	Conclusions	214
	References	215
	Index	217