



Learning Apache Spark with Python

Wenqiang Feng

September 03, 2019

CONTENTS

1	Preface	3
1.1	About	3
1.2	Motivation for this tutorial	4
1.3	Copyright notice and license info	4
1.4	Acknowledgement	4
1.5	Feedback and suggestions	5
2	Why Spark with Python ?	7
2.1	Why Spark?	7
2.2	Why Spark with Python (PySpark)?	8
3	Configure Running Platform	11
3.1	Run on Databricks Community Cloud	11
3.2	Configure Spark on Mac and Ubuntu	16
3.3	Configure Spark on Windows	19
3.4	PySpark With Text Editor or IDE	19
3.5	PySparkling Water: Spark + H2O	26
3.6	Set up Spark on Cloud	27
3.7	Demo Code in this Section	27
4	An Introduction to Apache Spark	29
4.1	Core Concepts	29
4.2	Spark Components	29
4.3	Architecture	32
4.4	How Spark Works?	32
5	Programming with RDDs	33
5.1	Create RDD	33
5.2	Spark Operations	37
5.3	<code>rdd.DataFrame</code> vs <code>pd.DataFrame</code>	39
6	Statistics and Linear Algebra Preliminaries	55
6.1	Notations	55
6.2	Linear Algebra Preliminaries	55
6.3	Measurement Formula	57
6.4	Confusion Matrix	58

6.5	Statistical Tests	59
7	Data Exploration	61
7.1	Univariate Analysis	61
7.2	Multivariate Analysis	74
8	Regression	81
8.1	Linear Regression	81
8.2	Generalized linear regression	93
8.3	Decision tree Regression	100
8.4	Random Forest Regression	106
8.5	Gradient-boosted tree regression	113
9	Regularization	121
9.1	Ordinary least squares regression	121
9.2	Ridge regression	121
9.3	Least Absolute Shrinkage and Selection Operator (LASSO)	122
9.4	Elastic net	122
10	Classification	123
10.1	Binomial logistic regression	123
10.2	Multinomial logistic regression	134
10.3	Decision tree Classification	145
10.4	Random forest Classification	154
10.5	Gradient-boosted tree Classification	164
10.6	XGBoost: Gradient-boosted tree Classification	164
10.7	Naive Bayes Classification	166
11	Clustering	179
11.1	K-Means Model	179
12	RFM Analysis	191
12.1	RFM Analysis Methodology	192
12.2	Demo	194
12.3	Extension	200
13	Text Mining	207
13.1	Text Collection	207
13.2	Text Preprocessing	215
13.3	Text Classification	217
13.4	Sentiment analysis	224
13.5	N-grams and Correlations	231
13.6	Topic Model: Latent Dirichlet Allocation	231
14	Social Network Analysis	249
14.1	Introduction	249
14.2	Co-occurrence Network	249
14.3	Appendix: matrix multiplication in PySpark	253
14.4	Correlation Network	256

15 ALS: Stock Portfolio Recommendations	257
15.1 Recommender systems	258
15.2 Alternating Least Squares	259
15.3 Demo	259
16 Monte Carlo Simulation	267
16.1 Simulating Casino Win	267
16.2 Simulating a Random Walk	269
17 Markov Chain Monte Carlo	279
17.1 Metropolis algorithm	279
17.2 A Toy Example of Metropolis	280
17.3 Demos	281
18 Neural Network	289
18.1 Feedforward Neural Network	289
19 Wrap PySpark Package	293
19.1 Package Wrapper	293
19.2 Package Publishing on PyPI	295
20 PySpark Data Audit Library	297
20.1 Install with pip	297
20.2 Install from Repo	297
20.3 Uninstall	297
20.4 Test	298
20.5 Auditing on Big Dataset	299
21 Zeppelin to jupyter notebook	309
21.1 How to Install	309
21.2 Converting Demos	310
22 My Cheat Sheet	315
23 PySpark API	319
23.1 Stat API	319
23.2 Regression API	325
23.3 Classification API	346
23.4 Clustering API	369
23.5 Recommendation API	385
23.6 Pipeline API	390
23.7 Tuning API	392
23.8 Evaluation API	397
24 Main Reference	403
Bibliography	405
Python Module Index	407



Welcome to my **Learning Apache Spark with Python** note! In this note, you will learn a wide array of concepts about **PySpark** in Data Mining, Text Mining, Machine Learning and Deep Learning. The PDF version can be downloaded from [HERE](#).

