Learning Apache Spark with Python

Release v1.0

Wenqiang Feng

July 11, 2018

1 Preface **3**
1.1 About . . . 3
1.2 Motivation for this tutorial . . . 4
1.3 Acknowledgement . . . 4
1.4 Feedback and suggestions . . . 4

2 Why Spark with Python ? **5**
2.1 Why Spark? . . . 5
2.2 Why Spark with Python (PySpark)? . . . 7

3 Configure Running Platform **9**
3.1 Run on Databricks Community Cloud . . . 9
3.2 Configure Spark on Mac and Ubuntu . . . 14
3.3 Configure Spark on Windows . . . 17
3.4 PySpark With Text Editor or IDE . . . 17
3.5 Set up Spark on Cloud . . . 19
3.6 Demo Code in this Section . . . 21

4 An Introduction to Apache Spark **23**
4.1 Core Concepts . . . 23
4.2 Spark Components . . . 23
4.3 Architecture . . . 26
4.4 How Spark Works? . . . 26

5 Programming with RDDs **27**
5.1 Create RDD . . . 27
5.2 Spark Operations . . . 31

6 Statistics Preliminary **33**
6.1 Notations . . . 33
6.2 Measurement Formula . . . 33
6.3 Statistical Tests . . . 34

7 Data Exploration **35**
7.1 Univariate Analysis . . . 35
7.2 Multivariate Analysis . . . 35

8 Regression **41**
8.1 Linear Regression . . . 41
8.2 Generalized linear regression . . . 48
8.3 Decision tree Regression . . . 53
8.4 Random Forest Regression . . . 58
8.5 Gradient-boosted tree regression . . . 58

9 Regularization **59**
9.1 Ridge regression . . . 59
9.2 Least Absolute Shrinkage and Selection Operator (LASSO) . . . 59
9.3 Elastic net . . . 59

10 Classification **61**
10.1 Logistic regression . . . 61
10.2 Decision tree Classification . . . 68
10.3 Random forest Classification . . . 75
10.4 Gradient-boosted tree Classification . . . 82
10.5 Naive Bayes Classification . . . 83

11 Clustering **85**
11.1 K-Means Model . . . 85

12 Text Mining **91**
12.1 Text Collection . . . 91
12.2 Text Preprocessing . . . 98
12.3 Text Classification . . . 100
12.4 Sentiment analysis . . . 106
12.5 N-grams and Correlations . . . 112
12.6 Topic Model: Latent Dirichlet Allocation . . . 112

13 Social Network Analysis **125**
13.1 Co-occurrence Network . . . 125
13.2 Correlation Network . . . 130

14 Neural Network **131**
14.1 Feedforward Neural Network . . . 131

15 My PySpark Package **135**
15.1 Hierarchical Structure . . . 135
15.2 Set Up . . . 136
15.3 ReadMe . . . 136

16 Main Reference **139**

Bibliography **141**

Index **143**

Welcome to our **Learning Apache Spark with Python** note! In these note, you will learn a wide array of concepts about **PySpark** in Data Mining, Text Mining, Machine Leanring and Deep Learning. The PDF version can be downloaded from HERE.