Course duration

- 3 days

Course Benefits

- Python essentials
- Capabilities of the Apache Spark platform and its machine learning module
- Terminology, concepts, and algorithms used in machine learning

Course Outline

1. Defining Data Science
    1. Data Science, Machine Learning, AI?
    2. The Data-Related Roles
    3. Data Science Ecosystem
    4. Business Analytics vs. Data Science
    5. Who is a Data Scientist?
    6. The Break-Down of Data Science Project Activities
    7. Data Scientists at Work
    8. The Data Engineer Role
    9. What is Data Wrangling (Munging)?
    10. Examples of Data Science Projects
    11. Data Science Gotchas
    12. Summary
2. Machine Learning Life-cycle Phases
    1. Data Analytics Pipeline
    2. Data Discovery Phase
    3. Data Harvesting Phase
    4. Data Priming Phase
    5. Data Cleansing
    6. Feature Engineering
    7. Data Logistics and Data Governance
    8. Exploratory Data Analysis
    9. Model Planning Phase
    10. Model Building Phase
    11. Communicating the Results
    12. Production Roll-out
    13. Summary
3. Quick Introduction to Python Programming
    1. Module Overview
    2. Some Basic Facts about Python

10. Saving Files
11. Summary

6. Quick Intro to Jupyter Notebooks
    1. Python Dev Tools and REPLs
    2. IPython
    3. Jupyter
    4. Jupyter Operation Modes
    5. Basic Edit Mode Shortcuts
    6. Basic Command Mode Shortcuts
    7. Summary
7. Data Visualization in Python using matplotlib
    1. Data Visualization
    2. What is matplotlib?
    3. Getting Started with matplotlib
    4. The matplotlib.pyplot.plot() Function
    5. The matplotlib.pyplot.scatter() Function
    6. Labels and Titles
    7. Styles
    8. The matplotlib.pyplot.bar() Function
    9. The matplotlib.pyplot.hist () Function
    10. The matplotlib.pyplot.pie () Function
    11. The Figure Object
    12. The matplotlib.pyplot.subplot() Function
    13. Selecting a Grid Cell
    14. Saving Figures to a File
    15. Summary
8. Data Science and ML Algorithms with PySpark
    1. In-Class Discussion
    2. Types of Machine Learning
    3. Supervised vs Unsupervised Machine Learning
    4. Supervised Machine Learning Algorithms
    5. Classification (Supervised ML) Examples
    6. Unsupervised Machine Learning Algorithms
    7. Clustering (Unsupervised ML) Examples
    8. Choosing the Right Algorithm
    9. Terminology: Observations, Features, and Targets
    10. Representing Observations
    11. Terminology: Labels
    12. Terminology: Continuous and Categorical Features
    13. Continuous Features
    14. Categorical Features
    15. Common Distance Metrics
    16. The Euclidean Distance
    17. What is a Model
    18. Model Evaluation
    19. The Classification Error Rate
    20. Data Split for Training and Test Data Sets

21. Data Splitting in PySpark
22. Hold-Out Data
23. Cross-Validation Technique
24. Spark ML Overview
25. DataFrame-based API is the Primary Spark ML API
26. Estimators, Models, and Predictors
27. Descriptive Statistics
28. Data Visualization and EDA
29. Correlations
30. Hands-on Exercise
31. Feature Engineering
32. Scaling of the Features
33. Feature Blending (Creating Synthetic Features)
34. Hands-on Exercise
35. The 'One-Hot' Encoding Scheme
36. Example of 'One-Hot' Encoding Scheme
37. Bias-Variance (Underfitting vs Overfitting) Trade-off
38. The Modeling Error Factors
39. One Way to Visualize Bias and Variance
40. Underfitting vs Overfitting Visualization
41. Balancing Off the Bias-Variance Ratio
42. Linear Model Regularization
43. ML Model Tuning Visually
44. Linear Model Regularization in Spark
45. Regularization, Take Two
46. Dimensionality Reduction
47. PCA and isomap
48. The Advantages of Dimensionality Reduction
49. Spark Dense and Sparse Vectors
50. Labeled Point
51. Python Example of Using the LabeledPoint Class
52. The LIBSVM format
53. LIBSVM in PySpark
54. Example of Reading a File In LIBSVM Format
55. Life-cycles of Machine Learning Development
56. Regression Analysis
57. Regression vs Correlation
58. Regression vs Classification
59. Simple Linear Regression Model
60. Linear Regression Illustration
61. Least-Squares Method (LSM)
62. Gradient Descent Optimization
63. Locally Weighted Linear Regression
64. Regression Models in Excel
65. Multiple Regression Analysis
66. Evaluating Regression Model Accuracy
67. The R>2

68. Model Score
69. The MSE Model Score
70. Hands-on Exercise
71. Linear Logistic (Logit) Regression
72. Interpreting Logistic Regression Results
73. Hands-on Exercise
74. Naive Bayes Classifier (SL)
75. Naive Bayesian Probabilistic Model in a Nutshell
76. Bayes Formula
77. Classification of Documents with Naive Bayes
78. Hands-on Exercise
79. Decision Trees
80. Decision Tree Terminology
81. Properties of Decision Trees
82. Decision Tree Classification in the Context of Information Theory
83. The Simplified Decision Tree Algorithm
84. Using Decision Trees
85. Random Forests
86. Hands-On Exercise
87. Support Vector Machines (SVMs)
88. Hands-On Exercise
89. Unsupervised Learning Type: Clustering
90. k-Means Clustering (UL)
91. k-Means Clustering in a Nutshell
92. k-Means Characteristics
93. Global vs Local Minimum Explained
94. Hands-On Exercise
95. Time-Series Analysis
96. Decomposing Time-Series
97. A Better Algorithm or More Data?
98. Summary

Class Materials

Each student will receive a comprehensive set of materials, including course notes and all the class examples.

Class Prerequisites

Experience in the following *is required* for this Spark class:

- General knowledge of statistics and programming.