Python With Data Science

Course Overview

View Course Dates & Register Today

This is a 2-day class

This course covers theoretical and technical aspects of using Python in Applied Data Science projects and Data Logistics use cases.

Who Should Attend

Data Scientists, Software Developers, IT Architects, and Technical Managers. Participants should have the general knowledge of statistics and programming and also be familiar with Python.

Course Objectives

NumPy, pandas, Matplotlib, scikit-learn;
Python REPLs;
Jupyter Notebooks;
Data analytics life-cycle phases;
Data repairing and normalizing;
Data aggregation and grouping;
Data visualization;
Data science algorithms for supervised and unsupervised; Machine Learning.

Course Outline

1 PYTHON FOR DATA SCIENCE

Using Modules
Listing Methods in a Module
Creating Your Own Modules
List Comprehension
Dictionary Comprehension
String Comprehension
Python 2 vs Python 3
Sets (Python 3+)
Python Idioms
Python Data Science “Ecosystem”
NumPy
NumPy Arrays
NumPy Idioms
pandas
Data Wrangling with pandas' DataFrame
SciPy
Scikit-learn
SciPy or scikit-learn?
Matplotlib
Python vs R
Python on Apache Spark
Python Dev Tools and REPLs
Anaconda
IPython
Visual Studio Code
Jupyter
Jupyter Basic Commands
Summary

Python With Data Science

2 APPLIED DATA SCIENCE

What is Data Science?
Data Science Ecosystem
Data Mining vs. Data Science
Business Analytics vs. Data Science
Data Science, Machine Learning, AI?
Who is a Data Scientist?
Data Science Skill Sets Venn Diagram
Data Scientists at Work
Examples of Data Science Projects
An Example of a Data Product
Applied Data Science at Google
Data Science Gotchas
Summary

3 DATA ANALYTICS LIFE-CYCLE PHASES

Big Data Analytics Pipeline
Data Discovery Phase
Data Harvesting Phase
Data Priming Phase
Data Logistics and Data Governance
Exploratory Data Analysis
Model Planning Phase
Model Building Phase
Communicating the Results
Production Roll-out
Summary

4 REPAIRING AND NORMALIZING DATA

Repairing and Normalizing Data
Dealing with the Missing Data
Sample Data Set
Getting Info on Null Data
Dropping a Column
Interpolating Missing Data in pandas
Replacing the Missing Values with the Mean Value
Scaling (Normalizing) the Data
Data Preprocessing with scikit-learn
Scaling with the scale() Function
The MinMaxScaler Object
Summary

5 DESCRIPTIVE STATISTICS COMPUTING FEATURES IN PYTHON

Descriptive Statistics
Non-uniformity of a Probability Distribution
Using NumPy for Calculating Descriptive Statistics Measures
Finding Min and Max in NumPy
Using pandas for Calculating Descriptive Statistics Measures
Correlation
Regression and Correlation
Covariance
Getting Pairwise Correlation and Covariance Measures
Finding Min and Max in pandas DataFrame
Summary

6 DATA AGGREGATION AND GROUPING

Data Aggregation and Grouping
Sample Data Set
The pandas.core.groupby.SeriesGroupBy Object
Grouping by Two or More Columns
Emulating the SQL's WHERE Clause
The Pivot Tables
Cross-Tabulation
Summary

7 DATA VISUALIZATION WITH MATPLOTLIB

Data Visualization
What is matplotlib?
Getting Started with matplotlib
The Plotting Window
The Figure Options
The matplotlib.pyplot.plot() Function
The matplotlib.pyplot.bar() Function
The matplotlib.pyplot.pie () Function
Subplots
Using the matplotlib.gridspec.GridSpec Object
The matplotlib.pyplot.subplot() Function
Hands-on Exercise
Figures
Saving Figures to File
Visualization with pandas
Working with matplotlib in Jupyter Notebooks
Summary

8 DATA SCIENCE AND ML ALGORITHMS IN SCIKIT-LEARN

Data Science, Machine Learning, AI?
Types of Machine Learning
Terminology: Features and Observations
Continuous and Categorical Features (Variables)
Terminology: Axis
The scikit-learn Package
scikit-learn Estimators
Models, Estimators, and Predictors

Common Distance Metrics
The Euclidean Metric
The LIBSVM format
Scaling of the Features
The Curse of Dimensionality
Supervised vs Unsupervised Machine Learning
Supervised Machine Learning Algorithms
Unsupervised Machine Learning Algorithms
Choose the Right Algorithm
Life-cycles of Machine Learning Development
Data Split for Training and Test Data Sets
Data Splitting in scikit-learn
Hands-on Exercise
Classification Examples
Classifying with k-Nearest Neighbors (SL)
k-Nearest Neighbors Algorithm
k-Nearest Neighbors Algorithm
The Error Rate
Hands-on Exercise
Dimensionality Reduction
The Advantages of Dimensionality Reduction
Principal component analysis (PCA)
Hands-on Exercise
Data Blending
Decision Trees (SL)
Decision Tree Terminology
Decision Tree Classification in Context of Information Theory
Information Entropy Defined
The Shannon Entropy Formula
The Simplified Decision Tree Algorithm
Using Decision Trees
Random Forests
SVM
Naive Bayes Classifier (SL)
Naive Bayesian Probabilistic Model in a Nutshell
Bayes Formula
Classification of Documents with Naive Bayes
Unsupervised Learning Type: Clustering
Clustering Examples
k-Means Clustering (UL)
k-Means Clustering in a Nutshell
k-Means Characteristics
Regression Analysis
Simple Linear Regression Model
Linear vs Non-Linear Regression
Linear Regression Illustration
Major Underlying Assumptions for Regression Analysis
Least-Squares Method (LSM)
Locally Weighted Linear Regression
Regression Models in Excel
Multiple Regression Analysis
Logistic Regression
Regression vs Classification
Time-Series Analysis

9 LAB EXERCISES