Data Science with Python_R_Spark

Objective: To learn data science step by step through real analytics examples.

Pre-requisites: Fundamentals of Python Programming and R programming

Session 1 (Introduction to R Programming)

- R Overview, Environment Setup, Basic Syntax
- Data Types, Variables, Operators
- Decision Making, Loops
- Functions, Strings, Vectors, Lists
- Matrices, Arrays, Factors
- Data Frames, Packages
- Reshaping

Session 2 (R Data Interfaces)

- Working with CSV Files
- Working with Excel Files
- Working with Binary Files
- Working with XML Files
- Working with JSON Files
- Working with WebData
- R Programming with Database

Session 3 (R Charts and Graphs)

- Pie Charts, Bar Charts, Boxplots
- Histograms, Line Graphs
- Scatter Plots
- Working with Shiny Apps and its Components

Session 4 (R Statistics Examples)

- Mean, Median and Mode
- Linear Regression
- Multiple Regression
- Logistic Regression
- Normal Distribution
- Binomial Distribution

- Poisson Regression
- Analysis of Covariance
- Time Series Analysis
- Nonlinear Least Square
- Decision Tree
- Random Forest
- Survival Analysis
- Chi Square Tests

Session 5 (Python Introduction)

- Basic Concepts, Data Types, Operators
- Control Statements, Looping, List, Tuple, Dictionary, Set
- Function and Types of Functions
- File Processing in Python

Session 6 (Working with Numpy)

- Introduction to Numpy, Environment Setup
- NDArray Object, Data Types, Array Attributes
- Array Creation Routines, Array from Existing Data
- Array From Numerical Ranges, Indexing and Slicing
- Iterating Over Arrays, Array Manipulation
- Binary Operators, String Functions, Mathematical Functions

Session 7 (More on Numpy with Matplotlib)

- Arithmetic Operators, Statistical Functions
- Sort, Search and Counting Functions, Byte Swapping
- Copies and Views, Matrix Library, Linear Algebra
- Matplotlib, Histogram Using Matplotlib
- I/O with Numpy
- Examples based on Numpy

Session 8 (Data Visualization using Matplotlib)

- Introduction to Matplotlib, Environment Setup
- Introduction to Anaconda and Jupyter Notebook
- PyPlot API, Simple Plot, PyLab Module
- Figure Class, Axes Class, Multiplots, Subplots() functions
- Subplot2grid() Function, Grids, Formatting Axis

- Setting Limits, Settings Ticks and Tick Labels

Session 9 (Data Visualization Plot Types)

- Bar Plot, Histogram, Pie Chart, Scatter Plot
- Contour Plot, Quiver Plot, Box Plot, Violin Plot
- 3D Contour Plot, 3D Wireframe Plot
- 3D Surface Plot, Working with Text and Images
- Working with Transforms
- Three Dimensional Plotting, Twin Axes

Session 10 (Working with Pandas)

- Introduction to Pandas, Data Structure
- Series, Data Frame, Panel and Basic Functionality
- Re-indexing, Iteration, Sorting using Pandas
- Working with Text Data, Statistical Functions
- Aggregations, Missing Data, GroupBy
- Merging/Joining, Concatenation, Date Functionality
- Time delta, Categorical Data, Visualization
- IO Tools, Sparse Data, Comparison with SQL

Session 11 (Mathematics for Data Science)

- Statistics
- Probability
- Calculus
- Linear Algebra

Session 12 (Fundamentals of Data Science)

- Introduction to Data Science, Basic Terminology
- Data Science Venn Diagram
- Data Science Case Study
- Working with Types of Data

Session 13 (The Five Steps of Data Sciences)

- Getting Problem Statements, Obtain the data
- Explore the Data, Model the Data
- Communicate and Visualize the Results

Session 14 (Linear Regression using Python)

- Scatter Diagram (Correlation Analysis)
- Scatter Diagram (Correlation Coefficient)
- Ordinary Least Squares
- Principles of Regression
- Splitting the data into training, validation and testing datasets
- Understanding Overfitting (Variance) vs Under Fitting (Bias)
- Generalization Error and Regularization Techniques
- Introduction to Simple Linear Regression, Heteroscedasticity/Equal Variance

Session 15 (Logistic Regression)

- Principles of Logistic Regression
- Types of Logistic Regression
- Assumption and Steps in Logistic Regression
- Analysis of Simple Logistic Regression Result
- Multiple Logistic Regression
- Confusion Matrix (False Positive, False Negative)

Session 16 (Apache Spark)

- Introduction to Apache Spark
- Installation
- Core Programming, Deployment
- How Apache Spark Works?

Session 17 (Project Work)