Applied Data Science with Python

Course ISI-1513 Two Days Instructor-Led, Hands-On

Introduction

This intensive training course provides theoretical and practical aspects of using Python in the realm of Data Science, Business Analytics, and Data Logistics. The coverage of the related core concepts, terminology, and theory is provided as well. This training course is supplemented by a variety of hands-on labs (the list of which is provided at the bottom of this outline) that help attendees reinforce their theoretical knowledge of the learned material.

Audience: Business Analysts, Developers, IT Architects, and Technical Managers

At Course Completion

In this course, students will learn the following:

- Applied Data Science and Business Analytics
- Common Data Science algorithms for supervised and unsupervised machine learning
- NumPy, pandas, Matplotlib, scikit-learn
- Python REPLs
- Jupyter notebooks
- Data analytics life-cycle phases
- Data repairing and normalizing
- Data aggregation and grouping
- Data visualization

Prerequisites

Participants should have a working knowledge of Python (or have the programming background and/or the ability to quickly pick up Python's syntax), and be familiar with core statistical concepts (variance, correlation, etc.)

Course Materials

The student kit includes a comprehensive workbook.

Course Outline

Module 1: Python for Data Science

- In-Class Discussion
- Importing Modules

- Listing Methods in a Module
- Creating Your Own Modules
- Random Numbers
- Zipping Lists
- List Comprehension
- Python Data Science-Centric Libraries
- NumPy
- NumPy Arrays
- Select NumPy Operations
- SciPy
- pandas
- Creating a pandas DataFrame
- Fetching and Sorting Data
- Scikit-learn
- Matplotlib
- Python Dev Tools and REPLs
- IPython
- Jupyter
- Jupyter Operation Modes
- Jupyter Common Commands
- Anaconda

Module 2: Applied Data Science

- What is Data Science?
- Data Science, Machine Learning, AI?
- Data Science Ecosystem
- Business Analytics vs. Data Science
- 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

Module 3: Data Analytics Life-cycle Phases

- Data Analytics Pipeline
- Data Discovery Phase
- Data Harvesting Phase
- Data Priming Phase
- Data Logistics and Data Governance
- Exploratory Data Analysis
- Model Planning Phase

Contact ISInc for more information at 916.920.1700 or by visiting our website at http://www.isinc.com

- Model Building Phase
- Communicating the Results
- Production Roll-out

Module 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

Module 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

Module 6: Data Grouping and Aggregation in Python

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

Module 7: Data Visualization with matplotlib

- Data Visualization
- What is matplotlib?

Contact ISInc for more information at 916.920.1700 or by visiting our website at http://www.isinc.com

- 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
- Figures
- Example of Using the figure() Function
- Saving Figures to a File
- Visualization with pandas
- Working with matplotlib in Jupyter Notebooks

Module 8: Data Science and ML Algorithms in scikit-learn

- In-Class Discussion
- Types of Machine Learning
- Terminology: Features and Observations
- Representing Observations
- Terminology: Labels
- Terminology: Continuous and Categorical Features
- Continuous Features
- Categorical Features
- Common Distance Metrics
- The Euclidean Distance
- What is a Model
- Supervised vs Unsupervised Machine Learning
- Supervised Machine Learning Algorithms
- Unsupervised Machine Learning Algorithms
- Choose the Right Algorithm
- The scikit-learn Package
- scikit-learn Estimators, Models, and Predictors
- Model Evaluation
- The Error Rate
- Feature Engineering
- Scaling of the Features
- Feature Blending (Creating Synthetic Features)
- The 'One-Hot' Encoding Scheme
- Example of 'One-Hot' Encoding Scheme
- Bias-Variance (Underfitting vs Overfitting) Trade-off
- The Modeling Error Factors
- One Way to Visualize Bias and Variance

- Underfitting vs Overfitting Visualization
- Balancing Off the Bias-Variance Ratio
- Regularization in scikit-learn
- Regularization, Take Two
- Dimensionality Reduction
- PCA and isomap
- The Advantages of Dimensionality Reduction
- The LIBSVM format
- Life-cycles of Machine Learning Development
- Data Split for Training and Test Data Sets
- Data Splitting in scikit-learn
- Hands-on Exercise
- Classification (Supervised ML) Examples
- Classifying with k-Nearest Neighbors
- k-Nearest Neighbors Algorithm
- k-Nearest Neighbors Algorithm
- Hands-on Exercise
- Regression Analysis
- Regression vs Correlation
- Regression vs Classification
- Simple Linear Regression Model
- Linear Regression Illustration
- Least-Squares Method (LSM)
- Gradient Descend Optimization
- Locally Weighted Linear Regression
- Regression Models in Excel
- Multiple Regression Analysis
- Linear Logistic (Logit) Regression
- Interpreting Linear Logistic Regression Results
- Decision Trees
- Decision Tree Terminology
- Properties of Decision Trees
- Decision Tree Classification in Context of Information Theory
- The Simplified Decision Tree Algorithm
- Using Decision Trees
- Random Forests
- Hands-On Exercise
- Support Vector Machines (SVMs)
- 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
- Global vs Local Minimum Explained
- Hands-On Exercise
- Time-Series Analysis
- Decomposing Time-Series
- A Better Algorithm or More Data?