Data Science with Anaconda

Course ISI-1503 5 Days Instructor-led, Hands on

Course Information

This five-day course for data scientists to develop a solution using Anaconda. Anaconda's commercial product, Anaconda Enterprise is software for enterprise organizations to collaborate, govern, scale, and secure Python and R data science and machine learning. Data scientists can securely collaborate on Notebooks and models with full enterprise directory-based access control and version control. IT can operate its own private Python and R package repository to mirror, filter, and manage packages to meet its governance and security policies. Quickly deploy ML models, live notebooks and dashboards to production compute clusters running Docker/Kubernetes, Hadoop, and Spark. Anaconda Enterprise runs in your own data centers and on AWS, Azure, and Google Cloud Platform. You can even run it in secure "air-gapped" environments that have no Internet access.

At Course Completion

Upon successful completion of this course, students will understand the following:

- Basic Python
- Natural Language Processing
- Essential Pandas
- Supervised Machine Learning
- Unsupervised Machine Learning
- Neural Networks and Deep Learning
- Big Data with Python
- Dask: Parallel & Distributed Computing
- Web Scraping and REST APIs

Prerequisites

This course requires that you should have the following experience:

- A concept of data science
- A background in R or Rapid Miner
- Prior programming knowledge, Python is most useful

Course Outline

Module 1: Basic Python

- Python fundamentals
    - Python ecosystem

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

- numbers: arithmetic, operators, relations strings: objects & member methods
- modules: using import & namespaces

- Flow control
  - logical expressions: bool , relational/logical operators branching: if -- else -- elif blocks
  - iteration: for & while blocks
  - exceptions: try -- except blocks
- Functions
  - signatures, arguments, docstring, return values positional/keyword arguments
  - lambda functions
- Data structures
  - list, dict, tuple, set methods/attributes for common data structures
  - idioms for slicing, indexing, & iteration
  - using comprehensions and builtins: sum, len, sorted
- File I/O
  - reading from & writing to files & file-like objects
  - using common file idioms & methods
  - context managers: using with

Module 2: Natural Language Processing

- NLP Overview
  - The importance of corpora
  - Clean first
  - Text Data
- Wrangling Tokenizing Sentences Words
  - Stop words
  - Statistics
  - Part-of-Speech tags
  - Stemming
- Classification
  - In corpora
  - By context
  - Feature extraction for machine learning
- Advanced topics
- Further reading

Module 3: Essential Pandas

(i.e. data pre-processing)

- Introduction
- Data Exploration
  - reading data sources
  - selections & summary statistics data analysis methods

- filtering data using logical conditions
- plotting in Jupyter notebooks

- Data Formats
  - Flat text files: CSV, TSV
  - Binary Formats: HDF5, SAS, Excel, databases
  - Structured files: JSON
- Pandas Data Structures: Index, Series, DataFrame
  - constructing new Pandas objects changing index columns modifying & sorting indexes working with non-unique indexes
  - index-aligned operations
- Time series
  - creating & using datetime indexes resampling time series
  - using rolling windows

Module 4: Supervised Machine Learning

- Training and validation
- Regression problems
  - Linear models
  - Support vector machines Decision trees Classification problems
  - K-nearest neighbors classification
  - Naive Bayes classification
  - Support vector machines
  - Decision trees and ensemble strategies
- Model building and scoring
  - Scoring functions & cross-validation
  - Feature selection
  - Feature extraction
- Pipelines
- Grid searches

Module 5: Unsupervised Machine Learning

- Feature extraction
- Clustering problems
  - K-means & hierarchical clustering
  - DBScan
- Dimensionality reduction (PCA, LDA, LFA, etc.)
- Detection & treatment of outliers

Module 6: Neural Networks and Deep Learning

- Neural nets as classifiers
- Perceptrons
- Convolutional neural networks
- Long short-term memory networks

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

Module 7: Big Data with Python

Python provides an easy facility for accessing large databases and on the Hadoop HDFS shared filesystem.

- Accessing Databases SQLAlchemy and ORMs The Python DBAPI
  - PyODBC
- PySpark
  - RDD files
  - DataFrame objects
  - HIVE tables and storage
  - Data Streaming

Module 8: Dask: Parallel & Distributed Computing

- Expressing delayed computations Choosing an appropriate scheduler Dask Delayed decorator
  - Working with Delayed objects
- Dask data structures
  - Dask DataFrame: connections to Numpy
  - Dask Array: connections to Pandas
  - Dask Bag: unstructured data
  - Storage formats
- Dask Distributed
  - Scheduler, worker processes, and connecting the client to submit actions
  - Scattering & gathering data
  - Nesting parallelism with Arrays & DataFrames

Module 9: Web Scraping and REST APIs

- Is web scraping legal?
- URLs and connections
  - requests vs urllib
- REST APIs
- Beautiful Soup
  - HTML, CSS, XML tag hierarchy
  - tag searching and
- navigation
- Scraping dynamic content
- Use Cases
  - WebDrivers and JavaScript
  - Page Crawling
  - Connect to REST API