Evolve I.T Hub Syllabus

Course I: Python for Data Science)

Total Estimated Time: 15 hours (may need more for beginners)

1. Environment Setup

Installation
Virtual Environments
Downloads
Connections
Hello World!

2. Python Basics

Why we Program?
Types
Syntax
Expressions and Variables
Strings
String Operations

3. Python Data Structures

Lists
Tuples
Dictionaries
Sets
Comprehensions

4. Python Programming Fundamentals

Conditions and Branching
Loops and Iterations
Functions
Objects and Classes

5. Working with data in Python

Reading Files with Open
Writing Files with Open
Loading Data with Pandas
Pandas: Working with and Saving Data

One Dimensional Numpy
Two Dimensional Numpy

Course II: Applied Data Science with Python

Total Estimated Time: 30 hours

1. Introduction to Data Science

This part of the course will introduce the learner to the basics of the Python programming environment, including fundamental python programming techniques such as lambdas, reading and manipulating CSV files, and the numpy library. It will introduce data manipulation and cleaning techniques using the popular python pandas data science library and introduce the abstraction of the Series and DataFrame as the central data structures for data analysis, along with tutorials on how to use functions such as groupby, merge, and pivot tables effectively. By the end of this, learners will be able to take tabular data, clean it, manipulate it, and run basic inferential statistical analyses.

Part I: Introduction

Introduction to Specialization
Data Science
The Jupyter Notebook
Python Basics
Python Functions
Python Types and Sequences
Python More on Strings
Python Demonstration: Reading and Writing CSV files (or Excels)
Python Dates and Times
Advanced Python Objects, map ()
Advanced Python Lambda and List Comprehensions Python
Demonstration: The Numerical Python Library (NumPy)

Part II: Data Cleansing and Processing with Pandas

The Series Data Structure
Querying a Series
The Data Frame Data Structure
Data Frame Indexing and Loading
Querying a Data Frame

Indexing Data frames
Missing Values
Merging Data frames

Pandas Idioms
Group by
Scales
Pivot Tables
Date Functionality

Part III: Statistical Techniques

Introduction
Distributions
More Distributions
Hypothesis Testing in Python

2. Data Visualisation and Representation in Python

This part of the course will introduce the learner to information visualization basics, with a focus on reporting and charting using the matplotlib library. The course will start with a design and information literacy perspective, touching on what makes a good and bad visualization, and what statistical measures translate into in terms of visualizations. It will focus on the technology used to make visualizations in python, matplotlib, and introduce users to best practices when creating basic charts and how to realize design decisions in the framework. This will be a tutorial of functionality available in matplotlib and demonstrate a variety of basic statistical charts helping learners to identify when a particular method is good for a particular problem.

Part I: Basic Plots

Introduction
Graphical heuristics
Tools Used
Matplotlib Architecture
Basic Plotting with Matplotlib
Scatterplots
Line Plots
Bar Charts

Part II: Visualisation Techniques

Subplots
Histograms
Box Plots
Plotting with Pandas

3. Machine Learning

This part of the course will introduce the learner to applied machine learning, focusing more on the techniques and methods than on the statistics behind these methods. It will start with a discussion of how machine learning is different than descriptive statistics, and introduce the scikit learn toolkit through a tutorial. The issue of the dimensionality of data will be discussed, and the task of clustering data, as well as evaluating those clusters, will be tackled. Supervised approaches for creating predictive models will be described, and learners will be able to apply the scikit learn predictive modelling methods while understanding process issues related to data generalizability (e.g. cross-validation, overfitting). The course will end with a look at more advanced techniques, such as building ensembles, and practical limitations of predictive models. By the end of this, learners will be able to identify the difference between a supervised (classification) and unsupervised (clustering) technique, identify which technique they need to apply for a particular dataset and need, engineer features to meet that need, and write python code to carry out an analysis.

Part I: Fundamentals (Intro to SciKit-Learn)

Introduction
Key Concepts in Machine Learning
Python Tools for Machine Learning
An Example Machine Learning Problem
Examining the Data
K-Nearest Neighbors Classification

Part II: Supervised Learning

Introduction to Supervised Machine Learning
Overfitting and Under fitting
Supervised Learning: Datasets
K-Nearest Neighbours: Classification and Regression
Linear Regression: Least-Squares
Linear Regression: Ridge, Lasso, and Polynomial Regression
Logistic Regression
Linear Classifiers: Support Vector Machines
Multi-Class Classification
Kernelized Support Vector Machines
Cross-Validation
Decision Trees

Part III: Model Selection and Evaluation

Model Evaluation & Selection

Confusion Matrices & Basic Evaluation Metrics
Classifier Decision Functions
Precision-recall and ROC curves
Multi-Class Evaluation
Regression Evaluation
Model Selection: Optimizing Classifiers for Different Evaluation Metrics

Part IV: Advanced learning methods (Optional)

Naive Bayes Classifiers
Random Forests
Gradient Boosted Decision Trees
Neural Networks
Data Leakage
Clustering
Conclusion

4. Text Classification (Optional)

This part of the course will introduce the learner to text mining and text manipulation basics. The course begins with an understanding of how text is handled by python, the structure of text both to the machine and to humans, and an overview of the nltk framework for manipulating text. It focuses on common manipulation needs, including regular expressions (searching for text), cleaning text, and preparing text for use by machine learning processes. We will apply basic natural language processing methods to text, and demonstrate how text classification is accomplished. Finally, it will explore more advanced methods for detecting the topics in documents and grouping them by similarity (topic modelling).

Introduction to Text Mining
Handling Text in Python
Regular Expressions
Basic Natural Language Processing
NLP tasks with NLTK
Text Classification
Identifying Features from Text
Naive Bayes Classifiers
Support Vector Machines
Learning Text Classifiers in Python
Case Study - Sentiment Analysis