Python and Data Science with Machine Learning Training

Prerequisite

1. Install Anaconda Distribution as per OS from https://www.anaconda.com/distribution/ (Python 3.7 version)
2. Sign Up for account creation on https://www.hackerrank.com/
3. Sign up for account creation on https://www.kaggle.com/
4. Sign up for account creation on https://github.com/
5. Git Bash Utility - https://git-scm.com/downloads

Module 1: Data Analysis using Numpy and Pandas

a. **Numpy**
   - Numpy Vector and Matrix
   - Functions – arange(), zeros(), ones(), linspace(), eye(), reshape(), random(), max(), min(), argmax(), argmin(), shape and dtype attribute
   - Indexing and Selection
   - Numpy Operations – Array with Array, Array with Scalars, Universal Array Functions

b. **Pandas**
   - Pandas Series
   - Pandas DataFrame
   - Missing Data (Imputation)
   - Group by Operations
   - Merging, Joining and Concatenating DataFrame.
   - Pandas Operations
   - Data Input and Output from wide variety of formats like csv, excel, db and html etc.

Module 2: Data Visualization using Matplotlib, Seaborn, Pandas-in built, Plotly and Cufflinks

a. **Matplotlib**
   - plot() using Functional approach
   - multi-plot using subplot()
   - plt.figure() using OO API Methods
   - add_axes(), set_xlabel(), set_ylabel(), set_title() Methods
   - Customization – figure size, impoving dpi, Plot appearance, Markers, Control over axis appearance and special Plot Types

b. **Seaborn**
   - **Distribution Plots** using distplot(), jointplot(), pairplot(), rugplot(), kdeplot()
   - **Categorical Plots** using barplot(), countplot(), boxplot(), violinplot(), stripplot(), swarmplot(), factorplot()
   - **Matrix Plots** using heatmap(), clustermap()
   - **Grid Plots** using PairGrid(), FacetGrid()

- **Regression Plots** using lmplot()
- **Styles and Colors** customization

c. **Plotly and Cufflinks**
   - **Interactive Plotting** using Plotly and Cufflinks

d. **Pandas Built-in**
   - Histogram, Area Plot, Bar Plot, Scatter Plot, Box-plot, Hex-plot, Kde-plot, Density Plot

e. **Choropleth Maps**
   - **Interactive World Map and US Map** using Plotly and Cufflinks

Module 3: Statistics and Probability

- Type of Data – Numerical, Categorical and Ordinal
- Mean, Median and Mode
- Variance and Standard Deviation
- Probability Density Function(PDF) and Probability Mass Function (PMF)
- Percentiles and Moments
- Covariance and Correlation
- Conditional Probability
- Bayes' Theorem

Module 4: GIT

- Distribution Version Control System
- How internally, GIT Manages Version Control on Changesets.
- Creating Repository
- Basic Commands like, git status, git add, git remove, git branch, git checkout, git log, git cat-file, git pull, git push, git commit
- Managing Configuration – System Level, User Level, Repository Level

Module 5: Jupyter Notebook

- Introduction, Basic Commands, Keyboard Shortcut and Magic Functions

Module 6: Linear Algebra and Calculus

- **Vector and Matrix,** basic operations
- Trigonometry
- Derivatives

Module 7: SQL

- **MySQL** Server and Client Installation
- **SQL Queries**
- **CRUD Operations**

Module 8: Machine Learning Introduction

- What is Machine Learning?
- Machine Learning Process Flow-Diagram
- Different Categories of Machine Leaning – Supervised, Unsupervised and Reinforcement
- Scikit-Learn Overview
- Scikit-Learn cheat-sheet

Module 9: Regression

- Linear Regression
- Robust Regression (RANSAC Algorithm)
- Exploratory Data Analysis (EDA)
- Correlation Analysis and Feature Selection
- Performance Evaluation – Residual Analysis, Mean Square Error (MSE), Co-efficient of Determination R^2, Mean Absolute Error (MAE), Root Mean Square Error (RMSE)
- Polynomial Regression
- Regularized Regression – Ridge, Lasso and Elastic Net Regression
- Bias-Variance Trade-Off
- Cross Validation – Hold Out and K-Fold Cross Validation
- Data Pre-Processing – Standardization, Min-Max, Normalization and Binarization
- Gradient Descent

Projects

1. Predicting Boston House Prices - https://www.kaggle.com/schirmerchad/bostonhoustingmlnd
2. Ecommerce Project – Company want to decide whether to focus their efforts on Mobile Experience or Website Experience.
3. USA Housing Prediction Project.
4. New York City Taxi Fare Prediction - https://www.kaggle.com/c/new-york-city-taxi-fare-prediction
5. Emergency 911 Calls - https://www.kaggle.com/mchirico/montcoalert

Module 10: Classification – Logistic Regression

- Sigmoid function
- Logistic Regression learning using Stochastic Gradient Descent (SGD)
- SGDClassifier

- Measuring accuracy using Cross-Validation, Stratified k-fold
- Confusion Matrix – True Positive (TP), False Positive (FP), False Negative(FN), True Negative (TN)
- Precision, Recall, F1 Score, Precision/Recall Trade-Off
- Receiver Operating Characteristics (ROC) Curve.

Projects

6. Digit Recognizer - https://www.kaggle.com/c/digit-recognizer
7. Titanic: Machine Learning from Disaster - https://www.kaggle.com/c/titanic
8. Advertising Project – Indicating whether a particular internet user will click on an advertisement or not.
9. Project on working on classified Data to predict the Target Class 0 or 1.
10. Another, Project on working on classified Data to predict the Target Class 0 or 1.

Module 11: Classification – k-Nearest Neighbor(KNN)

- Classification and Regression
- Application, Advantages and Disadvantages
- Distance Metric – Euclidean, Manhattan, Chebyshev, Minkowski
- Measuring accuracy using Cross-Validation, Stratified k-fold, Confusion Matrix, Precision, Recall, F1-score.

Projects

11. Breast Cancer Wisconsin (Diagnostic) Project using KNN - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
12. Iris Species - https://www.kaggle.com/uciml/iris

Module 12: Classification – SVM (Support Vector Machine)

- Classification and Regression
- Separating line, Margin and Support Vectors
- Linear SVC Classification
- Polynomial Kernel – Kernel Trick
- Gaussian Radial Basis Function (rbf)
- Grid Search to tune hyper-parameters.
- Support Vector Regression

Projects

13. Breast Cancer Wisconsin (Diagnostic) Project using KNN - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
14. Iris Species - https://www.kaggle.com/uciml/iris

Module 13: Classification –Decision Trees

- CART (Classification and Regression Tree)
- Advantages and Disadvantages and its applications
- Decision Tree Learning algorithms – ID3, C4.5, C5.0 and CART
- Gini Impurity, Entropy and Information Gain
- Decision Tree Regression
- Visualizing a Decision Tree using graphviz module.
- Regularization using tuning hyper-parameters using GridSearch CV.

Projects

15. IBM HR Analytics Employee Attrition and Performance - https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
16. Zomato Restaurants Data - https://www.kaggle.com/shrutimehta/zomato-restaurants-data
17. Predicting Bank Marketing Analysis - https://www.kaggle.com/kevalm/bank-marketingdataset
18. FIFA 18 Complete Player Dataset - https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

Module 14: Classification - Ensemble Methods

- Bootstrap Aggregating or Bagging
- Random Forest algorithm
- Extremely Randomized (Extra-Trees) Ensemble
- Boosting – AdaBoost (Adaptive Boosting), Gradient Boosting Machine (GBM), XGBoost (Extreme Gradient Boosting)

Module 15: Unsupervised Learning – Clustering

- Connectivity- based Clustering using Hierarchical Clustering.
- Ward's Agglomerative Hierarchical Clustering
- K-Means Clustering
- Elbow Method and Solhouette Analysis

Projects

19. Lending Club Loan Data Analysis - https://www.kaggle.com/wendykan/lending-club-loan-data
20. U.S. News And World Report's College Data - https://www.kaggle.com/flyingwombat/us-news-and-world-reports-college-data
21. Credit Card Dataset for Clustering - https://www.kaggle.com/arjunbhasin2013/ccdata

Module 16: Unsupervised Learning – Dimensionality Reduction

- Linear Principal Component Analysis (PCA) reduction.
- Kernel PCA
- Linear Discriminant Analysis (LDA) on Supervised Data.

Projects

22. Breast Cancer Wisconsin (Diagnostic) Analysis using PCA - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
23. Predicting Abalone's Sex - https://www.kaggle.com/yuridias/abalone-dataset
24. Wine Project - https://www.kaggle.com/zynicide/wine-reviews

Module 17: Natural Language Processing using NLTK

1. NLP with NLTK
2. NLTK Extensions and Explorations
3. Preprocessing: Tokenization
4. Preprocessing: Tokens to Vectors
5. Feturization: Word-Tuple, Term Frequency, TF-IDF
6. NLP Problems using Rule Based Models
7. NLP Problems using Machine Learning Algorithms

Projects

25. SMS Spam Collection Dataset Analysis - https://www.kaggle.com/uciml/sms-spam-collection-dataset
26. Auto Summarizing Text using Rule Based Model.
27. Yelp Business Rating Prediction - https://www.kaggle.com/c/yelp-recsys-2013