Data science curriculum

Introduction to Machine Learning in Python

- What is machine learning & why is it so important?
- Applications of machine learning across industries
- Machine Learning methodology
- Machine Learning Toolbox
- Tool of choice- Python: what & why?
- Course Components

Python

Introduction to Python

- Installation of Python framework and packages: Anaconda and pip
- Writing/Running python programs using Spyder, Command Prompt
- Working with Jupyter Notebooks
- Creating Python variables: Numeric, string and logical operations
- Basic Data containers: Lists, Dictionaries, Tuples & sets
- Practice assignment

Iterative Operations & Functions in Python

- Writing for loops in Python
- List & Dictionary Comprehension
- While loops and conditional blocks
- List/Dictionary comprehensions with loops
- Writing your own functions in Python
- Writing your own classes and functions as class objects
- Practice assignment

Data Summary; Numerical and Visual in Python

- Need for data summary
- Summarizing numeric data in pandas
- Summarizing categorical data
- Group wise summary of mixed data
- Need for visual summary
- Introduction to ggplot & Seaborn
- Visual summary of different data combinations
- Practice Exercise

Data Handling in Python using NumPy & Pandas

Data science curriculum

- Introduction to NumPy arrays, functions &properties
- Introduction to pandas
- Dataframe functions and properties
- Reading and writing external data
- Manipulating Data Columns

Machine Learning in Python

Basics of Machine Learning

- Business Problems to Data Problems
- Broad Categories of Business Problems
- Supervised and Unsupervised Machine Learning Algorithm
- Drivers of ML algorithms
- Cost Functions
- Brief introduction to Gradient Descent
- Importance of Model Validation
- Methods of Model Validation
- Introduction to Cross Validation and Average Error

Generalized Linear Models in Python

- Linear Regression
- Limitation of simple linear models and need of regularization
- Ridge and Lasso Regression (L1 & L2 Penalties)
- Introduction to Classification with Logistic Regression
- Methods of threshold determination and performance measures for classification score models
- Case Studies

Tree Models using Python

- Introduction to decision trees
- Tuning tree size with cross validation
- Introduction to bagging algorithm
- Random Forests
- Grid search and randomized grid search
- ExtraTrees (Extremely Randomized Trees)
- Partial Dependence Plots
- Case Studies
- Home exercises

Boosting Algorithms using Python

- Concept of weak learners

Data science curriculum

- Introduction to boosting algorithms
- Adaptive Boosting
- Extreme Gradient Boosting (XGBoost)
- Case study
- Home exercise

Support Vector Machines (SVM) and KNN in Python

- Introduction to idea of observation based learning
- Distances and Similarities
- K Nearest Neighbours (KNN) for classification
- Introduction to SVM for classification
- Regression with KNN and SVM
- Case study
- Home exercises

Unsupervised learning in Python

- Need for dimensionality reduction
- Introduction to Principal Component Analysis (PCA)
- Difference between PCAs and Latent Factors
- Introduction to Factor Analysis
- Patterns in the data in absence of a target
- Segmentation with Hierarchical Clustering and K-means
- Measure of goodness of clusters
- Limitations of K-means
- Introduction to density based clustering (DBSCAN)

Neural Networks

- Introduction to Neural Networks
- Single layer neural network
- Multiple layer Neural network
- Back propagation Algorithm
- Moment up and decaying learning rate in context of gradient descent
- Neural Networks implementation in Python

Text Mining in Python

- Quick Recap of string data functions
- Gathering text data using web scraping with urllib
- Processing raw web data with BeautifulSoup
- Interacting with Google search using urllib with custom user agent
- Collecting twitter data with Twitter API
- Introduction to Naive Bayes

Data science curriculum

- Feature Engineering for text Data
- Feature creation with TFIDF for text data
- Case Studies

Ensemble Methods in Machine Learning

- Making use of multiple ML models taken together
- Simple Majority vote and weighted majority vote
- Blending
- Stacking
- Case Study

List of Projects

- **Reviewing Customer Complaints:** Create a machine learning system which will automatically pick out customer complaints most likely to be unresolved and escalate them
- **Credit Card Fraud Detection:** Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification
- Pick any interesting project of your choice