Data Science - Curriculum

Module 1: Foundations of Data Science

Description: Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. In this first module we will introduce to the field of Data Science and how it relates to other fields of data like Artificial Intelligence, Machine Learning and Deep Learning.

- Introduction to Data Science
- High level view of Data Science, Artificial Intelligence & Machine Learning
- Subtle differences between Data Science, Machine Learning & Artificial Intelligence
- Approaches to Machine Learning
- Terms & Terminologies of Data Science
- Understanding an end to end Data Science Pipeline, Implementation cycle

Module 2: Math for Data Science, Machine Learning and Artificial Intelligence

Description: Mathematics is very **important** in the field of **data science** as concepts within **mathematics** aid in identifying patterns and assist in creating algorithms. The understanding of various notions of Statistics and Probability Theory are key for the implementation of such algorithms in **data science**.

- Linear Algebra
- Matrices, Matrix Operations
- Eigen Values, Eigen Vectors
- Scalar, Vector and Tensors
- Prior and Posterior Probability
- Conditional Probability
- Calculus
- Differentiation, Gradient and Cost Functions
- Graph Theory

Module 3: Statistics for Data Science

Description: This module focuses on understanding statistical concepts required for Data Science, Machine Learning and Deep Learning. In this module, you will be introduced to the estimation of various statistical measures of a data set, simulating random distributions, performing hypothesis testing, and building statistical models.

Descriptive Statistics

- Types of Data (Discrete vs Continuous)
- Types of Data (Nominal, Ordinal)
- Measures of Central Tendency (Mean, Median, Mode)

- Measures of Dispersion (Variance, Standard Deviation)
- Range, Quartiles, Inter Quartile Ranges
- Measures of Shape (Skewness and Kurtosis)
- Tests for Association (Correlation and Regression)
- Random Variables
- Probability Distributions
- Standard Normal Distribution
- Probability Distribution Function
- Probability Mass Function
- Cumulative Distribution Function

Inferential Statistics

- Statistical sampling & Inference
- Hypothesis Testing
- Null and Alternate Hypothesis
- Margin of Error
- Type I and Type II errors
- One Sided Hypothesis Test, Two-Sided Hypothesis Test
- Tests of Inference: Chi-Square, T-test, Analysis of Variance
- t-value and p-value
- Confidence Intervals

Module 4: Python for Data Science

Python for Data Science

- Numpy
- Pandas
- Matplotlib & Seaborn
- Jupyter Notebook

Numpy

NumPy is a Python library that works with arrays when performing scientific computing with Python. Explore how to initialize and load data into arrays and learn about basic array manipulation operations using NumPy.

- Loading data with Numpy
- Comparing Numpy with Traditional Lists
- Numpy Data Types
- Indexing and Slicing
- Copies and Views
- Numerical Operations with Numpy
- Matrix Operations on Numpy Arrays
- Aggregations functions
- Shape Manipulations
- Broadcasting
- Statistical operations using Numpy
- Resize, Reshape, Ravel
- Image Processing with Numpy

Pandas
Pandas is a Python library that provides utilities to deal with structured data stored in the form of rows and columns. Discover how to work with series and tabular data, including initialization, population, and manipulation of Pandas Series and DataFrames.

- Basics of Pandas
- Loading data with Pandas
- Series
- Operations on Series
- DataFrames and Operations of DataFrames
- Selection and Slicing of DataFrames
- Descriptive statistics with Pandas
- Map, Apply, Iterations on Pandas DataFrame
- Working with text data
- Multi Index in Pandas
- GroupBy Functions
- Merging, Joining and Concatenating DataFrames
- Visualization using Pandas

Data Visualization using Matplotlib
Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+

- Anatomy of Matplotlib figure
- Plotting Line plots with labels and colors
- Adding markers to line plots
- Histogram plots
- Scatter plots
- Size, Color and Shape selection in Scatter plots.
- Applying Legend to Scatter plots
- Displaying multiple plots using subplots
- Boxplots, scatter_matrix and Pair plots

Data Visualization using Seaborn
Seaborn is a data visualization library that provides a high-level interface for drawing graphs. These graphs are able to convey a lot of information, while also being visually appealing.

- Basic Plotting using Seaborn
- Violin Plots
- Box Plots
- Cat Plots
- Facet Grid
- Swarm Plot
- Pair Plot
- Bar Plot
- LM Plot
- Variations in LM plot using hue, markers, row and col

Module 5: Exploratory Data Analysis

Exploratory Data Analysis helps in identifying the patterns in the data by using basic statistical methods as well as using visualization tools to displays graphs and charts. With EDA we can assess the distribution of the data and conclude various models to be used.

Pipeline ideas

- Exploratory Data Analysis
- Feature Creation
- Evaluation Measures

Data Analytics Cycle ideas

- Data Acquisition
- Data Preparation
    - Data cleaning
    - Data Visualization
    - Plotting
- Model Planning & Model Building

Data Inputting

- Reading and writing data to text files
- Reading data from a csv
- Reading data from JSON

Data preparation

- Selection and Removal of Columns
- Transform
- Rescale
- Standardize
- Normalize
- Binarize
- One hot Encoding
- Imputing
- Train, Test Splitting

Module 6: Machine Learning

In machine learning, computers apply statistical learning techniques to automatically identify patterns in data. This module on Machine Learning is a deep dive to Supervised, Unsupervised learning and Gaussian / Naive-Bayes methods. Also you will be exposed to different classification, clustering and regression methods.

- Introduction to Machine Learning
- Applications of Machine Learning
- Supervised Machine Learning
    - Classification
    - Regression

- Unsupervised Machine Learning
- Reinforcement Learning
- Latest advances in Machine Learning
- Model Representation
- Model Evaluation
- Hyper Parameter tuning of Machine Learning Models.
- Evaluation of ML Models.
- Estimating and Prediction of Machine Learning Models
- Deployment strategy of ML Models.

Module 7: Supervised Machine Learning – Classification

Supervised learning is one of the most popular techniques in machine learning. In this module, you will learn about more complicated supervised learning models and how to use them to solve problems.

Classification methods & respective evaluation

- K Nearest Neighbors
- Decision Trees
- Naive Bayes
- Stochastic Gradient Descent
- SVM –
    - Linear
    - Non linear
    - Radial Basis Function
- Random Forest
- Gradient Boosting Machines
- XGboost
- Logistic regression

Ensemble methods

- Combining models
- Bagging
- Boosting
- Voting
- Choosing best classification method

Model Tuning

- Train Test Splitting
- K-fold cross validation
- Variance bias tradeoff
- L1 and L2 norm
- Overfit, underfit along with learning curves variance bias sensibility using graphs
- Hyper Parameter Tuning using Grid Search CV

Respective Performance measures

- Different Errors (MAE, MSE, RMSE)
- Accuracy, Confusion Matrix, Precision, Recall

Module 8: Supervised Machine Learning - Regression

Regression is a type of predictive modelling technique which is heavily used to derive the relationship between variables (the dependent and independent variables). This technique finds its usage mostly in forecasting, time series modelling and finding the causal effect relationship between the variables. The module discusses in detail about regression and types of regression and its usage & applicability

Regression

- Linear Regression
- Variants of Regression
  - Lasso
  - Ridge
- Multi Linear Regression
- Logistic Regression (effectively, classification only)
- Regression Model Improvement
- Polynomial Regression
- Random Forest Regression
- Support Vector Regression

Respective Performance measures

- Different Errors (MAE, MSE, RMSE)
- Mean Absolute Error
- Mean Square Error
- Root Mean Square Error

Module 9: Unsupervised Machine Learning

Unsupervised learning can provide powerful insights on data without the need to annotate examples. In this module, you will learn several different techniques in unsupervised machine learning.

Clustering

- K means
- Hierarchical Clustering
- DBSCAN

Association Rule Mining

- Association Rule Mining.
- Market Basket Analysis using Apriori Algorithm
- Dimensionality reduction using Principal Component analysis (PCA)

Module 10: Natural Language Processing

Natural language is essential to human communication, which makes the ability to process it an important one for computers. In this module, you will be introduced to natural language processing and some of the basic tasks.

- Text Analytics
- Stemming, Lemmatization and Stop word removal.
- POS tagging and Named Entity Recognition
- Bigrams, Ngrams and colocations
- Term Document Matrix
- Count Vectorizer
- Term Frequency and TF-IDF

Module 11: Advanced Analytics

Advanced Analytics covers various areas like Time series Analysis, ARIMA models, Recommender systems etc.

Time series

- Time series Analysis.
- ARIMA example

Recommender Systems

- Content Based Recommendation
- Collaborative Filtering

Module 12: Reinforcement Learning

Reinforcement learning is an area of **Machine Learning** which takes suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.

- Basic concepts of Reinforcement Learning
- Action
- Reward
- Penalty Mechanism
- Feedback loop
- Deep Q Learning

Module 13: Artificial Intelligence

Artificial intelligence (AI) is the ability of a computer program or a machine to think and learn. It is also a field of study which tries to make computers "smart"

Artificial Neural Networks

- Neural Networks & terminologies
- Non linearity problem, illustration
- Perceptron learning
- Feed Forward Network and Back propagation

- Gradient Descent

Mathematics of Artificial Neural Networks
- Gradients
- Partial derivatives
- Linear algebra
  - Li
  - LD
  - Eigen vectors
  - Projections
- Vector quantization

Overview of tools used in Neural Networks
- Tensor Flow
- Keras

Module 14: Deep Learning

Deep learning is part of a broader family of machine learning methods based on the layers used in artificial neural networks. In this module, you'll deep dive in the concepts of Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Auto Encoders and many more.

Deep Learning
- Tensorflow & keras installation
- More elaborate discussion on cost function
- Measuring accuracy of hypothesis function
- Role of gradient function in minimizing cost function
- Explicit discussion of Bayes models
- Hidden Markov Models (HMM)
- Optimization basics
- Sales Prediction of a Gaming company using Neural Networks
- Build an Image similarity engine.

Deep Learning with Convolutional Neural Nets
- Architecture of CNN
- Types of layers in CNN
- Different Filters and Kernals
- Building an Image classifier with and without CNN

Recurrent neural nets
- Fundamental notions & ideas
  - Recurrent neurons
  - Handling variable length sequences
- Training a sequence classifier
- Training to predict Time series

Module 15: Cloud Computing for Data Science

Cloud computing is massively growing in importance in the IT sector as more and more companies are eschewing traditional IT and moving applications and business processes to the cloud. This section covers detailed information about how to deploy Data Science models on Cloud environments.

Topics

- Introduction to Cloud Computing
- Amazon Web Services Preliminaries - S3, EC2, RDS
- Big data processing on AWS using Elastic Map Reduce (EMR)
- Machine Learning using Amazon Sage Maker
- Deep Learning on AWS Cloud
- Natural Language processing using AWS Lex
- Analytics services on AWS Cloud
- Data Warehousing on AWS Cloud
- Creating Data Pipelines on AWS Cloud

Module 16: DevOps for Data Science

DevOps play a pivotal role in bridging the gap between Development and Operational teams. This section covers key DevOps tools which a Data Scientist need to be aware of for doing their day to day data science work.

Topics

- Introduction to DevOps for Data Science
- Tasks in Data Science Development
- Deploying Models in Production
- Deploying Machine Learning Models as Services
- Running Machine Learning Services in Containers
- Scaling ML Services with Kubernetes

Projects:

Build an XML to CSV converter using Python

Description: This python project which will help students to brush up their basic python skills to build a real-world XML to CSV Converter.

Industry: Data Operations

Tools: Python, XML, CSV, Data Ingestion

Building a Photo Editor from scratch using Flask and Numpy

Description: Numpy is a versatile package to do data operations on matrices, numbers and number operations. With this project we will learn doing number operations on Images and apply filters, cropping, flipping and resizing using Numpy.

Industry: Media and Photography

Tools: Python, Numpy, Image Processing, Matplotlib

Exploratory Data Analysis on Retail Shop Sales data.

Description: Performing exploratory data analysis to find patterns in data which will determine the approach to take in Machine Learning. Exploratory data analysis will help identifying features which can be used or discarded in the Machine Learning approach.

Industry: Retail

Tools: Exploratory Data Analysis, Data Preprocessing, Scaling Techniques, Missing Value Imputations.

Fruit type prediction using K Nearest Neighbors algorithm.

Description: K Nearest Neighbor algorithm is used to predict the type of fruit given its mass, height, width and color score. We will understand on how to use KNN algorithm and how to tune the hyperparameters such as n_neighbors and k value.

Industry: Food and Beverages

Tools: Supervised Machine Learning, K Nearest Neighbor Classifier, KNN Algorithm, Pandas, Numpy, Matplotlib

Predict Malignancy in Mammographic Masses using Decision Tree Classifiers

Description: Mammographic masses is a public dataset from UCI machine learning repository which has information on mass shape, structure, density and age of a person. With this project we will use Decision Tree classifier to detect Malignancy (either Malignant or Benign) of the Mammographic Mass

Industry: Healthcare

Tools: Decision Tree Classifier, Sklearn, Pandas, Numpy, Graphviz, Matplotlib

Predict whether a candidate will be shortlisted in H1B Visa process using Random Forest Algorithm.

Description: Every year there are close to 1 lakh 65 thousand applications for H1B Visa processing and only few get shortlisted in the process. A lot of information is used for scrutinizing candidates in the selection process. In this project we will use Random Forest algorithm on past five years of H1B processing data to identify if a H1B application will be selected or rejected in the application process.

Industry: US H1B Visa - USCIS

Tools: Random Forest, Pandas, Sklearn, Numpy, Seaborn, Matplotlib

Predict Breast Cancer using Support Vector Machine algorithm.

Description: Wisconsin Breast Cancer dataset has 569 sample of Breast cancer observations determining Malignancy or Benign state of breast mass. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. In this project we will use SVM algorithm to detect malignancy in the cell.

Industry: Healthcare

Tools: SVM, Numpy, Pandas, Sklearn

HR Analytics for Attrition Prediction using Logistic Regression

Description: Employee Attrition is an important subject to gauge the satisfaction of the employee in a company. HR departments take various measures to arrest employee attrition. In this project we will use Logistic regression to predict who is the potential employee who is in a verge of leaving the company.

Industry: Human Resources

Tools: Logistic Regression, Sklearn, Numpy, Pandas, Matplotlib

Churn Analysis in Telecommunication:

Description: A customer can be called as a “churner” when he/she discontinue their subscription in a company and move their business to a competitor. Prediction as well as prevention of customer churn brings a huge additional revenue source for every business.

Here, we use a telecom customer data set to classify the set of possible customers who are likely to churn

Industry: Telecommunications

Tools: SVM, K Nearest Neighbors Classifier, Random Forest, Logistic Regression, Pandas, Numpy, Sklearn, Seaborn, Matplotlib

Predicting Housing prices using Regression:

Description: Predict the sales price for each house based on input features provided for the house.

Industry: Real Estate

Tools: Linear Regression, Multi Linear Regression, Sklearn, Pandas, Numpy, Matplotlib

Social Network Ads based Prediction:

Description: Predicting if a user buys a specific product or not based on the ad populated on the Social Network. Data uses specific demographic information about the user to do the prediction.

Industry: Social Media

Tools: Hierarchical Clustering, Scipy, Sklearn, Pandas, Numpy, Seaborn

Retail Customer segmentation based on spending patterns

Description: Customer analysis plays a crucial role in determining the profitability of Retail companies. Segmentation of the customer based on their purchase patterns helps Retail companies to cluster their user base and serve them effectively.

Industry: Retail

Tools: K-Means, Clustering, Sklearn, Pandas, Numpy, Seaborn

Market Basket Analysis using Apriori Algorithm

Description: Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. A co-occurrence is when two or more things take place together.

The technique determines relationships of what products were purchased with which other product(s).

Industry: Retail

Tools: Apriori Algorithm, Numpy, Pandas

SMS Spam Detection using Natural Language Processing

Description: In our day-to-day lives, we receive a large number of spam/junk messages either in the form of Text (SMS) or E-mails. It is important to filter these spam messages since they are not truthful or trustworthy.

In this case study, we apply various machine learning algorithms to categorize the messages depending on whether they are spam or not.

Industry: Social Media

Tools: NLP, Sklearn, Logistic Regression

Sentiment analysis on Restaurant Reviews using Natural Language processing and Supervised Learning

It involves in identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Industry: Food and Beverages, E-Commerce

Tools: NLP, Sklearn, Support Vector Machine

Image Classification using Deep Learning

Description: Classifying Images based on the features is a tough problem. With Deep Learning algorithms like CNN it has become fairly easy. In this project we will learn to classify two faces using Convolutional Neural Networks.

Industry: Media

Tools: Deep Learning, Neural Networks, Tensorflow, Keras, CNN

Content Based Recommender Engine using Deep Learning

Description: Content based recommender systems use the content in the data to segment items and then use them for recommending similar items. In this project we will use Deep Learning and Convolutional Neural Networks to perform content-based recommender engine.

Industry: E-Commerce, Retail

Tools: Tensorflow, Keras, CNN, Transfer learning, Deep Learning

Chatbots using Recurrent Neural Networks and Deep Learning

Description: A chatbot is an artificial intelligence (AI) software that can simulate a conversation (or a chat) with a user in natural language through messaging applications, websites, mobile apps or through the telephone. In this project we will build a conversational chatbot using Deep Learning.

Industry: Retail, Banking, E-Commerce, Media

Tools: Tensorflow, Keras, RNN, LSTM, Numpy

Stock market price prediction

Description: This project deals with the predictions of stock market prices using history of Data. It also considers the physical factors vs. psychological, rational and irrational behavior etc. Machine learning techniques implemented in Python acts as game changer for the predictions. Algorithms including Linear regression, LSTM and ARIMA model are used for the same.

Industry: Finance

Tools: Tensorflow, Keras, LSTM, RNN

Exploratory Data analysis of Crime records in Boston

Description: This project analyses data using quantitative prediction of crimes in Boston and drawing visualizations of Trends in the data over the years. Exploratory Data Analysis is carried on the crimes data using lots of techniques from Linear model to Stochastic gradient boosting.

Industry: Defense and Law Enforcement

Tools: Linear Regression, XGBoost, Numpy, Sklearn, Pandas, Matplotlib