

Foundations of AI and ML

Introduction to Data Science and AI & ML

- Data Science, AI & ML
- Use Cases in Business and Scope
- Scientific Method
- Modelling Concepts
- CRISP-DM Method

Python

- Install python
- jupyter Notebook
- <http://jupyter.readthedocs.io/en/latest/install.html>
- Spyder
- <https://pypi.python.org/pypi/spyder>

- Introduction to Python
- Basic Python Syntax
- Data types
- <https://docs.python.org/3/library/datatypes.html>
- Function
- Modules
- Packages
- Machine Learning Python libraries
- NumPy
- SciPy
- Pandas etc.

Descriptive Statistics

- Data exploration (histograms, bar chart, box plot, line graph, scatter plot)
- Qualitative and Quantitative Data
- Measure of Central Tendency (Mean, Median and Mode),
- Measure of Positions (Quartiles, Deciles, Percentiles and Quantiles),
- Measure of Dispersion (Range, Median, Absolute deviation about median, Variance and Standard deviation), Anscombe's quartet
- Other Measures: Quartile and Percentile, Interquartile Range

Statistical Analysis Initial Data Analysis

- Relationship between attributes: Covariance, Correlation Coefficient, Chi Square
- Measure of Distribution (Skewness and Kurtosis), Box and Whisker Plot (Box Plot and its parts, Using Box Plots to compare distribution) and other statistical graphs Probability
- Probability (Joint, marginal and conditional probabilities)
- Probability distributions (Continuous and Discrete)
- Density Functions and Cumulative functions

Data Management

Data Acquisition

- Gather information from different sources
- Internal systems and External systems

- Web APIs, Open Data Sources, Data APIs, Web Scrapping
- Relational Database access (queries) to process/access data
- Data Pre-processing and Preparation
- Data Munging, Wrangling
- Plyr packages
- Cast/Melt

Data Quality and Transformation

- Data imputation
- Data Transformation (minmax, log transform, z-score transform etc,)
- Binning, Classing and Standardization
- Outlier/Noise& Anomalies

Handling Text Data -NLP concepts

- Bag-of-words
- Regular Expressions
- Sentence Splitting and Tokenization
- Punctuations and Stop words, Incorrect spellings
- Properties of words and Word cloud
- Lemmatization and Term-Document TxD computation
- Sentiment Analysis (Case Study)
- Principles of Big Data
- Introduction to Big Data
- Challenges of processing Big Data (Volume, Velocity and Variety perspective)
- Use Cases
- Big Data Frameworks – Hadoop, Spark and NoSQL
- Processing, Storage and Programming Framework
- Hadoop eco-system Components and their functions
- Essential Algorithms (Word count, Page Rank, IT-IDF)
- Spark: RDDs, Streaming and Spark ML
- NoSQL concepts (CAP, ACID, NoSQL types)

Statistical Decision Making

Data Visualization

Visualizing and Communicating clearly and effectively about the patterns we find in data is a key skill for a successful data professional This module focuses on the design and implementation of complementary visual and verbal representations of patterns and analyses in order to convey findings, answer questions, drive decisions, and provide persuasive evidence supported by data

- Science of Visualization
- Visualization Periodic Table
- Aesthetics and Story telling
- Concepts of measurement - scales of measurement
- Design of data collection formats with illustration
- Principles of data visualization - different methods of presenting data in business analytics

- Concepts of Size, Shape, Color
- Various Visualization types
- Bubble charts
- Geo-maps (Chloropleths)
- Gauge charts
- Tree map
- Heat map
- Motion charts
- Force Directed Charts etc,

Sampling and Estimation

- Sample versus population
- Sample techniques (simple, stratified, clustered, random)
- Sampling Distributions
- Parameter Estimation
- Unbalanced data treatment

Inferential Statistics

- Develop an intuition how to understand the data, attributes, distributions
- Procedure for statistical testing, etc
- Test of Hypothesis (Concept of Hypothesis testing, Null Hypothesis and Alternative Hypothesis)
- Cross Tabulations (Contingency table and their use, Chi-Square test, Fisher's exact test), One Sample t test (Concept, Assumptions, Hypothesis, Verification of assumptions, Performing the test and interpretation of results)
- Independent Samples t test
- Paired Samples t test
- One-way ANOVA (Post hoc tests: Fisher's LSD, Tukey's HSD)
- z-test and F-test

Machine Learning

Predictive Analytics

Linear Regression

- Regression basics: Relationship between attributes using Covariance and Correlation
- Relationship between multiple variables: Regression (Linear, Multivariate) in prediction
Residual Analysis
- Identifying significant features, feature reduction using AIC, multi-collinearity
- Non-normality and Heteroscedasticity
- Hypothesis testing of Regression Model
- Confidence intervals of Slope
- R-square and goodness of fit
- Influential Observations – Leverage

Multiple Linear Regression

- Polynomial Regression
- Regularization methods

- Lasso, Ridge and Elastic nets
- Categorical Variables in Regression

Non-Linear Regression

- Logit function and interpretation
- Types of error measures (ROCR)
- Logistic Regression in classification
- Forecasting models
- Trend analysis
- Cyclical and Seasonal analysis
- Smoothing; Moving averages; Box-Jenkins, Holt-winters, Auto-correlation; ARIMA
- Examples: Applications of Time Series in financial markets

Foundations for ML

- ML Techniques overview
- Validation Techniques (Cross-Validations)
- Feature Reduction/Dimensionality reduction
- Principal components analysis (Eigen values, Eigen vectors, Orthogonality)

Clustering

- Distance measures
- Different clustering methods (Distance, Density, Hierarchical)
- Iterative distance-based clustering;
- Dealing with continuous, categorical values in K-Means
- Constructing a hierarchical cluster
- K-Medoids, k-Mode and density-based clustering
- Measures of quality of clustering

Classification

Naïve Bayes Classifier

- Model Assumptions, Probability estimation
- Required data processing
- M-estimates, Feature selection: Mutual information
- Classifier

K-Nearest Neighbours

- Computational geometry; Voronoi Diagrams; Delaunay Triangulations
- K-Nearest Neighbour algorithm; Wilson editing and triangulations
- Aspects to consider while designing K-Nearest Neighbour

Support Vector Machines

- Linear learning machines and Kernel space, Making Kernels and working in feature space SVM for classification and regression problems

Decision Trees

- ID4, C45, CART
- Ensembles methods
- Bagging & boosting and its impact on bias and variance
- C50 boosting

Random forest

- Gradient Boosting Machines and XGBoost

Association Rule mining

The applications of Association Rule Mining: Market Basket, Recommendation Engines, etc

A mathematical model for association analysis; Large item sets; Association Rules

Apriori: Constructs large item sets with mini sup by iterations; Interestingness of discovered association rules;

Application examples; Association analysis vs classification

FP-trees

Artificial Intelligence

Foundations for AI

- AI: Application areas
- AI Basics (Divide and Conquer, Greedy, Branch and Bound, Gradient Descent)
- NN basics (Perceptron and MLP, FFN, Backpropagation)

Convolution Neural Networks

- Image classification
- Text classification
- Image classification and hyper-parameter tuning
- Emerging NN architectures

Recurrent Neural Networks

- Building recurrent NN
- Long Short-Term Memory
- Time Series Forecasting

Deep Learning

- Auto-encoders and unsupervised learning
- Stacked auto-encoders and semi-supervised learning
- Regularization - Dropout and Batch normalization
- Py torch models like Vggnet , Resnet, Densenet, Alexnet , Inception blocks Etc.
- Keras models (yolo models)
- Computer vision

Additional Skills:

- English communication skills.
- Personality Development.

== > Real time Projects 3 with explanation