Data Science Course Curriculum

Module 1 – Introduction, Data Science Overview, Recommender Overview

- Introduction
- Data Science Overview
- Use Cases
- Project Lifecycle
- Data Acquisition
- Evaluating Input Data
- Data Transformation
- Data Analysis and Statistical methods
- Fundamentals of Machine Learning
- Recommender Overview
- Basic Introduction to Apache Mahout

Module 2 – Use Cases, Project Lifecycle

- What is Data Science?
- What Kind of Problems can you solve?
- Data Science Project Life Cycle
- Data Science-Basic Principles
- Data Acquisition
- Data Collection
- Understanding Data- Attributes in a Data, Different types of Variables
- Build the Variable type Hierarchy
- Two Dimensional Problem
- Co-relation b/w the Variables- explain using Paint Tool
- Outliers, Outlier Treatment
- Boxplot, How to Draw a Boxplot

Module 3 – Data Acquisition

- Discussion on Boxplot- also Explain
- Example to understand variable Distributions

- What is Percentile? – Example using Rstudio tool
- How do we identify outliers?
- How do we handle outliers?
- Outlier Treatment : Using Capping/Flooring General Method
- Distribution- What is Normal Distribution?
- Why Normal Distribution is so popular?
- Uniform Distribution
- Skewed Distribution
- Transformation

Module 4 – Machine Learning

- Discussion about Boxplot and Outlier
- Goal: Increase Profits of a Store
- Areas of increasing the efficiency
- Data Request
- Business Problem: To maximize shop Profits
- What are Interlinked variables
- What is Strategy
- Interaction b/w the Variables
- Univariate analysis
- Multivariate analysis
- Bivariate analysis
- Relation b/w Variables
- Standardize Variables
- What is Hypothesis?
- Interpret the Correlation
- Negative Correlation
- Machine Learning

Module 5 – Data Analysis and Statistical Methods, Implementing Recommenders with Apache Mahout, Data Transformation

- Correlation b/w Nominal Variables
- Contingency Table
- What is Expected Value?
- What is Mean?
- How Expected Value is differ from Mean

- Experiment – Controlled Experiment, Uncontrolled Experiment
- Degree of Freedom
- Dependency b/w Nominal Variable & Continuous Variable
- Linear Regression
- Extrapolation and Interpolation
- Univariate Analysis for Linear Regression
- Building Model for Linear Regression
- Pattern of Data means?
- Data Processing Operation
- What is sampling?
- Sampling Distribution
- Stratified Sampling Technique
- Disproportionate Sampling Technique
- Balanced Allocation-part of Disproportionate Sampling
- Systematic Sampling
- Cluster Sampling
- 2 angels of Data Science-Statistical Learning, Machine Learning

Module 6 – Experimentation and Evaluation, Production Deployment and Beyond

- Multi variable analysis
- linear regration
- Simple linear regration
- Hypothesis testing
- Speculation vs. claim(Query)
- Sample
- Step to test your hypothesis
- performance measure
- Generate null hypothesis
- alternative hypothesis
- Testing the hypothesis
- Threshold value
- Hypothesis testing explanation by example
- Null Hypothesis
- Alternative Hypothesis
- Probability
- Histogram of mean value
- Revisit CHI-SQUARE independence test

- Correlation between Nominal Variable

Module 7 – Various Algorithms on Business, Simple approaches to Prediction, Model Building, Deploy the model

- Machine Learning
- Importance of Algorithms
- Supervised and Unsupervised Learning
- Various Algorithms on Business
- Simple approaches to Prediction
- Predict Algorithms
- Population data
- sampling
- Disproportionate Sampling
- Steps in Model Building
- Sample the data
- What is K?
- Training Data
- Test Data
- Validation data
- Model Building
- Find the accuracy
- Rules
- Iteration
- Deploy the model
- Linear regression

Module 8 – Prediction & Analysis Segmentation

- Clustering
- Cluster and Clustering with Example
- Data Points, Grouping Data Points
- Manual Profiling
- Horizontal & Vertical Slicing
- Clustering Algorithm
- Criteria for take into Consideration before doing Clustering
- Graphical Example

- Clustering & Classification: Exclusive Clustering, Overlapping Clustering, Hierarchy Clustering
- Simple Approaches to Prediction
- Different types of Distances: 1.Manhattan, 2.Euclidean, 3.Consine Similarity
- Clustering Algorithm in Mahout
- Probabilistic Clustering
- Pattern Learning
- Nearest Neighbor Prediction
- Nearest Neighbor Analysis

Module 9 – Project 1-Cold Start Problem in Data Science, Project 2-Movie Recommendation, Conclusion

- Recommendation Algorithms
- Two Ways of Recommendation
- Recommendation Types-Collaborative Filtering Based Recommendation, Content Based Recommendation
- Cold Start Problem in Data Science
- Project 2-Movie Recommendation
- Prediction – Rating Prediction, Item Prediction
- Two Basic Approaches: Memory Based and Model Based
- What is User Based Methods in K-Nearest Neighbor?
- What is Item Based Method?
- Matrix Factorization
- Singular Value Decomposition
- Discuss on Data Science Project
- Collaboration Filtering
- What are the Business Variables?

Module 10 – Integrating R with Hadoop

- R introduction
- How R is typically used
- Features of R
- Introduction to Big data
- R+Hadoop
- Ways to connect with R and Hadoop
- Products

- Case Study
- Architecture
- Steps for Installing RIMPALA
- How to create IMPALA packages