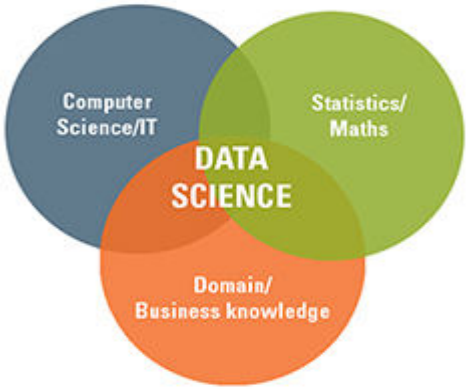

DATA SCIENCE

Machine Learning Index

Machine Learning

Machine Learning Types
Supervised
Un Supervised
Reinforcement

Data Management
Data Sources
Data Types
Data Labels

Model Building
Data
Model
Algorithm
Model Parameters (Weight, Bias)

Feature Management
Wrangling (Clean, Transform, Summarize)
Scaling and Standardizing
Selecting Important Features(RFE, Select K Best)
Extracting New Features

Machine Learning

Dimensionality Reduction (PCA, LDA)
Binning and one hot encoding

Model Error Management
Bias / Under Fitting
Variance / Over Fitting
Irreducible

Model Evaluating
Training Set
Test Set
Validation Set (K - 5, k-10 fold)
Bagging
Boosting

Model Improving
Cost Function
Gradient Descent Optimization (Batch, Stochastic, Mini)
Generalization
Regularization
Model Parameter (Weight, Bias)
Hyper Parameter (Grid, Random)
Learning Rate
Difference between Optimization and Regularization

Machine Learning

Model Automating
Pipe line
Future Union

Model Deployment
Modules (Pickle, Joblib)
Web API (Django, Flask)
REST API (web service)
Virtualization (VM ware, Docker)
Cloud (Cloud Service)

Supervised Modeling – Regression

Liner Regression (Simple, Multiple, Polynomial)
Variance, Co Variance, Correlation
Multi Co linearity
Cost Function MSE
Ordinary Least Square(OLS)
Linear Regression Rules
Metrics (EVS, RSS,MAE, MSE,R - Square, Adj R Square)
Model selection
Regularization Techniques (L1, L2, Elastic net)

Machine Learning Index

Machine Learning

Supervised Modeling – Classification

Logistic Regression
Logit / Sigmoid
Cost Function
Entropy
Maximum Log likelihood Estimation (MLE)
Gradient Descent optimization

Logistic Regression Types
Binomial,
Multinomial
Naive Bayes
LDA and QDA

Metrics (Confusion Matrix)
Model Selection (ROC AUC)
Regularization Technique

Machine Learning

Supervised Modeling - Neighbor
KNN
Lazy Learning
KNN Rules
Selecting K value
Regularization Technique

Supervised Modeling - Network
SVM
MMH (Linear case)
SVC
Kernel (Non Linear case)
Regularization Technique (C, kernel types)

Supervised Modeling - Decision Tree

Decision Tree
Parts of DT
Finding Root Node
Entropy
Information Gain
Gini Index
Variance Reduction
Regularization Technique (Pruning)

Machine Learning

Supervised Modeling - Ensemble Ensemble

Random Forest Tree
Extremely Randomized Tree (Ex Tree)
variable importance
Regularization Technique
Bagging (Sample, Boot Strap)
Boosting (Ada, Stochastic, Gradient, Xboost, Catboost)
Voting (Hard, Soft)
Stacking (Base Learner, Meta Learner)
Regularization Technique

Un Supervised Learning - Cluster

K Means (Divisive)
Hierarchical (Agglomerative)
Find K value (Elbow, Average Silhouette)
Linkage (Complete, Average, Ward (minimum))

Machine Learning Index

Machine Learning

Un Supervised Learning - Association

Rule Mining
Market Basket Association (MBA)
Apriori Algorithm (Support, Confidence, Lift)

Recommendation
Content-Based Filtering (CF)
Collaborative Based Filtering(CBF)
User Based Collaborative Filter (UBCF)
Item Based Collaborative Filter (IBCF)
Hybrid Recommendation
Matrix Factorization
Measures (Euclidian, Cosine Similarity)

Machine Learning

Time Series data

Time Series Forecasting
Component of Time Series
Types of Time Series Models
Stationary Data
Non Stationary Data
Decomposition
Differencing
Lag Term
ARIMA Model (P, D, Q)
AR(Lag) and MA (Error) Terms
ACF and PACF
ARIMA Building Steps
Time Series Analysis

Machine Learning

Text Mining

Text Mining
Toolkits
Text Processing
Word Embedding (BOW, TI-IDF)
Word 2 Vec
Text Classification
Text Summary
Sentiment Analysis
Text Clustering

Topic Modeling
LSA
LDA
NMF

DIFFERENT ALGORITHMS USES

Supervised - Regression

Both input and output labels must be provided
Use this when out put want in continues format
Useful when data in linear

Supervised - Classification

Both Input and Output labels must be provided
Use when the output want in discrete format
Useful when data is in non linear format

Neighbor - KNN

Can be used for both regression and Classification
This is Purely based on distance.
Selection of K value is important K = Sqrt(N)
normally used not easy to find optimal K clusters

Network - SVM

Neural Network based
Use this when data is too much complex to separate
Use MMH when data can be separated Linearly
Use Kernel trick when data to be separated non linear way

Tree - Decision

Splitting is based on most homogeneous variable
Follows the binary split
Entropy, Information Gain, Gini Index are the measures to check purity and impurity of the data
One on advantage data scaling not required for this

Ensemble – Random forest

Multiple Trees will be build on multiple samples, unlike a single tree built on data
Average score will taken from these models
One advantage you can use for variable selection also

DIFFERENT ALGORITHMS USES

Ensemble - Bagging

Radom sample will be collected for training the model with replacement
This is parallel approach
Normally helpful in controlling the High variance

Ensemble- Boosting

Boosting coverts the weak learners as strong learners as part of the output received from previous model
Sequential approach, best in controlling high bias

Un Supervised - Clustering

Only input labels to provided
It will create the cluster based on the k value specified
Best K value can find using Elbow, Shilheotti
Two Types of clustering K Means , Hierarchical

Unsupervised- Association

Rule mining is the technique to find out the relation between items to able to predict to what items purchased
Recommendations technique is used to recommend the products based on user or item or genre

Time Series Analysis

To forecast the time series data, we use ARIMA model
To forecast data must be in stationary format, from the data we need to separate stationary data
PACF, ACF charts used find the parameters o required by ARIMA model

Text Mining

With This you can perform Text summary, Text Classification, Text clustering, Topic modeling, sentiment analysis etc
Once data is converted to numbers, you can use this for Classification and clustering problems

Machine Learning Notes

MACHINE LEARNING SUMMARY

Wrangle | Scale | Standard

Data Wrangling - Collecting, Selecting, Cleaning, Transforming, Summarizing, visualizing, reporting data

Data Wrangling Steps

Understanding Data- Understanding structure and shape of the data
Filtering Data - Cleaning a dataset involves tasks such as removing/handling incorrect or missing data, handling outliers, and so on
Typecasting -Typecasting or converting data into appropriate data types
Transform - Transform existing columns or derive new attributes based on requirements of the use case or data itself.
Handling Categorical Data - One hot encoding and other encodings can be handled
Data Summarization - Data summarization refers to the process of preparing a compact representation of raw data at hand.
Normalizing - normalization is the process of standardizing the range of values (0 mean, 1 SD).

Feature Construction Steps

Feature Engineering- Creating new features from raw data
Feature Scaling / Standard - Keep the raw data between 0 and 1 range as well as 0 mean and 1 SD
Feature Selection- Selecting the important features (using statistical or machine learning methods)
Feature Extraction- Drawing new features from existing features by adding or removing (age from DOB)