ANALYTIXLABS

Recap

Python Foundation – Key Topics

- What is Python?
- Why python is for data science?
- Important libraries of Python for data science?
- Basic Programming elements of python?
- Data Structures and Data Types in Python?
- Indentation concept?
- What is the difference between loops & apply functions?
- Lambda function vs. built in function vs. User Defined function?
- Comprehensions (List/Dictionary)?
- Methods vs. Functions?
- What are Important data manipulation steps?
- What is Data visualization?
- Important Charts

Python Foundation – Key Topics – Basic Statistics

- Descriptive stats vs. inferential statistics?
- Population vs. sample?
- Mean vs. Median vs. Mode
- Variance vs. Standard Deviation vs. Coefficient of Variation
- Uni-Variate analysis vs. Bivariate analysis vs. Multivariate analysis?
- Estimation of population?
- Types of sampling?
- Probability Distributions (Discrete Vs. Continuous)
- Normal distribution & standardized normal distributions
- Relationships between variables?
- Central Limit Theorem
- Standard Deviation vs. Standard Error
- P-value?
- Confidence intervals
- Hypothesis testing
- Statistical Methods (t-test/z-test, f-test/anova, chi-square, correlations)

ANALYTIXLABS

Machine Learning – Key Topics – Predictive Modeling

- Analytics project stages?
  - Data → Data Importing → Data Manipulation → Analysis (descriptive/predictive/prescriptive analytics → Visualization → Reporting → Implementation
  - Descriptive analytics → Diagnostic Analytics → Predictive analytics → Prescriptive Analytics → Cognitive Analytics
  - Identify business problem → Understand possible reasons for the problem → Identify solution for the problem → Identify optimal solution → Self healing solution/Automated/Dynamic solution
- Types of statistical tests?
- What scenarios we can leverage statistical tests?
- How to identify relationship between different variables using statistical methods?
- Machine Learning vs. Traditional Learning?
- Predictive modeling?
- Types of Business problems (examples) & Characteristics
  - Regression vs. Classification vs. Segmentation vs. Forecasting vs. Optimization vs. Others
  - Strategic vs. Operational
  - Supervised vs. Unsupervised
- Role of Machine Learning?

ANALYTIXLABS

Some of the Key analytics techniques & business problems

	Business Examples	Analytics Techniques
Regression Problems Predicting a value	• Predict CTR for each ad to determine placement, based on Historical CTR, Keyword match etc • Which of the retail image levers drives footfalls or conversions? • What drives satisfaction among branch users? • What causes high performance of bank branch on the basis of financial parameters?	• Linear Regression / GLM • Decision Trees (CART) • Support vector regression • Ensembles (RF, GBM, BAGGING) • ANN
Classification Problems Predicting an event or category	• Credit approval: classify credit application as low risk, high risk, or average risk • Fraud Detection: Fraud Vs. Not Fraud • Collections: Identify cardholders that are likely to default and thus need collection effort (Payment Projection Models) • Insurance: Identify claims that are Fraud or Not Fraud • campaigns(Response/Non Response, Buying/Not Buying • Operations: Models to identify to employees who attrite(Attrition/ Retention)	• Logistic Regression /Probit Regression • Naïve Bayes/KNN • Decision Trees (CART, CHAID, C5.0) • Support Vector Machines (SVM • ANN • Ensembles (RF, GBM, BAGGING)
Segmentation Problems Classifying data into unknown number of groups	• Improve customer retention by providing products tailored for specific segments • Increase profits by leveraging disposable incomes and willingness to spend • Grow you business quicker by focusing marketing campaigns on segments with higher propensity to buy • Improve customer lifetime value by identifying purchasing patterns and targeting customers when they are in the market	• K-means/K-Medians • Hierarchical clustering • Spectral clustering • DBSCAN
Forecasting Problems Predicting Future value (depends on time)	• Call volume demand in call centers • Average handle time trends • Demand for seasonal maintenance • Event based demand for field services • Estimation of cash requirement in ATMs and Branches • Number of transactions for tellers	• Averages / Smoothening • Decomposition • ARIMA/SARIMA • ARIMAX • ARCH/GARCH/VAR

ANALYTIXLABS

Machine Learning – Key Topics – Predictive Modeling

- What is predictive modeling?
- What is model / Linear Model / Non-Linear Model?
- What is linear Regression / OLS Regression/Simple Regression/Multivariate Regression?
- What is best fit line? How system will identify best fit line?
- What is hypothetical relationship (general equation) in Linear Regression?
- Terminology:
  - Regression Analysis
  - Regression Modeling
  - Dependent Variable/Y-variable/Objective Variable/Response Variable/Target variable
  - Independent variable/X-variable/Features/Exogenious variables/Explanatory variables
  - Interpretation of beta (slope/gradient/estimate/coeficient), constant (intercept) in linear regression equation
  - SSE (Sum of squares of error), SSD(sum of squares of deviation)
- Metrics to evaluate regression model?
  - R-Square/Coefficient of determination/Goodness of fit
  - MAPE/RMSE/MSE
  - Corr(Actual, predicted)

ANALYTIXLABS

Machine Learning – Key Topics – Predictive Modeling

- How business problem solved in Machine Learning approach?
  - Business Problem → Statistical problem → Optimization problem → Solving optimization problem (Model building/Estimating the betas) → Convert optimization solution into statistical solution → Convert stats solution into business solution (Implementation)
  - How to estimate betas in linear regression
    - Converting stats problem into optimization problem (using Gradient descent)
    - Converting stats problem into linear algebra problem (using matrices operations)
- How to solve optimization problem?
  - Using Gradient Descent Algorithm / Stochastic Gradient descent algorithm
  - Objective function/Cost Function?
- Key drivers/variable importance/positive drivers/negative drivers/significant variables /hypothesis testing for model is possible or not/hypothesis testing for variable is significant or not
- What is the difference between Linear Models (LM) and Generalized Linear Models (GLM)
- Examples of GLM's?

Machine Learning – Key Topics – Modeling Stages

Typical Modeling Stages & activities in each stage: Pre-Modeling

- Identify the business problem
  - Based on the client hypothesis
  - Using Diagnostics of data
- Validate the hypothesis using Descriptive/Diagnostics
- Convert business problem into stats problem (identify type of problem)
  - (Regression vs. classification vs. segmentation vs. forecasting vs. others) or (Operations vs. Strategic) or (supervised vs. unsupervised)
- Define the Y & X
- Choose right technique (based on type of problem)
- Collect the data from multiple sources
- Consolidate data into single file by doing aggregation
  - Customer level (Customer 360 File) or Store level or product level or Car Model level etc.

Machine Learning – Key Topics – Modeling stages

Typical Modeling Stages & activities in each stage: Pre-Modeling

- Triangulate the numbers & do quick checks (Data Audit Report)
- File Level
  - Number of rows, columns
  - is the data sample/population?
  - Quick check between sample metrics vs. population metrics if the data is sample
- Variables Level
  - Data types mismatch (categorical vs. numerical vs. date vs. Boolean vs. string etc.)
  - Missing's
  - Outliers
  - Variables with low variance
  - Data with constants (like 99999) or what is meaning for zero's
  - Having special values - NA, N/A, inf, -inf, #Null, #error etc

Machine Learning – Key Topics – Modeling Stages

Typical Modeling Stages & activities in each stage: Modeling

- Data preparation-1 (Handle data related problems)
  - Handling missing's (Dropping/Imputation)
  - Handling outliers (Dropping/capping/flooring)
  - Data type conversions if required (as per the data dictionary)
  - Replace constants with some value, special values with right values
  - Creating new variables
    - Converting categorical variables into numerical
    - With calculations
  - Dropping variables with low variance (near zero variance)

Machine Learning – Key Topics – Modeling stages

Typical Modeling Stages & activities in each stage: Modeling

Data preparation-2: Assumptions check (Technique specific -Linear Regression assumptions)

- Y should follow normal (Errors follows normal distribution)
  - if it is not following, Convert Y by applying transformation such that new variables follows normality
- Y & each X should follow linear relationship
  - if x is not having linear relationship with Y, then apply transformation on x such that Y & transformed x have linear relationship with Y
- No outliers should be present in the data (both Y & X)
  - Treat the outliers if they present
- Homoscedasticity (constant variance in the data)
  - if the data is not following homoscedastic, then apply transformation on y such that it will homoscedastic
- No multicolinierity (No Relationship with in x's variables)
  - if there is multicolinierity, you need to drop the variables using correlations or VIF

Machine Learning – Key Topics – Modeling Stages

Typical Modeling Stages & activities in each stage: Modeling

Data Preparation Step-3: Feature engineering (Variable reduction/Dimension reduction/Feature reduction) - some of these steps related feature engineering we do in step 1 & 2

- Statistical methods (t-test, anova, correlation, chi-square)
- F- Regression
- Factor analysis (PCA)
- Variable Clustering
- SVD - Singular Value Decomposition
- RFE - Recursive feature elimination
- Select KBest
- Using VIF - variance inflation factor or CI

Data preparation step-4: Split the data into train (dev) and test (validation)

- Random sample of 70%/80% - training (Training & testing - Mutual exclusive)
- Random sample of 30%/20% - testing

Machine Learning – Key Topics – Modeling stages

Typical Modeling Stages & activities in each stage: Model Building & validation

- Model building on the train data
  - Finalize the mathematical equation
  - For train data, calculate metrics like R-square/RMSE/MSE/Corr(Actual, pred)
- Validate your model using test or new data
  - Using the equation, Predicting values for test data (called as scoring)
  - For test data, calculate metrics like R-square/RMSE/MSE/Corr(Actual, pred)
- Compare metrics between train & test
  - Metrics are coming similar for both train & test, you can proceeds further
  - if not, iterate the process

Machine Learning – Key Topics – Modeling stages

Typical Modeling Stages & activities in each stage: Model building & validation

- Training accuracy is high & testing accuracy is low (Over fitting) - Possible reasons for over fitting
  - Having lots of variables in the model --> high complexity --> high variance
  - Having lot of transformations --> high complexity --> high variance
  - Not following assumptions
  - Data preparation is not proper
  - Less Data (less number of observations)
- Training accuracy is it self low (Under fitting) - Possible reasons for under fitting
  - Not chosen right technique for given data
  - Not following assumptions of model
  - Data preparation is not proper
  - Not tuned the hyper parameters properly

Machine Learning – Key Topics – Modeling stages

Typical Modeling Stages & activities in each stage: Post Modeling

- Implementation code (data preparation, scoring etc.)
- Pros & cons of model & Assumption of the model
- Documentation of entire process
- Create Deck of outputs
  - Data audit report
  - Data preparation steps
  - Variable reduction steps (Need to provide, which variables got removed from each stage)
  - Final equation
  - Variable importance or key drivers (positive/negative)
  - Validation metrics
- Business validation
  - Is the equation making sense?
  - Are the feature's coefficients making sense in terms of negative/positive
- Tool for implementation of model (Excel Based/Web Based/Package/Create Containers etc…)
- Tracking the model (Model validation)
  - Model maintenance (Calibrate the model)

ANALYTIXLABS

Machine Learning – Key Topics – Modeling stages

- Why do we need to take care of multicolinierity (Consequences of multicolinierity)?
- Why do we need to worry about outliers presence in the data (Consequences of outliers)?
- How to perform business validation of predictive model? (Why it is required)?
- How do we implement predictive model at business?
- How do we deploy the models at enterprise level?
  - Using Docker Containers (image of entire environment)
  - Using Pickle objects (pickle is package in python which can help to export model into objects, so that it can be imported when ever you need it with out running the model)
  - Program of mathematical equation (implementation code)
- How to track the model impacting business over the time? (Model Monitoring)
- What is model maintenance?

ANALYTIXLABS

Machine Learning – Key Topics – Modeling Stages

- Bias & Variance in the data? How these will effect?
- How to handle Bias & Variance in the data?
- What is cross validation?
  - K-Fold validation
  - Leave one out Validation
  - Train vs. Test vs. Validate
- What is regularization of models?
- Regularization techniques (Ridge vs. Lasso vs. Elastic net)?
- What is meaning of tuning the parameters?

ANALYTIXLABS

Machine Learning – Key Topics – Classification Problem

- What is Classification problem? Types of classification problems?
- Example of classification problems?
- Why not linear regression for solving classification problem?
- What is logistic regression?
- Linear Regression vs. Logistic Regression? How they are different from each other?
  - Y definition / Assumptions/ Variable reduction techniques /Betas Estimation (MLE)/ Metrics
- What is odds ratio? How to interpret odds ratio?
- Assumptions of logistic regression?
- What is logit/log(odds)/WOE?
- How did we derived logit transformation for solving classification problem as part of GLM? (By assuming relationship between Y & X as sigmoid-curve)

ANALYTIXLABS

Machine Learning – Key Topics – Classification Problem

- General equation of sigmoid curve? How to achieve sigmoid curve?
- Hypothetical relationship between Y & X in logistic regression?
- What are the metrics used for evaluating classification model?
  - Based on probability (concordance, Discordance, Ties, Somerce-D (Gini), Gamma, AUC)
  - Based on category as prediction (Misclassification metrics/Confusion Metrics/ Classification Report, Sensitivity, Specificity, Accuracy, Precision, Recall, f1-score etc)
  - Other metrics like KS, Lift, rank ordering, gains table etc.
- How to decide best cut-off for converting probability into prediction category?
  - Percentage of 1's in training data
  - Based on sensitivity, specificity, accuracy, sensitivity + specificity, ROC curve etc
  - Based on KS Table (KS Value)
  - Business Logic

Machine Learning – Key Topics - Segmentation

- What is segmentation?
- Applications of segmentation
- Types of segmentation
  - Heuristic vs. Scientific
  - Objective vs. subjective
  - Behavioral vs. Need based
- Types of techniques
  - Value based vs. life stage vs. RFM
  - Kmeans vs. Kmedians vs. Kmodes vs. Ward (Hierarchical) vs. DBSCAN Vs. Spectral
- Process of segmentation (steps for segmentation)
- What is profiling & How to do it? Use of Profiling?
- How to identify characteristics of segmentation?
- What are the pre-requisites of Kmeans Clustering?
- Types of distance algorithms
  - Euclidian vs. City Distance vs. Cosine similarity
- How to decide K value in Kmeans segmentation?
  - Using Metrics like Silhoute coefficient, Pseudo F-value (Elbow method), Dendogram
  - Profiling to identify characteristics
  - Best practices like Distribution of segmentation

Machine Learning – Key Topics - Segmentation

- What are the Typical modules in python for implementing clustering?
- How to implement segmentation
- What are the dimensions reduction techniques used till now?
  - VIF, F-regression, RFE, Uni-variate regression, WOE, Statistical tests
  - Principle component analysis (PCA/Factor analysis)
- What is factor analysis? How it will be helpful?
- Difference between factor analysis vs. Cluster analysis?
- What is eigen value?
- What is eigen vector?
- Prerequisites for factor analysis?
- How to choose number of factors?
  - Based on Eigen value / Based amount of variance explained
- What are factor loadings? How it helps?

Machine Learning – Key Topics - Segmentation

- Types of clustering algorithms
  - Kmeans vs. DBSCAN vs. Hierarchical vs. Spectral
- The basic idea of density-based clustering
- What is DBSCAN?
- Terminology for DBSCAN?

epsilon neighborhood, core point, boundary point, noise point/outlier point, min points, high density, low density, density reachability, densely connected points,

- The two important parameters and the definitions of neighborhood and density in DBSCAN
- DBSCAN's pros and cons
- What is Hierarchical clustering? Basic idea of clustering?
- How to decide number of clusters in Hierarchical clustering?
  - Dendogram

Machine Learning – Key Topics - Forecasting

- Types of data?
  - Cross Sectional Data vs. Time Series Data vs. Panel Data
- What is Forecasting?
- Key characteristics of time series?
- Key terminology of time series
  - Time Series, Time Series Model, Time Series analysis
  - Components of time series (Base, Trend, Seasonality, Cyclicity, irregular)
  - Decomposition (Additive, multiplicative)
  - Lag, Lead, Difference
  - Auto correlation, Auto correlation function(ACF), Partial auto correlation function(PACF), Auto Regression, White noise, Random walk
  - Stationary Series, ADF Test (Augmented Dicky Fuller Test)/Unit Root Test
- How to check series is stationary or not?
  - Using line chart
  - Using ADF test (H0: Series is not stationary, Ha: Series is stationary)
  - Over the time, calculate mean & variance and compare them
- How to convert non-stationary series into stationary series?
  - Log()
  - Differencing
  - De-Trending(removing trend from data)
  - De-Seasonlization(removing seasonality from data)

ANALYTIXLABS

Machine Learning – Key Topics - Forecasting

- Types of techniques available for solving forecasting problems
  - Uni-Variate Time Series:
    - Basic: Averages (MA, WMA, CMA), ETS Models (Holt winter models – Exponential smoothening (single, double, triple), Decomposition
    - Medium: ARIMA Family (AR, MA, ARMA, ARIMA, SARIMA)
    - Complex: ARCH, GARCH, VAR, Wavelets
  - Multivariate Time Series:
    - ARIMAX, SARIMAX, Regression, ANN
  - Panel Data:
    - Hierarchical Models (Mixed Models), Fixed effect models, Random effect models
- Metrics for evaluating forecasting model?
  - MAPE, MSE, RMSE
- How to split the data into train & test?
  - Based on the time
- Key packages in python for forecasting?
  - Pandas, Statsmodels, prophet

ANALYTIXLABS

Machine Learning – Key Topics - Forecasting

- ARIMA vs. SARIMA vs. ARIMAX vs. SARIMAX
- What is SARIMAX?
  - Seasonal ARIMA with X variables (Seasonal Auto Regressive Integrated moving average with X variables)
- What are the parameters of SARIMAX()?
- Stages of ARIMA model?
  - Identification (identify values for p, d, q)
  - Estimation (Estimate the beta's)
  - Forecasting (Forecast after estimation)
- What are the uses of ACF?
  - To check series is white noise/random walk?
  - To check model is possible or not?
  - To check series is stationary or not?
  - To check Seasonality exists or not?
  - To find model which model is it? (AR, MA, ARMA, SARIMA)?
- How to identify p, d, q values?
  - Box-Jenkins methodology?
  - Using iterative process (which ever combination of p, d, q values gives minimum AIC)
  - Hybrid approach

ANALYTIXLABS

Machine Learning – Key Topics – Introduction to ML

- What is Machine Learning?
- Machine Learning vs. Traditional Learning?
- Dynamic models vs. Static Models?
- Bias & Variance?
- Objectives of machine learning?
  - Dynamic Models
  - Automated
  - High accuracy
  - Low Bias & Low Variance
  - No over fitting (stable model), No under fitting
- How to handle over fitting problem?
  - Cross validation
  - Regularization
- How to handle under fitting problem?
  - Tuning hyper parameters
- What is GridsearchCV?

ANALYTIXLABS

Machine Learning – Key Topics – Introduction to ML

- Different types of optimization algorithms
  - Gradient Descent Algorithm
  - Stochastic Gradient Descent Algorithm
  - Mini Gradient (Batch wise – Batch size ranges from 50-256 points)
- Types of estimation process
  - MSE (Minimize sum of squares of errors)
  - MLE (Maximum Likelihood estimation)
- What is pipelining in python? How it will be helpful?
- What are the use cases covered?
  - Brest cancer prediction, Prediction of employee leave or not, image classification
- What are the python modules available and used?
  - ANN: sklearn.neural_network, sklearn.preprocessing, keras, theano, tensorflow
  - SVM: sklearn.svm
  - Pipeline: sklearn.pipeline
  - Inbuilt data sets: sklearn.datasets

Machine Learning – Key Topics – ML Steps

Machine Learning Steps (Objective):

- We're trying to build models that neither Over fit or Under fit
- Over fitting/Under fitting is directly affected by the model's complexity
  - A highly complex model might over fit. (High Variance)
  - A too simple model might under fit. (High Bias)
- Model Complexity is controlled by its Hyper parameters.
- So, we have a SEARCH PROBLEM at hand.
- Want to find the optimum combination of model hyper parameters that neither over fit nor under fit the data.
- The only way to do this is with Trial and Error.
- So, we set up a grid containing all possible (reasonable) combinations of Hyper parameter values.
- We then fit models to each combination, and find the (cross validated) Performance of each combination of hyperparamter values.
- GridSearchCV *automates* this process for me. All I need to specify are –
  - The model to use, The Grid, The performance metric to use (depending on the model), The cross validation scheme to use (3- or 5- or 10- fold CV)
- GridSearchCV finds the Best Model that optimizes the Performance Metric.

Machine Learning – Key Topics – Decision Trees

- What is Decision Tree? & Uses of Decision trees?
- Types of decision trees?
  - Regression Trees vs. Classification trees
- Terminology in decision trees
  - Root node, child node, leaf node, splitting criteria, stopping criteria, growing tree, pruning tree, lift, feature importance
- Splitting criteria for decision trees?
  - Regression trees – Based on ANOVA, Minimize MSE
  - Classification trees – Based on Chisqure, Gini, Entropy, Information Gain etc.
- Stopping criteria's for decision trees?
  - maxdepth
  - max_leaf_nodes
  - max_features
  - min_samples_split
  - min_samples_leaf
  - Improvement in splitting criteria
  - Complexity parameter
- Types of decision trees algorithms
  - CHAID (chisqure automatic interaction detection trees)
  - CART (classification & regression trees)
  - C5.0 (Classification trees) (ID3 – Iterative Dichotomizer)
  - Quest (Quick, Unbiased and Efficient Statistical Tree.)

Machine Learning – Key Topics – Decision Trees

- Advantages with decision trees
  - Easy to implement
  - Easy to explain
  - Don't need lot's of data preparation
  - Processing is relatively faster
  - Easy to score at enterprise level with rules
  - It will be used as intermediate technique
    - Identify the important variables (variable reduction)
    - Convert categorical with more number of categories into small number of categories
    - Convert numerical variable into bins (categorical)
- Disadvantages with decision trees
  - Because of high variance, over fits
  - Group of observations predicted as same value, so we are unable to differentiate among them

Machine Learning – Key Topics – Ensemble Learning

- What is ensemble learning?
- Types of ensemble learning?
  - Homogeneous vs. Heterogeneous
- Types of Ensemble algorithms using trees? & Their tuning parameters
  - Bagging (Bootstrap aggregating)
  - Random Forest (RF)
  - Adaboost (Adaptive boosting)
  - Gradient Boost Machine (GBM)
  - Extreme Gradient Boost (XgBoost)
- Bagging vs. Random Forest
- Adaboost vs. Gradient Boost vs. Xgboost
- Parallel models vs. sequential models?
- Key packages in Python to implement decision trees & ensemble learning algorithms?
  - Sklearn.tree, sklearn.ensemble, sklearn.metrics, pydotplus, pandas etc.

ANALYTIXLABS

Machine Learning – Key Topics - SVM

- What is SVM?
- Key Terminology in SVM
  - Margin, maximum margin classifier, Hard margin, Soft margin, Kernel function, Support vectors, Hyper plane, linearly separable, linearly inseparable, optimization
- Types of SVM?
  - Linear SVM vs. Non Linear SVM vs. Kernel SVM
- What is Kernel? Why do we need it?
- Types of Kernels
  - Linear vs. sigmoid vs. polynomial of power P vs. radial basis function (rbf/Gaussian) vs. chi-square, string etc.
- What is optimization function in LSVM?
- Tuning parameters for SVM
  - Kernel: shape of separators (linear vs. rbf vs. poly). RBF is the default setting and is used to create a non-linear hyper plane. The same goes for poly. Linear is used to create a linear hyperplance
  - gamma: defines influence of a single training example. a higher gamma value allows close data points to greatly influence the decision boundary. The opposite is true for low gamma values.
  - C: Defines tradeoff between smooth decision boundary and classifying points correctly. A large C value leads to a higher accuracy.

ANALYTIXLABS

Machine Learning – Key Topics - ANN

- What is ANN? Mapping of different ANN algorithms with Human Brain functionalities
- Types of ANN
  - Single layer vs. Multi Layer vs. recurrent
  - ANN vs. DNN (Deep Neural Networks)
  - ANN vs. CNN vs. RNN vs. LSTM
- Terminology in ANN
  - Perceptron, network, Neuron/logistic unit/computation unit/nodes, Activation function, adder function, input layer, output layers, hidden layers, MLP, threshold, learning rate
- What is activation function? Types of activation functions?
  - RELU, Sigmoid, Threshold, hyperbolic tangent, linear
- Key optimization techniques (solver for weight optimization)
  - 'lbfgs' is an optimizer in the family of quasi-Newton methods.
  - 'sgd' refers to stochastic gradient descent.
  - 'adam' refers to a stochastic gradient-based optimizer
  - 'adam' works pretty well on relatively large datasets (with thousands of training samples or more) in terms of both training time and validation score. For small datasets, however, 'lbfgs' can converge faster and perform better.
- Types of Betas estimation process
  - Backward propagation
  - Forward propagation
- Tuning parameters in ANN
  - Learning Rate, Activation function, Optimization function, number of layers & number of neurons

ANALYTIXLABS

Machine Learning – Key Topics – Text Mining

- Types of data?
- What is Text Mining?
- Data Analytics vs. Text Analytics
- Applications of Text Mining in different industries?
- Descriptive mining vs. predictive mining in text mining
- Key terms in text mining.
  - Information Extraction, Corpus, Documents/Words/Tokenization, TF, TF-IDF, Cosine Similarity, DTM/TDM/DFM/ Vectorization, BOW, n-grams, non-textual data (white spaces, numbers, punctuations etc.), Disambiguation, POS Tagging, Stemming, Lemmatization, Stop words, Sparse terms/Rare Terms, NER, Entity, sentiment/Polarity/ Opinion mining/Contextual Text Mining, Intent, word cloud, NLP/DNLP, NLG, NLU, LDA/LSA, topic modeling/Concepts mapping, SNA, Community detection, categorization
- Text mining process steps?
- What is NLP?

Types of Analysis

- Intent Analysis
- Sentiment analysis(Lexicon/Classification)
- Document Classification
- Text Summarization
- Segmentation - Identify inherent themes
- Social Network analysis - Community detection

ANALYTIXLABS

Machine Learning – Key Topics – Text Mining

- **Types of text Analytics we performed**
  - Supervised: Classification (sentiment, intent, document classification etc.)
  - Unsupervised: Segmentation (Kmeans, topic modeling, PCA etc)
  - Visualizations: Wordclouds, BarGraphs, Frequency Analysis etc…

- **Data Processing - What are some of the lower level components?**¶
  - **Tokenization**: breaking text into tokens (words, sentences, n-grams)
  - **Stopword removal**: a/an/the
  - **Stemming and lemmatization**: root word
  - **TF-IDF**: word importance
  - **Part-of-speech tagging**: noun/verb/adjective
  - **Named entity recognition**: person/organization/location
  - **Spelling correction**: "New Yrok City"
  - **Word sense disambiguation**: "buy a mouse"
  - **Segmentation**: "New York City subway"
  - **Language detection**: "translate this page"
  - **Machine learning**

ANALYTIXLABS

Machine Learning – Key Topics – Text Mining

Why is NLP hard?

- **Ambiguity**:
  - Hospitals are Sued by 7 Foot Doctors
  - Juvenile Court to Try Shooting Defendant
  - Local High School Dropouts Cut in Half
- **Non-standard English**: text messages
- **Idioms**: "throw in the towel"
- **Newly coined words**: "retweet"
- **Tricky entity names**: "Where is A Bug's Life playing?"
- **World knowledge**: "Mary and Sue are sisters", "Mary and Sue are mothers"
- NLP requires an understanding of the **language** and the **world**.

ANALYTIXLABS

Machine Learning – Key Topics – Text Mining

Feature Engineering

- ***TF-IDF Vectors as features***
- TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
- IDF(t) = log_e(Total number of documents / Number of documents with term t in it)
- TF-IDF Vectors can be generated at different levels of input tokens (words, characters, n-grams)
  - a. Word Level TF-IDF : Matrix representing tf-idf scores of every term in different documents
  - b. N-gram Level TF-IDF : N-grams are the combination of N terms together. This Matrix representing tf-idf scores of N-grams
  - c. Character Level TF-IDF : Matrix representing tf-idf scores of character level n-grams in the corpus

Text / NLP based features

- Word Count of the documents – total number of words in the documents
- Character Count of the documents – total number of characters in the documents
- Average Word Density of the documents – average length of the words used in the documents
- Punctuation Count in the Complete Essay – total number of punctuation marks in the documents
- Upper Case Count in the Complete Essay – total number of upper count words in the documents
- Title Word Count in the Complete Essay – total number of proper case (title) words in the documents
- Frequency distribution of Part of Speech Tags:
  - Noun Count, Verb Count, Adjective Count, Adverb Count, pronoun Count

ANALYTIXLABS

Machine Learning – Key Topics – Text Mining

Model Building

- Naive Bayes Classifier
- Linear Classifier
- Support Vector Machine
- KNN
- Bagging Models
- Boosting Models
- Shallow Neural Networks
- Deep Neural Networks
  - Convolutional Neural Network (CNN)
  - Long Short Term Model (LSTM)
  - Gated Recurrent Unit (GRU)
  - Bidirectional RNN
  - Recurrent Convolutional Neural Network (RCNN)
  - Other Variants of Deep Neural Networks

ANALYTIXLABS

Machine Learning – Key Topics – Text Mining

- Packages in python for text mining?
  - textmining1.0: contains a variety of useful functions for text mining in Python.
  - NLTK: This package can be extremely useful because you have easy access to over 50 corpora and lexical resources
  - Tweepy: to mine Twitter data
  - scrappy: extract the data you need from websites
  - urllib2: a package for opening URLs
  - requests: library for grabbing data from the internet
  - Beautifulsoup: library for parsing HTML data
  - re: grep(), grepl(), regexpr(), gregexpr(), sub(), gsub(), and strsplit() are helpful functions
  - wordcloud: to visualize the wordcloud
  - Textblob: package to create blob object and perform text processing

Machine Learning – Key Topics – Recommender systems

- Recommender System vs. Recommendation Engine
- Beer-Diapers Story
- Target Cross Sell Story related to Pregnancy
- Content/context Based Recommender system
- Memory based/model based
- Market Basket analysis/Collaborative filtering

Machine Learning – Key Topics – Recommender systems

- Market Basket analysis/Collaborative filtering
- Association analysis/Association Rules
- Support, Lift, Confidence, conviction
- Apriori, Eclat, FP-Growth algorithms
- Memory based/model based
- Similarities(jaccard, cosine, correlations)
- K-Nearest Neighbors (KNN)
- User Based CF, Item Based CF
- Ratingmetrix/realratingmetrix (user behavior metrix), user-feature vector, item-feature metrix
- SVD/Matrix factorization
- Precision, Recall, F1-Score, Sensitivity, Specificity, AUC, ROC Curve, RMSE, MSE, MAE

ANALYTIXLABS

Contact us

Visit us on: http://www.analytixlabs.in/

For course registration, please visit: http://www.analytixlabs.co.in/course-registration/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 88021-73069

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/

ANALYTIXLABS