What is machine learning?

**What are the applications of Machine Learning?**

Companies like Google, Facebook, Baidu, IBM, Microsoft uses ML extensively to push their respective ads to the relevant users.

Here are a few applications that you should know:

1. Banking & Financial services: ML can be used to predict the customers who are likely to default from paying loans or credit card bills. This is of paramount importance, as machine learning would help the banks to identify the customers who can be granted loans and credit cards.
2. Healthcare: It is used to diagnose deadly diseases (e.g. cancer) based on the symptoms of patients and tallying them with the past data of similar kind of patients.
3. Retail: It is used to identify products which sell more frequently (fast moving) and the slow moving products which help the retailers to decide what kind of products to introduce or remove from the shelf. In addition, machine-learning algorithms can be used to find which two / three or more products sell together. This is done to design customer loyalty initiatives, which in turn helps the retailers to develop and maintain loyal customers.

**Types of Machine learning**

**Support vector Machines**

**Hidden Markov model**

**RandomForest**

**GradientBoostingTrees (**random set of features or all used, fraction of samples for overfitting**)**

Deep Learning

Deep learning is a family of advanced machine learning algorithms which used the concept of human brain neurons to model arbitrary functions to capture the real world non-linear complexities which traditional machine learning algorithms are not capable of. These algorithms require lot of data to perform.

Libraries like Theano and Tensorflow help in specific machine learning number-crunching operations like derivatives on huge matricies with large efficiency

Tree based Modeling

#TODO - https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

**Deep Learning, NLP and intersection**:-

Deep Learning – is a subfield of machine learning. it is a family of advanced machine learning algorithms which used the concept of human brain neurons to model arbitrary functions to capture the real world non-linear complexities which traditional machine learning algorithms are not capable of. These algorithms require lot of data to perform.

It attempts to learn representations and output. It can learn unsupervised(from raw text) or supervised(with specific labels)

Deep learning in speech

Deep learning in vision

NLP – Natural Language Processing is a field at the intersection of computer science, artificial intelligence and linguistics.

Reason for Intersection of Deep Learning and NLP is that it will help computers to understand natural language like humans to perform useful tasks e.g. Question-Answering

Following steps can be performed in general:-

1. Speech (phonetic or phonological analysis) or text(OCR/tokenization)
2. Morphological Analysis
3. Syntactic Analysis
4. Semantic Interpretation
5. Discourse Processing

Applications of NLP – spell checking, keyword search, finding synonyms, extracting info from websites such as product prices, names, date, location etc., classification on reading level of text , sentiment analysis and identification on positive or negative. More complex tasks like machine translation, complex question answering or spoken dialogue systems.

NLP in industry –

1. Search (written and spoken)
2. Online Advertisements
3. Automated translation
4. Sentiment Analysis for marketing/trading/elections
5. Speech recognition
6. Automating customer support

Why NLP is complex?

NLP disambiguos examples

Deep Learning + NLP = Deep NLP

Earlier semantics were represented by Lambda Calculus but with deep learning its just a vector and calculations

Also traditional approaches involve using bag of words representation or hand designated negation features.

Question Answering –

A lot of feature engineering to capture real world e.g. regular expressions

**Word Vectors**

**Count based(traditional) vs Prediction based(unconventional)**

1. Document word concurrence matrix - lead to general topics
2. Window based concurrence matrix – captures both syntactic and semantic information

Dimensionality reduction on concurrence matrix X with the help of singular value decomposition

A word can be represented by a dense vector. How many dimensions to use is something an important decision based on the nature of the problem.

SVD doesn’t scale and computation time increases quadrtically

Word2vec is something as state of the art. Predict surrounding words in a window of length m of every word

**Sampling**

It is a statistical procedure concerned with the selection of the individual observation to make statistical inferences about the population.

Types of sampling –

Random sampling – simple random sampling, equal probability systematic sampling, stratified simple random sampling, multistage stratified random sampling, cluster sampling, Multistage cluster sampling,

Non-random sampling – availability sampling, quota sampling, expert sampling,

Probability or non-probability sampling

**Important Concepts:-**

**26-01-2017**

1. Mahalanobis distance – It is best suited to detect outliers in n-dimensions where n >1. Mahalanobis’ distance is a statistical measure of the extent to which cases are multivariate outliers, based on a chi-squared distribution. R's mahalanobis() function provides a simple means of detecting outliers in multidimensional data.

#TODO - <https://www.r-bloggers.com/outlier-detection-with-mahalanobis-distance/>

1. Goodness-of-fit - A goodness-of-fit test, in general, refers to measuring how well do the observed data correspond to the fitted (assumed) model

**Chi Square test, Kolmogorov–Smirnov test, Cramér–von Mises criterion**

#TODO - <https://www.r-bloggers.com/goodness-of-fit-test-in-r/>

1. Bootsrap sampling - If we don’t have enough data to train our algorithm then we can increase the size of our training set by randomly selecting items and duplicating them (with replacement).

#TODO - <http://www.statisticssolutions.com/sample-size-calculation-and-sample-size-justification/sampling/>

1. Adjusted Rand Score – Given the knowledge of the ground truth class assignments labels\_true and our clustering algorithm assignments of the same samples labels\_pred, the adjusted Rand index is a function that measures the similarity of the two assignments, ignoring permutations and with chance normalization. ARI requires knowledge of the ground truth classes while is almost never available in practice or requires manual assignment by human annotators

ARI = (RI - Expected\_RI) / (max(RI) - Expected\_RI)

#TODO - <http://scikit-learn.org/stable/modules/clustering.html#adjusted-rand-score>

1. For K-fold cross validation, what k should be selected?

#TODO - <http://stats.stackexchange.com/questions/61783/variance-and-bias-in-cross-validation-why-does-leave-one-out-cv-have-higher-var>

<https://www.quora.com/For-K-fold-cross-validation-what-k-should-be-selected>

1. How to deal with multi-collinearity situation in linear regression model? - To check multicollinearity, we can create a correlation matrix to identify & remove variables having correlation above 75% (deciding a threshold is subjective). In addition, we can use calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.But, removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variable so that the variables become different from each other. But, adding noise might affect the prediction accuracy, hence this approach should be carefully used.
2. Threshold in classification problems(Confidence Splitting criteria) - <http://nerds.airbnb.com/confidence-splitting-criterions/>
3. Accuracy paradox in an imbalanced dataset – Accuracy should not be used as the measure of model performance because accuracy might be only predicting major class correctly but we are interested only in the minor class. Therefore, we should use sensitivity i.e. true positive rate (also called as recall), specificity i.e. true negative rate, and precision i.e. positive predicted value). F measure to determine class wise performance of the classifier.

#TODO - <https://en.wikipedia.org/wiki/Sensitivity_and_specificity>

#TODO - <https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/>

1. Ensemble Learning – use multiple algorithms for prediction and they combine output of multiple machine learning algorithms for getting more robust or generalized output which outperform all the individual models

#TODO - <https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/>

#TODO - <https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning/>

#TODO - <http://mlwave.com/kaggle-ensembling-guide/>

1. R-squared and adj. R-squared – R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why we must assess the residual plots. However, R-squared has additional problems that the adjusted R-squared and predicted R-squared are designed to address.Every time you add a predictor to a model, the R-squared increases or remains same.

#TODO - <https://discuss.analyticsvidhya.com/t/difference-between-r-square-and-adjusted-r-square/264/3>

**27-01-2017**

1. Entity disambiguation –

#TODO - <http://www.theatlantic.com/technology/archive/2011/03/does-anne-hathaway-news-drive-berkshire-hathaways-stock/72661/>