Fake News Detection Using ML

Yash Mirge (172010021)

Introduction

- With the growing popularity of mobile technology and social media, information is accessible at one’s fingertips.
- Mobile applications and social media platforms have overthrown traditional print media in the dissemination of news and information.
- Fake news is typically published with an intent to mislead or create bias to acquire political or financial gains. Hence it may tend to have during headlines or interesting content to increase viewership.
- The main challenge throughout the project has been to build a set of uniform clean data and to tune parameters of our algorithms to attain the maximum accuracy.

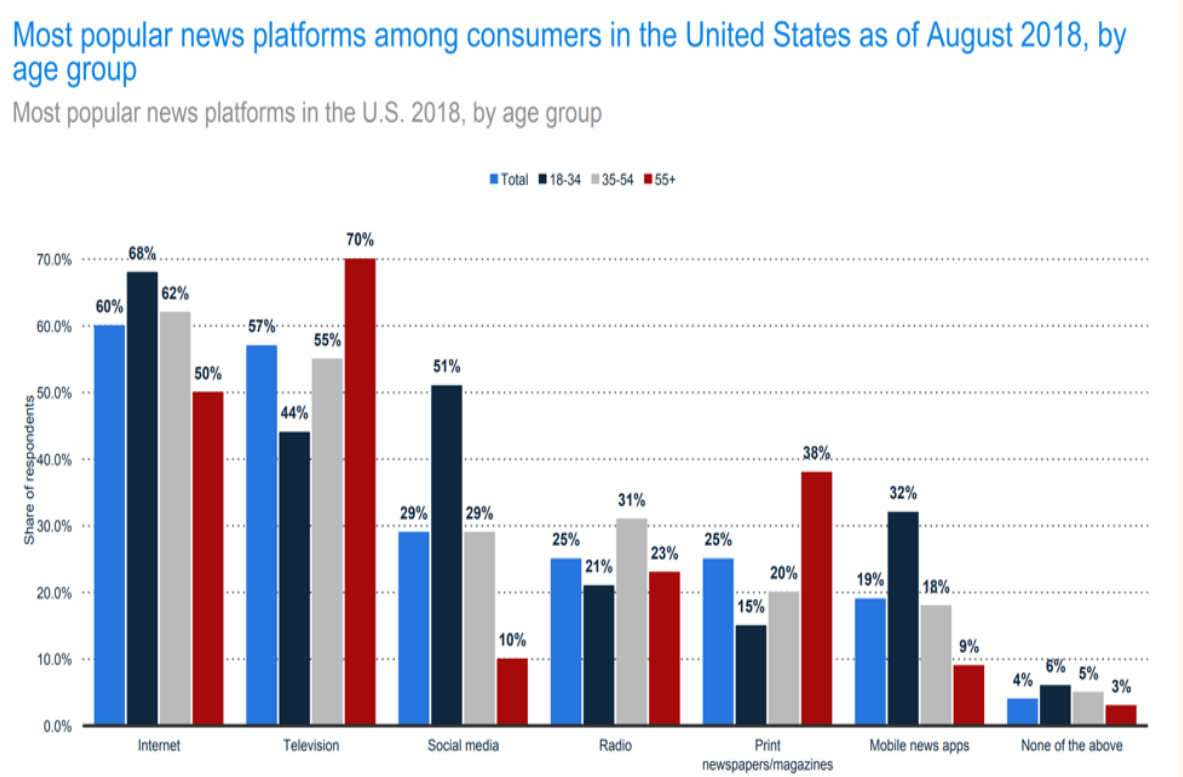
Literature Reviews

- In general , Fake news could be categorized into three groups :
  - The First group is fake news , which is news that is completely fake and is made by the writers of the articles.
  - The second group of fake news are those whose main purpose is to provide humour to the reader.
  - The third group is poorly written news articles , which have some degree of real news , but they are not entirely accurate.

- In short , it is news that uses, for example, quotes from political figures to report a fully fake story. Usually, this kind of news is designed to promote certain agenda or biased opinion.

Importance of the study: How to spot it.

- Sources of News
- Evaluate and Verify
  - Who is creator ?
  - What is the message ?
  - Why was this created ?
- No intention to cause harm but has potential to fool
- Data analysis and machine learning can play important role in detection.



Dataset Overview

- Most researches on this topic use dataset from kaggle .
- The dataset comprised of 5 attributes (Id , Title , Author , text , Label)
  - id: unique id for a news article
  - title: the title of a news article
  - author: author of the news article
  - text: the text of the article; could be incomplete
  - label: a label that marks the article as potentially unreliable
    - 1: unreliable
    - 0: reliable
- This dataset is training dataset.

Major contribution of this study

- Extraction of classified accuracy useful for fake news detection.
- Removal of redundant and irrelevant data .
- Comparison of different machine learning algorithm on fake news dataset.
- Identification of the best performance-based algorithm for fake news detection.

Approaches

- Our approach is trained on the large-scale, noisy dataset, using different machine learning algorithms.
- All of those methods expect the representation of a tweet as a vector of features.
- consider five different groups of features.
  - User-level features
  - Tweet-level features
  - Text features
  - Topic features
  - Sentiment features.

User-Level features

- For the user, we first collect all features that the Twitter API directly returns for a user, e.g., the number of followers. Furthermore, we use the API to create additional statistics modeling the user's behavior on Twitter, e.g., the frequency of tweets or the ratio of retweets.

Tweet-Level features

- For tweet-level features, we again use the Twitter API to first collect all information directly available (e.g., number of retweets), and add meta information (e.g., weekday and time) as well as statistical information on the contents (e.g., word count, ratio of question and exclamation marks).

Text features

- For representing the textual contents of the tweet, we used a bag of words (BOW) model

Topic features

- We used topic feature to separate similar topic in one single group.
- we also apply topic modeling for creating features from the tweets.

Sentiment features

- we used **SentiWordNet** to compute the polarity of tweets in terms of ratio of positive, negative, and neutral words.

Evaluation

- Setting 1: Cross-validation on Training Dataset
  - Remove of redundant and irrelevant data .
  - we expect the results of actual performance on a correctly labeled training dataset.
  - For achieving appropriate result we used various classification algorithm.

- Setting 2: Validation against Standard
  - Validation against standard dataset was collected data independently from the training dataset and is never used for training, feature selection, or parameter optimization, we can safely state that our approach is not overfit to that dataset.

Methodologies : General Approaches

1. Logistic Regression
2. K-Nearest Neighbors (KNN)
3. Support Vector Machine (SVM)
4. Random Forest
5. Naive Bayes

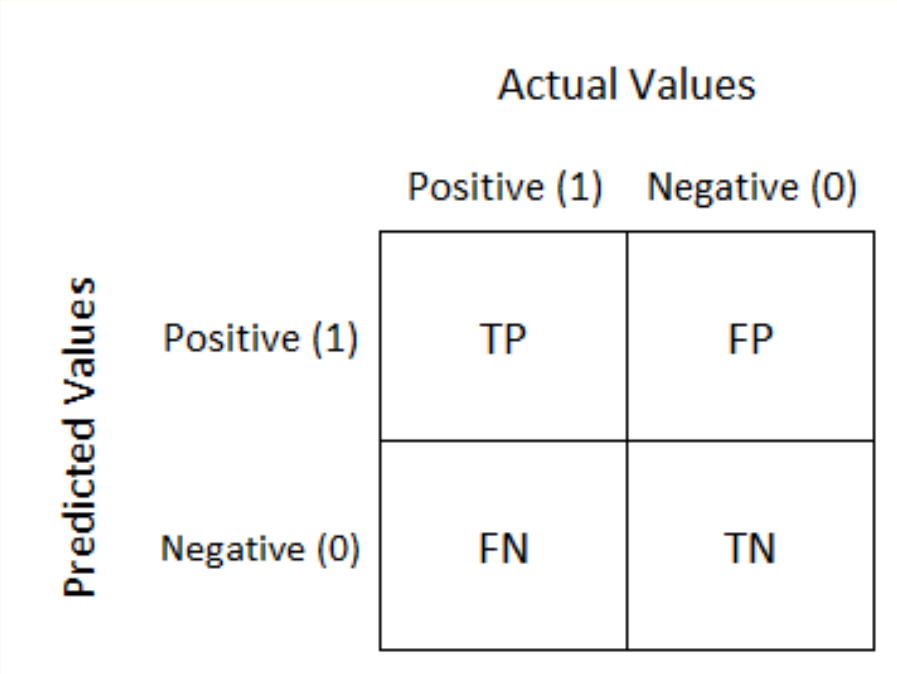
General Approaches : Logistic Regression

- Logistic regression is used to find the probability of event (Success / Failure).
- Logistic Regression can be used for binary classification.
- Logistic Regression uses a decision boundary to map data to probabilities.
- The decision as to the data point lies in which class is made on the basis of a decision boundary.
- To avoid over fitting and under fitting , we used logistic regression

When we make a binary prediction, there can be 4 types of outcomes:

- We predict **FAKE** while we should have the class is actually **FAKE**: this is called a **True Negative**.
- We predict **FAKE** while we should have the class is actually **TRUE** this is called a **False Negative**.
- We predict **TRUE** while we should have the class is actually **FAKE** this is called a **False Positive**.
- We predict **TRUE** while we should have the class is actually **TRUE**: this is called a **True Positive**.

Confusion Matrix



SVM and KNN

- SVM and KNN are supervised learning classification algorithms.
- SVM performs a classification by finding the hyperplane that differentiate the two classes.
- Can be tuned by using regularization parameters.
- In KNN, predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors).

Random Forest Algorithm

- The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves.
- It merges the decisions of multiple decision trees in order to find an answer, which represents the average of all these decision trees.
- The random forest algorithm is a supervised learning model; it uses labeled data to "learn" how to classify unlabeled data.
- The Random Forest Algorithm is used to solve both regression and classification problems

Naive Bayes Classification

- Bayes' theorem with an assumption of independence between predictors.
- In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayesian model is easy to build and particularly useful for very large data sets.
- For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple.

Future Work

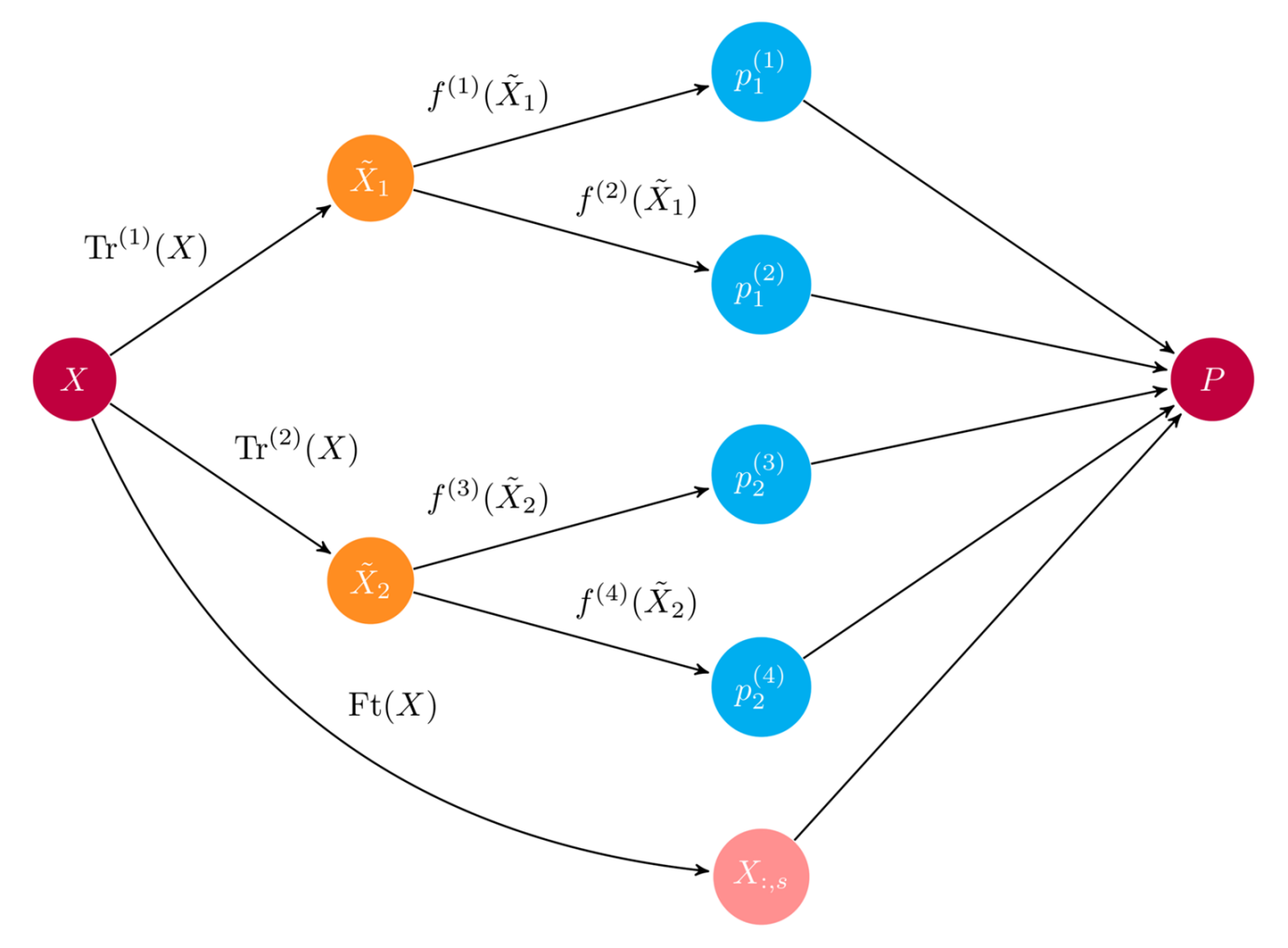
Ensemble Techniques in machine learning

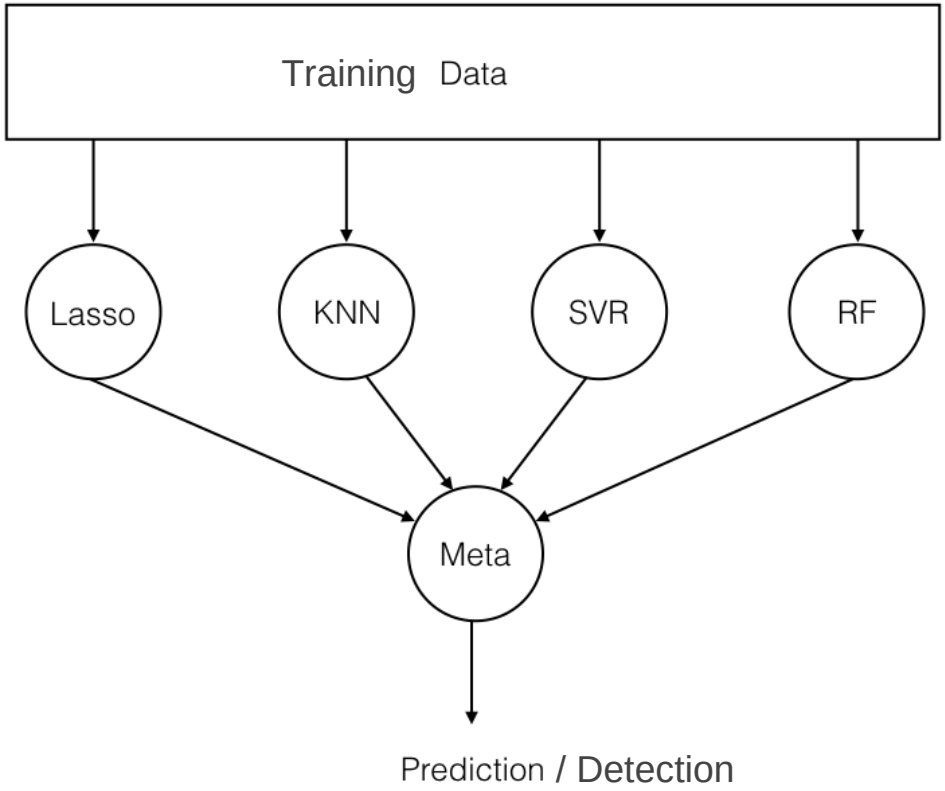
Ensemble Techniques

- Bagging Algorithm
- Boosting Algorithm
- Stacking Algorithm

Ensemble Learning

- Ensemble learning techniques attempt to make the performance of the predictive models better by improving their accuracy.
- Ensemble Learning is a process using which multiple machine learning models (such as classifiers) are strategically constructed to solve a particular problem.
- An ensemble is the art of combining a diverse set of learners (individual models) together to improvise on the stability and predictive power of the model.


X
$\mathrm{Tr}^{(1)}(X)$
$\tilde{X}_1$
$f^{(1)}(\tilde{X}_1)$
$f^{(2)}(\tilde{X}_1)$
$p_1^{(1)}$
$p_1^{(2)}$
$\mathrm{Tr}^{(2)}(X)$
$\tilde{X}_2$
$f^{(3)}(\tilde{X}_2)$
$f^{(4)}(\tilde{X}_2)$
$p_2^{(3)}$
$p_2^{(4)}$
$\mathrm{Ft}(X)$
$X_{:,s}$
P


Training Data
Lasso
KNN
SVR
RF
Meta
Prediction / Detection

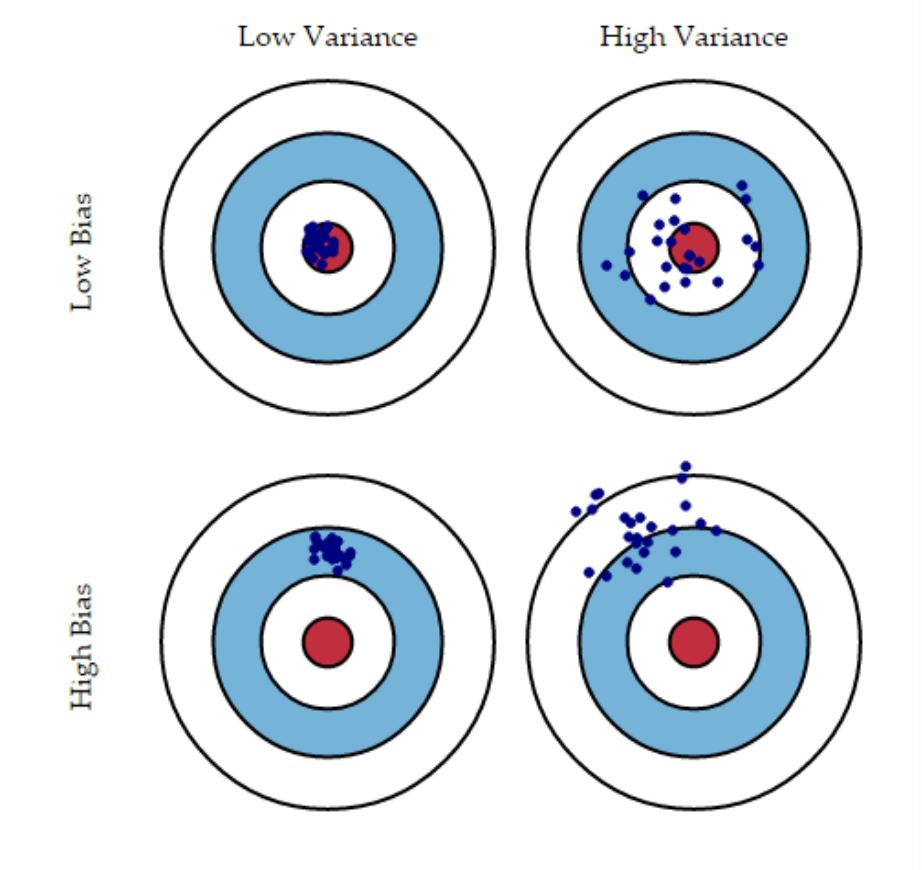
Model error and reducing this error with ensembles :

- The error emerging from any machine model can be broken down into three components mathematically. Following are these component:

Bias + Variance + Irreducible error

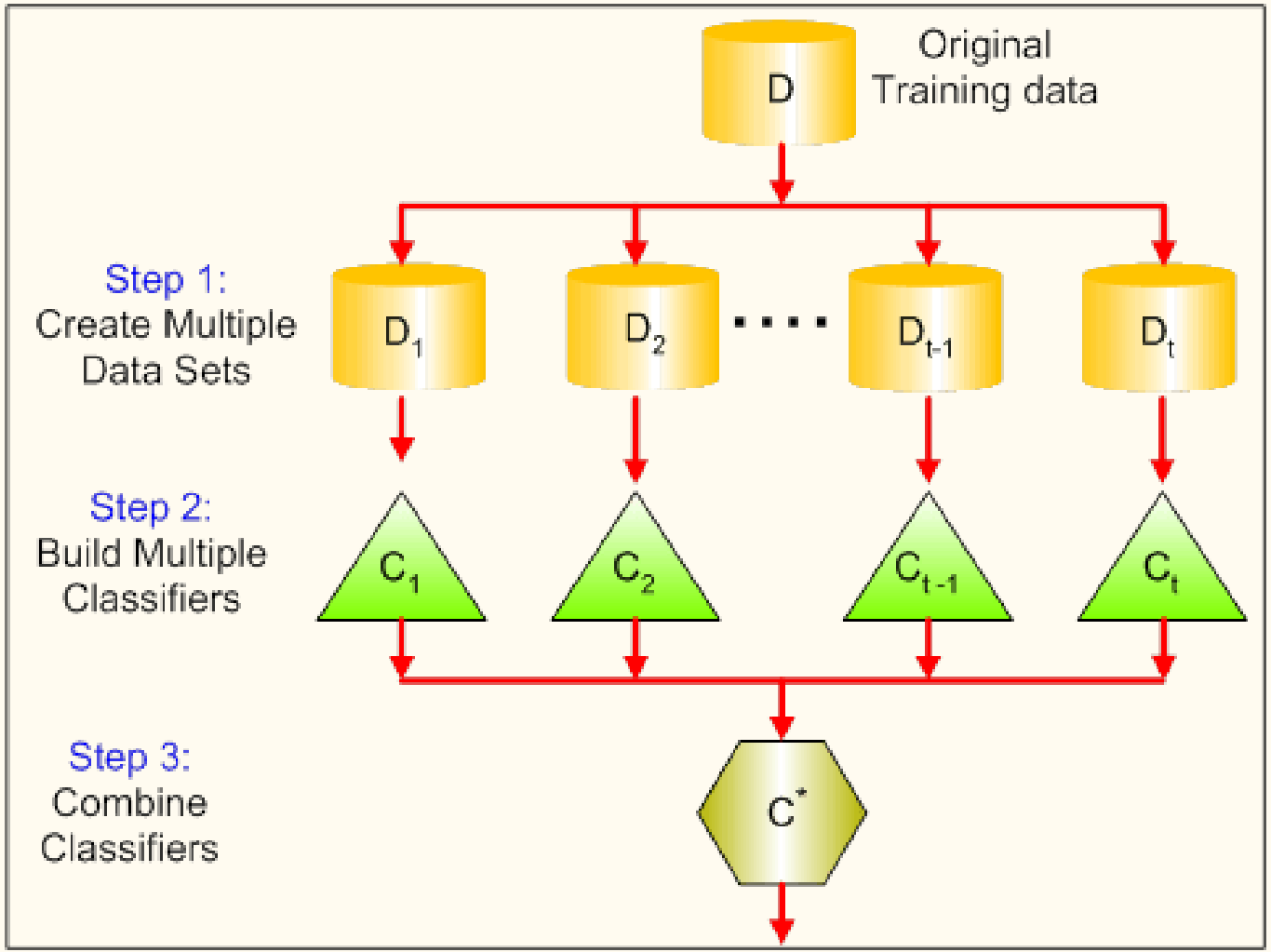
- **Bias error** is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have an under-performing model which keeps on missing essential trends.

- **Variance** on the other side quantifies how are the prediction made on the same observation different from each other.
- A high variance model will over-fit on your training population and perform poorly on any observation beyond training.
- Following diagram will give you more clarity (Assume that red spot is the real value, and blue dots are predictions):



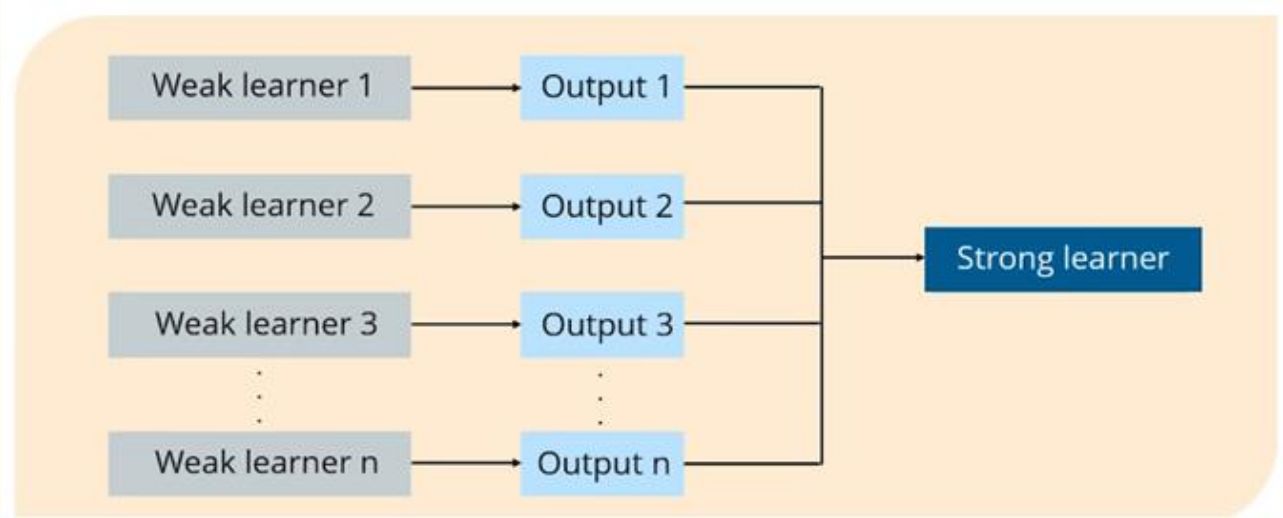
Bagging Algorithm

- Bagging is a form of ***parallel learning*** technique.
- Bagging is one of the Ensemble construction techniques which is also known as ***Bootstrap Aggregation***.
- Bootstrap establishes the foundation of Bagging technique.
- Bootstrap is a sampling technique in which we select “n” observations out of a population of “n” observations. But the selection is entirely random, i.e., each observation can be chosen from the original population so that each observation is equally likely to be selected in each iteration of the bootstrapping process.
- After the bootstrapped samples are formed, separate models are trained with the bootstrapped samples.
- The final output prediction is combined across the projections of all the sub-models.


D
Original
Training data
Step 1:
Create Multiple
Data Sets
D1
D2
Dt-1
Dt
Step 2:
Build Multiple
Classifiers
C1
C2
Ct-1
Ct
Step 3:
Combine
Classifiers
C*

Boosting Algorithm

- Boosting is a form of ***sequential learning*** technique.
- Boosting is the process that use a set of Machine learning Algorithm to combine weak learner to form strong learners in order to increase the accuracy of the model.



Algorithm : Steps

Step 1 : The base algorithm reads the data and assign equal weight to each sample observation .

Step 2 : False prediction are assigned to the next base learner with a higher weightage on these incorrect predictions.

Step 3 : Repeat step 2 until the algorithm can correctly classify the output.

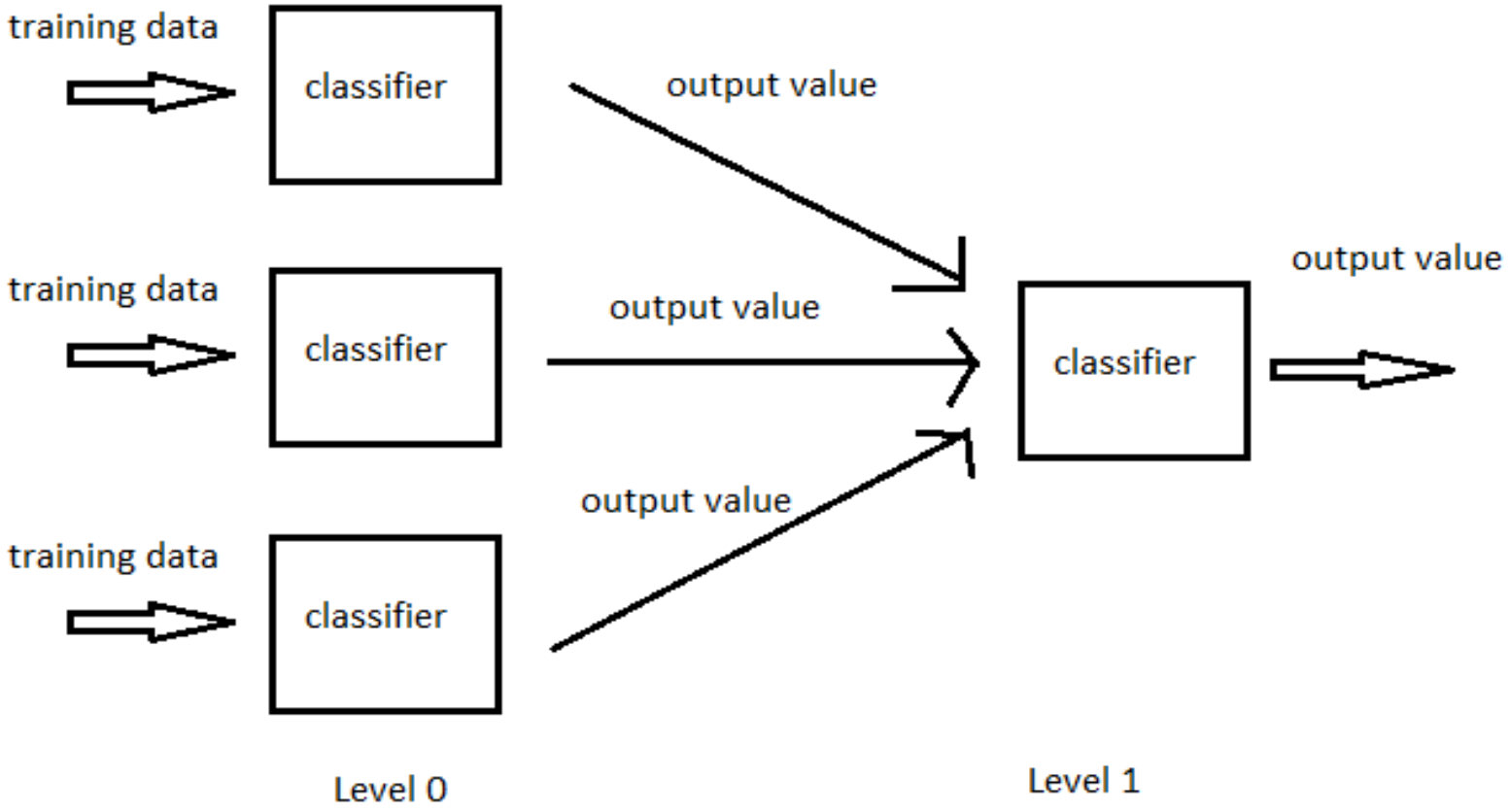
Stacking Algorithm

- Stacking is ***Meta Modeling Technique*** or ***split learning technique*.**
- In Stacking there are two types of learners called Base Learners and a Meta Learner.Base Learners and Meta Learners are the normal machine learning algorithms like Random Forests, SVM, Perceptron etc.
- Base Learners try to fit the normal data sets where as Meta learner fit on the predictions of the base Learner.

Stacking Technique involves the following Steps:-

1. Split the training data into 2 disjoint sets
2. Train several Base Learners on the first part
3. Test the Base Learners on the second part and make predictions
4. Using the predictions from (3) as inputs,the correct responses from the output,train the higher level learner or meta level Learner

Concept Diagram of Stacking



References

- https://machinelearningmastery.com/naive-bayes-classifierscratch-python/
- www.analyticsvidhya.com
- En.wikipedia.org. Support vector machine. https://en.wikipedia.org/wiki/Support_vector_machine
- En.wikipedia.org. Random Forest. https://en.wikipedia.org/wiki/Random_forest
- Stat.berkeley.edu. Random forests - classification description. https://www.stat.berkeley.edu/~breiman/RandomForests/ cc_ home.htm.
- http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html
- http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- http://scikit-learn.org/stable/modules/svm.html

- Lecture Notes (Naïve Bayes) of CS 5100: Fundamentals of Artificial Intelligence, CCIS, Northeastern University by Prof. David Smith.

- Datasets:
  https://www.kaggle.com/c/fake-news/data

- For Bagging and Boosting Algorithm
  https://www.datacamp.com/community/tutorials/ensemble-learning-python
  https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e

- For Stacking Algorithm
  https://medium.com/@gurucharan_33981/stacking-a-super-learning-technique-dbed06b1156d

Reference Research Paper

1. Fake News Detection by Akshay Jain and Amey Kasbe

2. Weakly Supervised Learning for Fake News Detection on Twitter by. Stefan Helmstetter , Heiko Paulheim.

Thank

You !!!