MY RESEARCH ON ---ML, NLP, PYTHON

Hi, this is Sandeep, working as Software Engineer involved in Python, NLP, Machine Learning, Deep Leaning with TensorFlow 2.0 etc.…

In this article I am exploring into Basic Machine Learning and NLP techniques. Let's starts…

Basic question is ***What is Machine Learning?***

"Machine Learning is a method of Data Analysis that automates analytical model building. "

Using algorithm that iteratively learn from data, machine learning allows computers to find the hidden insights without Explicitly program where to look.

How Algorithm learns?

Learning Algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithms learn by comparing its predicted outputs with correct outputs to find errors.

Depends on the results we can modifies the model.

In Machine Learning we have three categories

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning.

This article will walk you through Supervised Learning models.

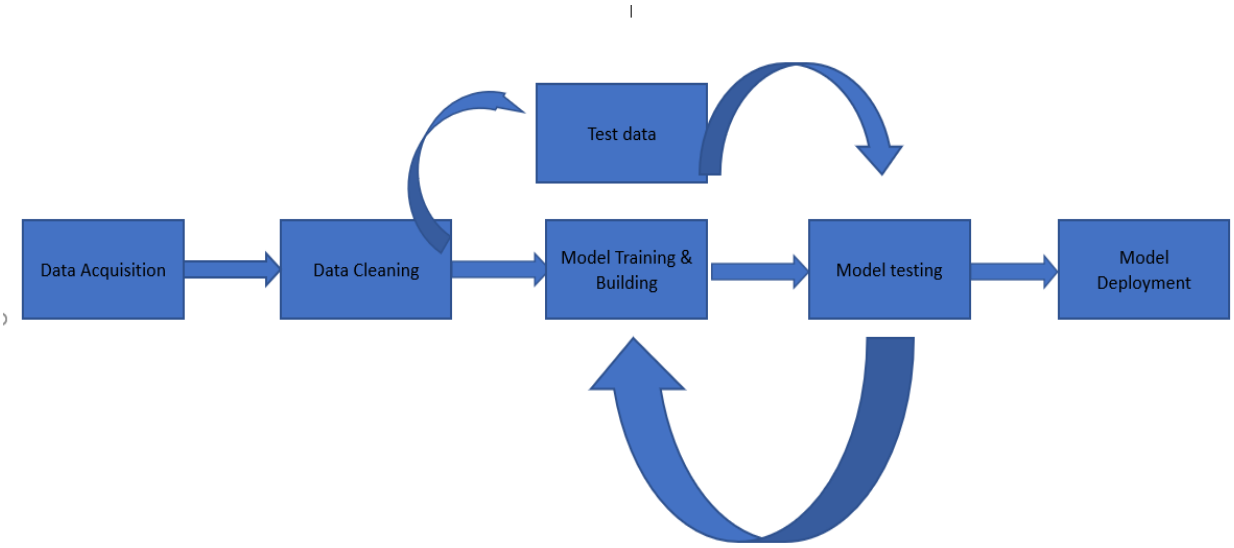
Supervised Learning:

Supervised learning is commonly used in applications where historical data predicts likely future events.

In simple terms, the data which has ***Labeled***

If we go through Machine Learning life cycle the below picture represents.

My Site: https://sandeepkumar16nlp.github.io/sandeepmyportfolion.github.io/



Data Acquisition: Its simply gathering the data. The source of data should be anything, we can collect from customer, company, different Api's whatever.

Data Cleaning (Data Preprocessing): It's the one of the important task in machine learning algorithms building. There are lot of way we can define this, here Data cleaning is first understanding the data, is this data labeled or not. Is that data is numerical data or text data, what are the features we have, features means columns. There are different techniques to identify relation between two or more columns each other. Finally, we understand the what are the features are important and not, removal duplicates, remove null values, this process depends on requirements.

Example to get relation between columns in Data... >> Data.corr()

Model Training & Building:

Training the model with Trained Data. Here we are passing the data through preexisting models from Sci-kit learn.

Model Testing: Where we can test the model is model working fine or not if any changes required to model build.

Test data: with Test data testing model.

Deployment model: After all modifications over we can deploy model in production.

Okie, we gone through all the term which we have in diagram representations. Lets question our self what is Training data and Test data.

Example:

My Site: https://sandeepkumar16nlp.github.io/sandeepmyportfolion.github.io/

First how we are testing the data?

If we have data with 1000 points (rows), and then if we pass this data to model as an input then how we can test.

For this, we need to split the data into two parts i.e., Training Data, Test data. In general, we split this in the ration of 70:30 (Train and Test)

So, from 1000 points we are giving 700 points to model with question and answers to learn what the hidden insights are there, and 300 points to test with only questions. Remember we have answers for all the 1000 points so when we test model with 300points to get to know results are checking with already existing answers.

So, question is how can we measure, **how can I trust this model?**

In Machine Learning model, there are different techniques we have, so we will investigate basic and important.

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall
5. Error rate
6. F1 Score:

Confusion Matrix:

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

The above diagram represents confusion matrix.

From the above we have total 165 data points are there in test data, when we give input to the model, in 165 data points result labels are YES and NO.

From that Actually YES are 105 points and Actually NO is 60

Now let's see how model predicted,

YES labels: Correctly Predicted 100 wrong prediction 5

NO labels: Correctly Predicted 50 wrong Prediction 10

True Positive

True Negative

False Positive

False Negative.

Type1 error: False Positive

Type2 error: False Negative.

Accuracy:

Accuracy in the classification problems is

"Number of Correct Predictions made by the model divided by the Total number of Predictions"

Accuracy Model is good is good for Balanced data set, its not suitable for imbalanced data sets.

(TP + TN) / Total points >>

Recall:

"Number of True Positives divided by the number of True Positive plus the Number of False Negative"

TP / (TP + FN)

Precision:

"Number of True Positive divided by Number of True Positive plus the number of False Positive"

TP / (TP + FP)

Error rate:

(FP + FN) / # no points