Introduction to MACHINE LEARNING

Linear Regression :- ~~Predictions~~ Predictions using the simple yet powerful L.R. algorithm.

Logistic Regression :- Predicting classes for events using Probability of odds.

D.T. & R.F. :- Question based approach to Prediction.

PCA :- Simplify data to lower dimensions.

KNN :- Classifies a data point based on its nearest neighbours.

Naive Bayes :- Classifies data based on Conditional Probability.

K-means :- Cluster a set of objects based on measure of similarity.

SVM :- Predict or classify using support vectors.

Time Series Analysis :- Analysing time series and forecasting future occurences.

Linear Regression :- ①

Logistic Regression :- used when there is a linear relationship bw variables

DT & RF :- There is no linear relation & Completly non-linear.

SVM :- Good ; when No. of features are higher than the no. of rows.

KNN :- Good ; when everything is mathematical, because it looks at Euclidian distance / manhatten distance, need to normalise the data really well.

PCA :- Good, when the features have not a lot of interdependence bw them & dealing with uncorrelated features.

LDA :- Advanced version of P.C.A used for classification.

Recomendation Engine

Apriori algo :- How to come up with recomendation based on time slices. Particular transaction with particular time what item the customer will buy

Anamoly detection :- detect what is normal & identify what is abnormal. We use T-digest, windowing algo.

How to win kaggle Competition → Ensemble learning

Boosting → Adaboost
Gradient Boosting
XG Boost

Optimization :- Linear Problem - use minimize SSE etc

Non-linear problem - we use gradient descent

which direction we need to go

Batch GD
takes all the samples & feedbacks @ once, Processes & adjust the direction

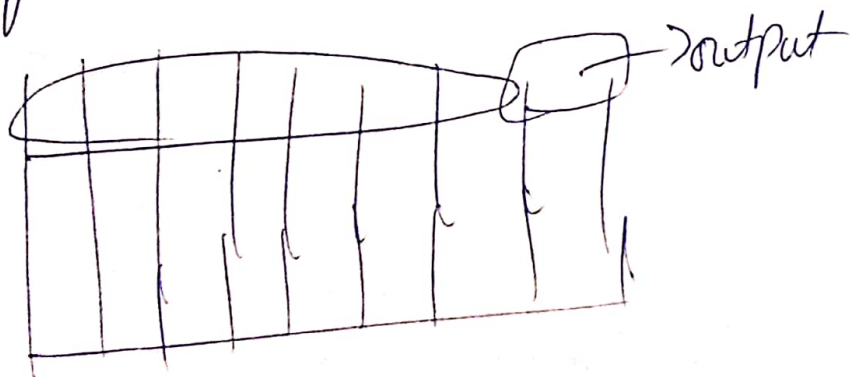
Mini Batch GD

Better Both world

S.G.D.
(one sample @ a time)

Top ML Terminologies

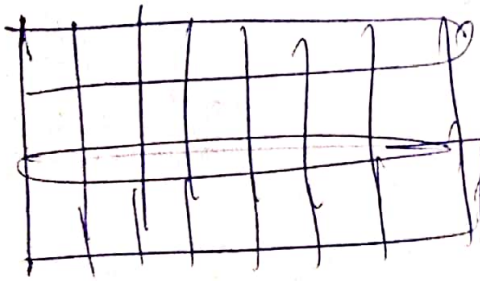
Term - 1 = Features : The No of distinct traits that can be used to describe each item in a quantitative manner.



Term-2 = A sample is an item to process. For instance, it could be a document, picture, sound, or a raw database.

Term 3: Feature Vector

A feature vector is an n-dimensional vector of numerical features that represent some object

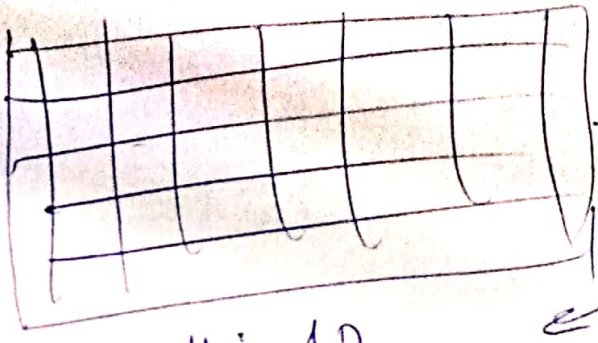


→ feature vector

Term 4 :- Feature Extraction.

↳ Preperation of feature vector.

↳ Transforms the data from high dimensional space to a space of fewer dimensions.



→ Transforming 3D to 1D

Sweet Juicy Ripe

all in 1D

Term 5 : Training Set

Set of data to discover potentially predictive relationships.

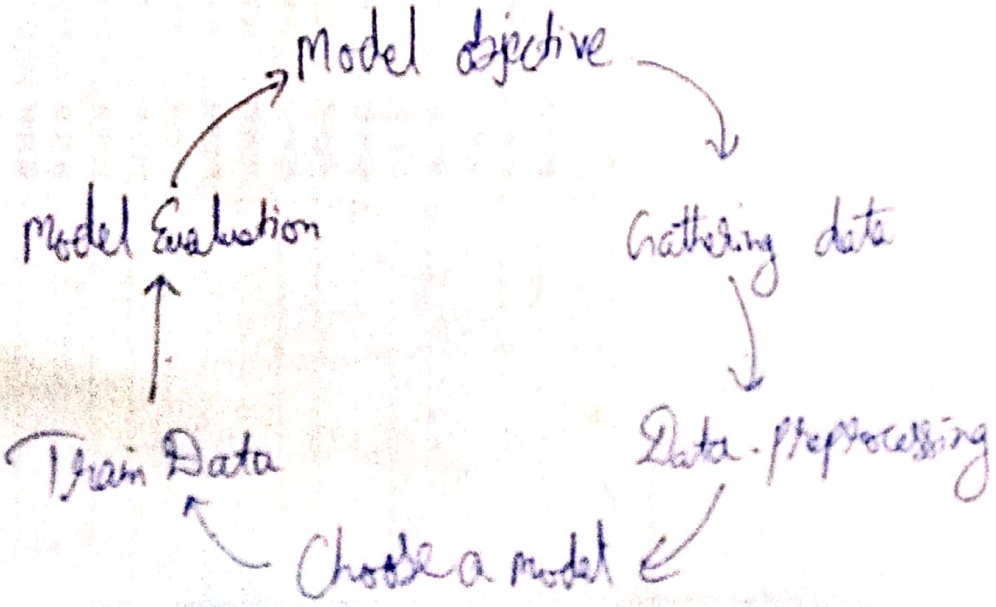
Term 6 : Test Set

Set of data to validate the predictive relationship

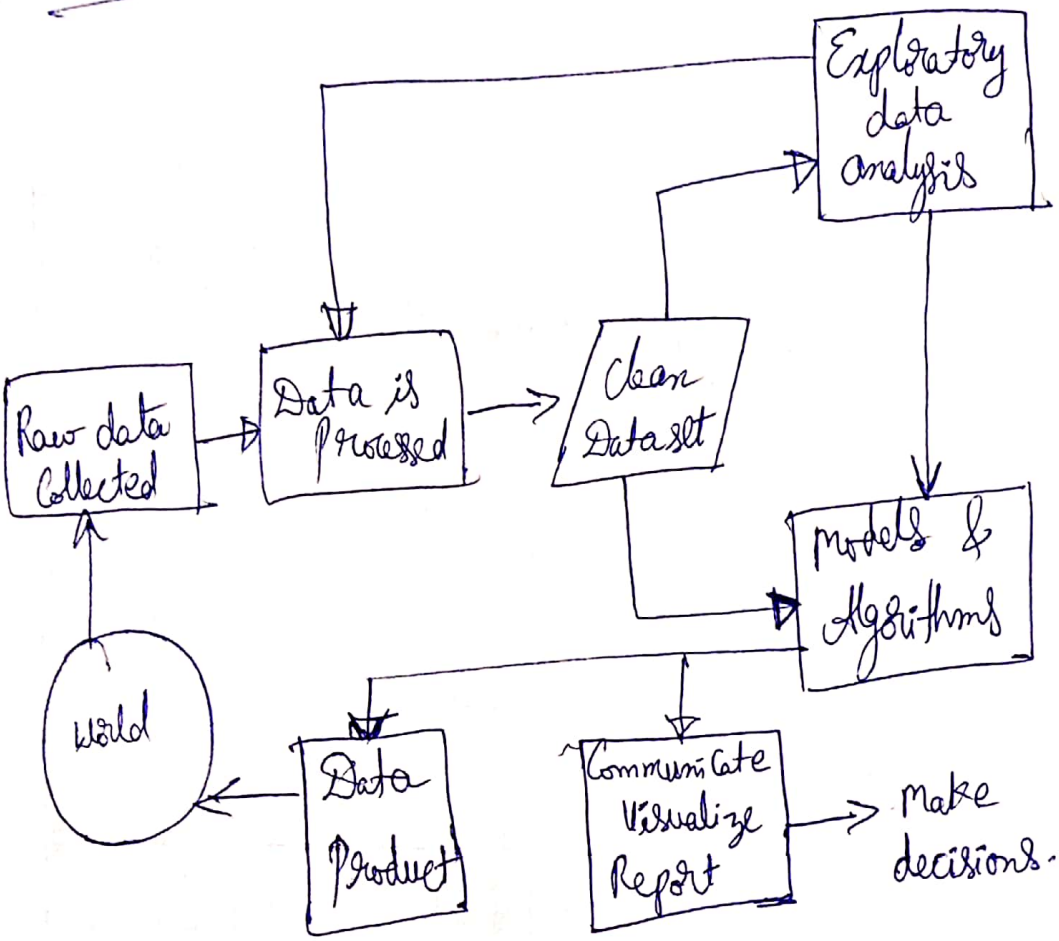
Term 7 : Scoring

Evaluating the performance of a machine learning based on any statistical metric.

ML Workflow



Putting it all together



Step 1: Collect Raw Data

To solve a given problem, as a data scientist you need data.

Step 2: Store Raw Data

Raw data means data that has not been changed since acquisition. This raw data is stored in your storage systems.

Step 3: Data Pre-processing.

As a Data Scientist, a lot of your time will go in Data Pre-processing. Also known as Data cleaning.

This step includes

- Removing outliers
- Replacing missing data
- Malicious Data.
- Erroneous Data
- Irrelevant Data
- Inconsistant Data
- Formatting.

Once Data is cleaned, it needs to be processed to make it ready for use.

- This stage includes
  - → Sorting
  - → Summerization.
  - → Aggregation
  - → Validation
  - → Classification

→ Data Preprocessing (Data cleaning) is at times considered to be part of Data Processing

Step 4: Exploratory Data Analysis

What are the **Key Concepts** of about EDA?

2 types of Data Analysis

- Confirmatory data analysis
- Exploratory data analysis

4 objectives of EDA

- Discover patterns
- Spot Anomalies
- Frame Hypothesis
- Check Assumptions

2 methods for Exploration

- Univariate analysis
- Bivariate analysis

Stuffs done during EDA:

- Trends
- Distributions
- Mean
- Median
- outliers
- Spread measurement (SD)
- Correlations
- Hypothesis testing
- visual exploration

Objectives of EDA

↳ Discover patterns
↳ Spot anomalies
↳ Frame hypotheses
↳ Check assumptions

Step 5: Modelling & Algorithms

① Split data into train and test.

② Build multiple models to solve the Business problem.

③ Evaluate models using test data and Error metrics.

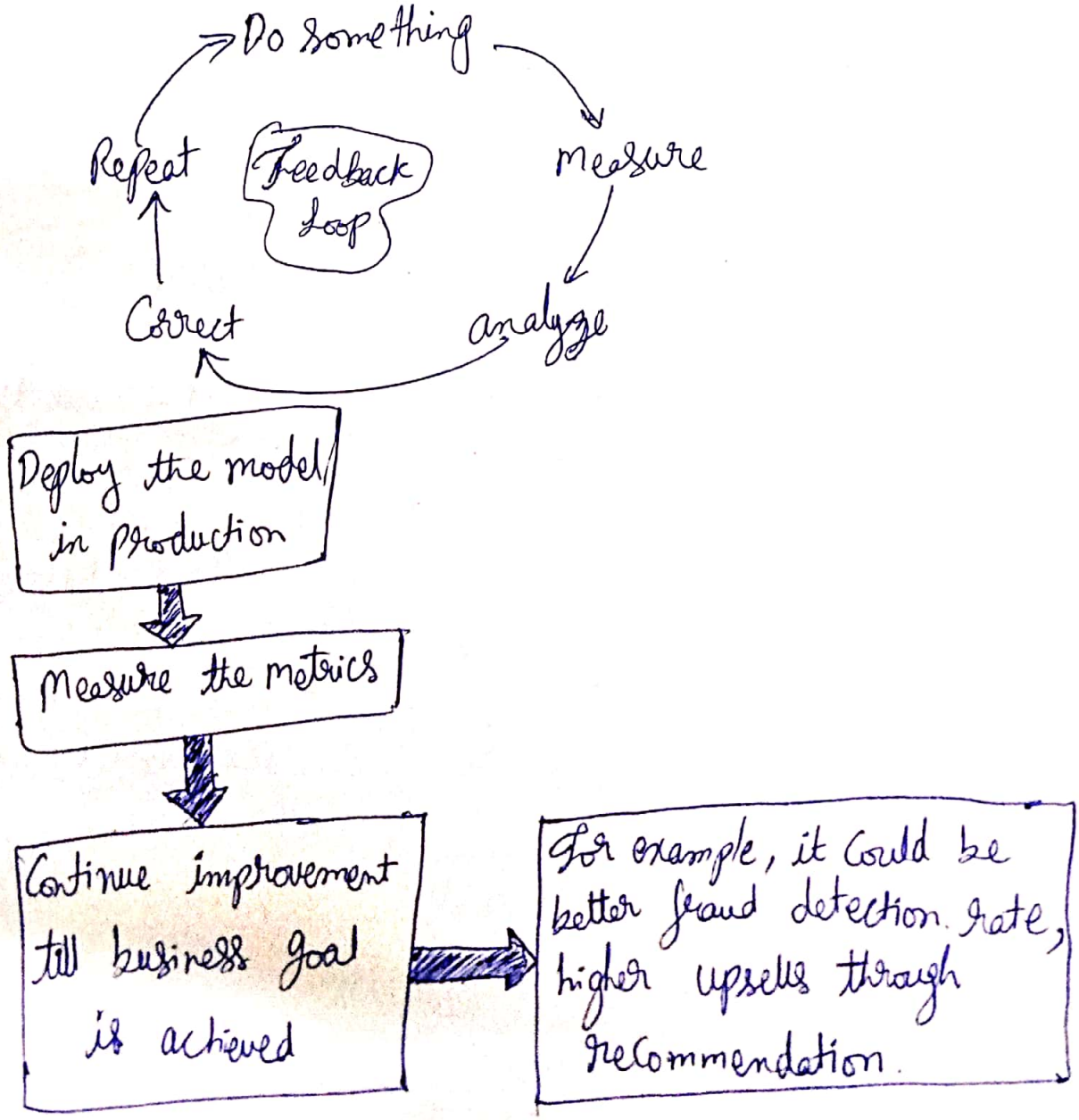
④ Choose "Best" model by Comparing models to see which one comes closest to answering business objective.

Step 6: Communicate visualize & Report

↳ Brainstorm with Management & Showcase the benefits, the analysis, and models bring to the plate.

↳ Seek management's consideration for deploying the solution to real world to help make the business more optimized & beneficial.

Step 7: Take Action & deploy the findings in real world



Types of Machine learning

Supervised Learning enables Computers to learn from labeled data without being explicitly Programmed.

UnSupervised Learning draws inferences from data without labeled responses.

Supervised	Unsupervised
Direct feed back	No feedback
Labeled data	No Labeled data
Predictions as output	Find hidden structure in data

Reinforcement Learning :- Learn by interaction with environment.