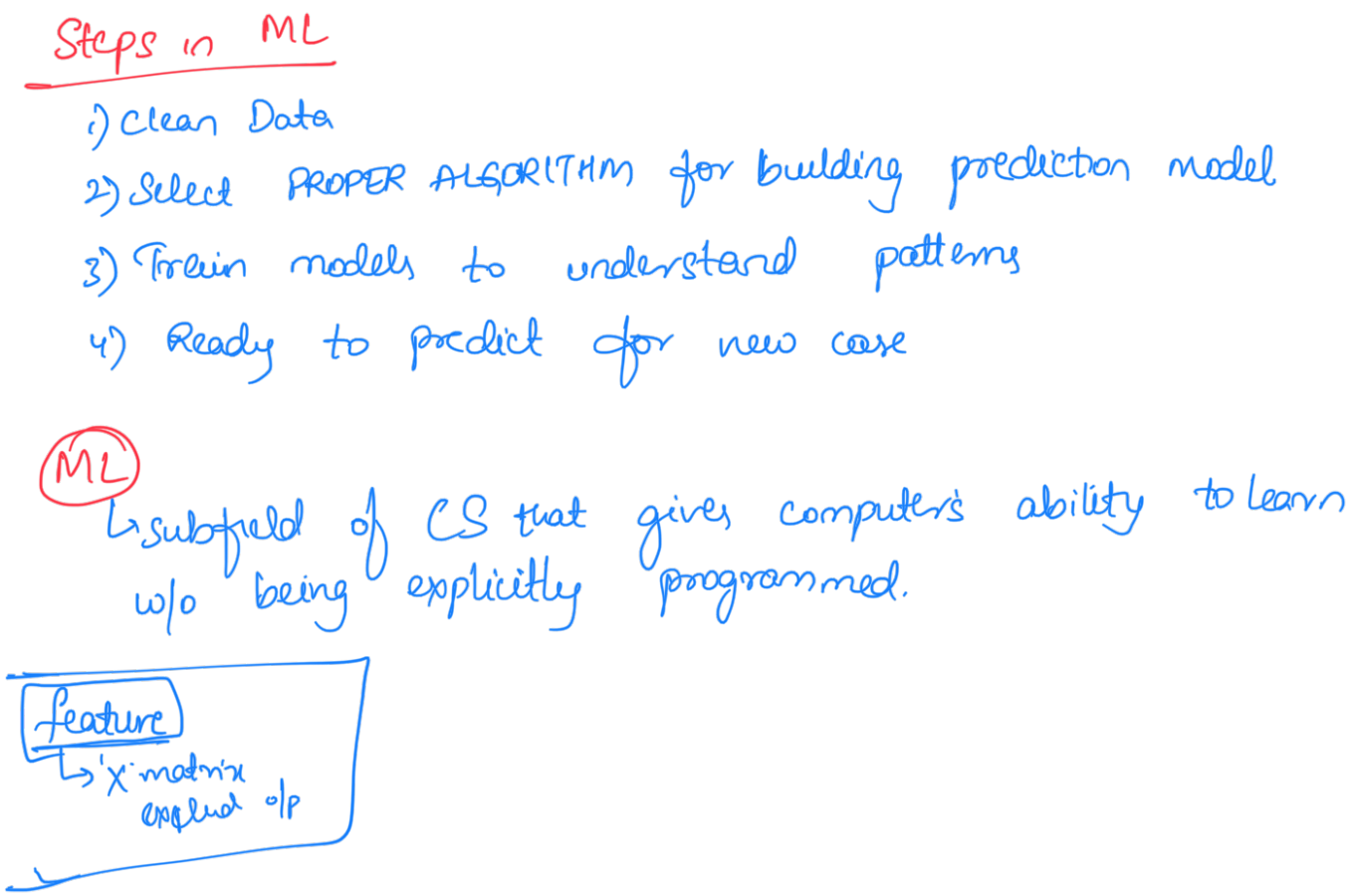
ML Theory FSN



MAJOR MACHINE LEARNING TECHNIQUES

- ***Supervised ML approach***
  - We teach the model with labelled / historical data, then with that knowledge , it can predict unknown or future instances of unlabelled data
  - **REGRESSION / ESTIMATION = process of Predicting a continuous value**
    - y = θx + c
      - θ can estimated by Ordinary Least squares and optimisation algorithm
    - **Linear Regression** ✅👨🏻‍💻
      - (CO2 Emission Vs Engine Size)
    - **Multiple Linear Regression**
      - (**CO2 Emission Vs Engine Size , No of Cylinders**)
    - **Non Linear Regression**
      - (Linear or non linear based on relation between dependent and independent variable)
      - **Polynomial Regression**
      - ***Ex : Data points corresponding to Chinas GDP from 1960 - 2014***
  - **CLASSIFICATION = process of predicting/ categorising discrete class labels or categories , target attribute is categorical variable**

- **(Ex : Predicting cancer cell benign or malignant , customer churn : whether customer switches to other service provider ,)**
  - **Classification Accuracy**
    - **Jaccard Index =** (like %. Predicted correct / total reduction)
    - **F1 score = (2 * precision * recall) / (precision + recall) (Ideally 1)**
      - **Precision = TP / (TP +** FP **)**
      - **Recall = TP / (TP +** FN **)**
    - **Log loss = performance of a classifier where the predicted output is probability value between 0 and 1**
  - **KNN (K Nearest Neighbours)** ✅✅ 👨🏻‍💻
    - **Ex : Customer category Service (Basic , E , Plus & Total) based on age , income & education**
    - **can be used for Regression as well as for Classification** but mostly used for Classification.
    - method for classifying cases based on their similarity to other cases
    - Cases that are near to each other are said to be neighbours
    - Based on similar cases with same class labels are near to each other
    - **Steps :**
      1. Pick a value of k
      2. Calculate distance of unknown data case from all cases
      3. Select k observations in training data that are nearest to unknown data case
      4. Predict the response of the unknown data point using the most popular response value from the k nearest neighbours
      5. Find best value of k
  - **Decision Trees** ✅✅ 👨🏻‍💻
    - Intuition is to map out all possible decision paths in form of a tree
    - It is built by splitting the training data set into distinct nodes
    - Ex : Drug ABC Vs Age , Sex , BP , Cholestrol
    - Steps :
      1. Chose An attribute from dataset
         - Which attribute is best
           - Cholesterol : High 11100 Normal 11000 ❌
           - Sex : Male : 11110 Female 10000 ✅
             - More - predictive , purity
             - Less - Impurity , Entropy
           - Entropy : measure of randomness / uncertainty ,
             - less entropy , less uniform , More Pure Node
           - Information gain :
             - tree with high information gain after splitting

        - Information gain - info that can increase level of certainity after splitting
        - = entropy before split - weighted (avg) entropy after split
    2. Calculate significance of attribute in splitting of data
    3. Split data based on the value of best attribute
    4. Go to step 1
- **Random Forest ✅**
    - Used in both classification and regression , but better in classification
    - Ex : above drug dataset
    - It builds decision trees on different samples and takes
        - majority vote for classification & average in case of regression
    - It uses Bagging ensemble technique = (combining multiple models) ,
    - thus collection of models is used to predict than an individual model
    - **Bagging / Bootstrap aggregation**
        - It is technique used to reduce variance of our predictions by combining the result of multiple classifier models on different sub-samples of same data set
        - It's parallely)
        - (Boosting , Not used in RF , it's sequentially)
        - Bootstrapping is statistical method to create sample without leaving the properties of actual data set , individual samples of data are called bootstrap samples
    - Steps :
        1. Create n number of trees == (forest)
        2. Take test features , use randomly created decision tree to predict outcome & store predicted outcome (target)
        3. Calculate votes for each predicted target
        4. Consider high voted predicted target as final prediction from RF algorithm
- **Logistic Regression**
    - Used for classification
    - It is classification algorithm for classification variables
    - Similar to linear regression , but takes categorical instead of numeric
    - It measures the probability of a case belong to specific class
    - Ex : Customer Churn : Yes or No based on Tenure , Age , income , type of service , employment
    - **Linear Vs Logistic**
        - Linear not good for binary classification
        - Logistic employs **Sigmoid function** (Heart of Log Reg **)**
        - Predicting churn (binary classification)using Linear Regression

has some issues , Ex if x < 0.5 No , else Yes , it's like digital 0 or 1
        - So sigmoid function is used in Logistic regression
        - = 1/(1+e power X) , it's S like smooth curve , correctly differentiate and gives exact probability of case in range of 0-1
        - Output : p (Y=1| x) === ex like 0.8
        - Ex p(Churn =1 | Income, Age) = 0.8
    - Support Vector Machines SVM ✅👨🏻‍💻
        - supervised algorithm that classifies cases by finding a separator
            - 1. Mapping data into high dimensional feature space
            - 2. Finding a seperator
        - Kernelling : mapping data into higher dimensional space , in way Data Not linearly separable to Linearly separable using linear / polynomial / rbf / sigmoid functions
    - Naive Bayesian
- *Un-Supervised ML approach*
    - **Where model works on its own to discover information**
    - **CLUSTERING = Is grouping of data points / objects that are somehow similar by Find structure of data / summarisation (Customer segmentation in banking)**
        - K means ✅👨🏻‍💻
            - Randomly placing k centroids , one for each cluster ,
            - Calculate the distance of each point from each point from each centroid ,
            - Assign each data point to closest centroid , creating a cluster ,
            - Recalculate position of 4 centroids
            - Repeat steps 2-4 , until, the centroids no longer move
            - Choosing k , elbow point
    - **DIMENSION REDUCTION = Reducing size of data**
    - **DENSITY ESTIMATION**
    - **MARKET BASKET ANALYSIS**
    - **ASSOCIATIONS = Associating frequent Co occurring items / events (Groceries items bought together by particular customer)**
    - **SEQUENCE MINING = Predicting next events (clickstream in websites)**
    - **RECOMMENDATION SYSTEMS = people preferences with others who have similar tastes (YouTube , Amazon)**

- *Semi-Supervised Machine Learning*
    - **ANAMOLY DETECTION = Discover abnormal and unusual cases (credit card fraud detection)**

- *Reinforcement Learning*

Artificial Intelligence

refers to simulation of human intelligence in machines that are programmed to think like humans & mimic their actions.

Deep Learning

is a ML technique that teaches computers to do learn by example like humans.

Ex: Driverless car
- Drive (repeat) learn
- Recognize stop sign
- distinguish pedestrian from lamppost.

Errors

- MAE
- MSE
- RMSE
- RAE
- RSE
- R^2 = 1- RSE

MODEL EVALUATION Approach

- Train and test split on same data set
- Train / Test split
- K fold cross validation : ex : 4 fold = 4 different 25% test split , then 4 diff average

Error of model = difference between data points and trend line generated by the algorithms

To improve model accuracy
Bagging
Boosting
Bootstrapping

Sir , Tips

Yes no SVM
no label , group K means clustering /
KNN for training data , for new data KNN
Decision tree, sequence of decisions to make

Data distribution
Correlation
1 significant , kde plot

Pair plot
Correlation matrix

Vectorise if categorical ,

Correlation matrix

Categories group ,
If dealing with clusters , better to normalise

Score
Predict
Train on significant columns ,
Score vary based on new columns , passing