## Machine learning

RECAP:

→ write a computer prog to learn from examples

$$y = f(x)$$

features   o/p   output   ↙   ↖ feature or representation.
relation.              prediction
                       function
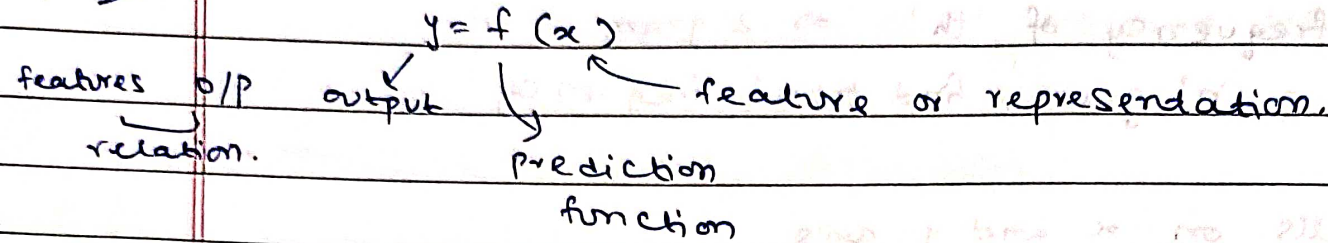
categorization of ML based on learning.

ML
↓
Supervised    unsupervised    Semi-supervised    Re-inforcement
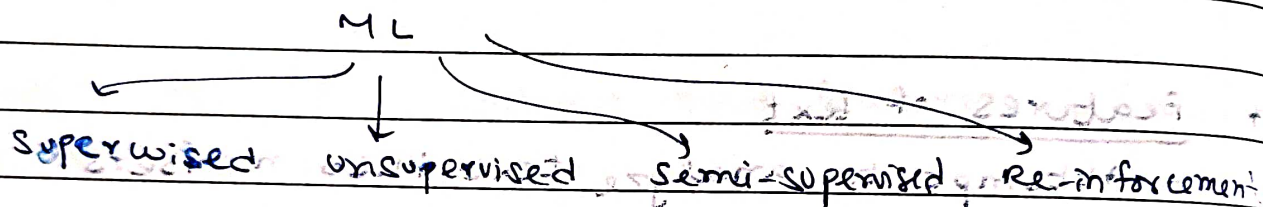
\* ## Supervised learning

(feature vectors, o/p values)   dataset→kaggle.

- ML that is designed to learn by examples.

- It is trained with labelled data
    · feature vectors : $x_{ij}$
    · output values.

- It maps the input to an output based on
  previous input-output pairs
  
  $y = f(x)$
  ↙
  Categorical   Numerical

<u>classification</u> → predicts a discrete value

$y = f(x)$

if $y$ is discrete ⇒ classification dataset

belongs to a particular class; or no

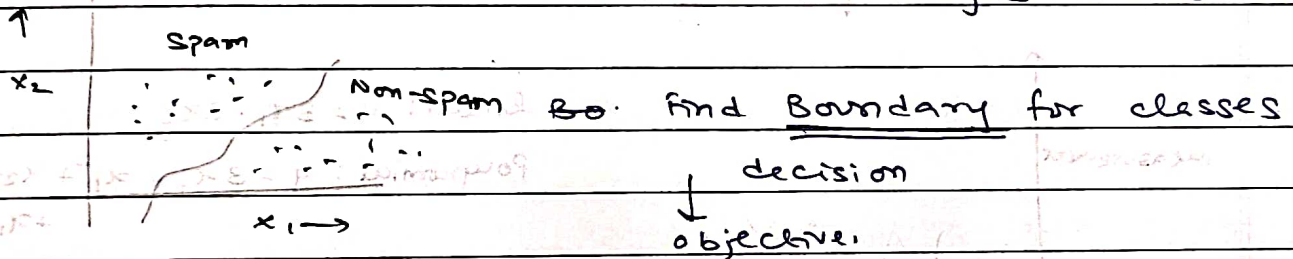- spam filtering
- classify if two docs are original / copied.
- which image class$^n$ → dogs, cats

defining set of categories, assign datapoints.

Classification algos : • Decision Trees
- Support vector Machines
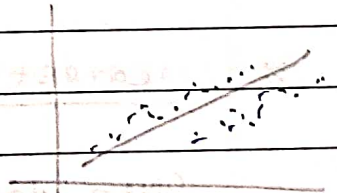- Naive Bayes classifier.



Bo. find <u>Boundary</u> for classes

decision
↓
objective.

<u>Regression</u> → That predicts a continuous value.

if $y$ is cont value for ex: Price of house.

$y = 100, 101, 100.295$ take any cont value.

No Boundary, find best way to describe all points.

Regress$^n$ algos • linear regress$^n$
- Polynomial regress$^n$



find a straight line best fitting most points.

Ex: Predicting sales for a business
price of a house.

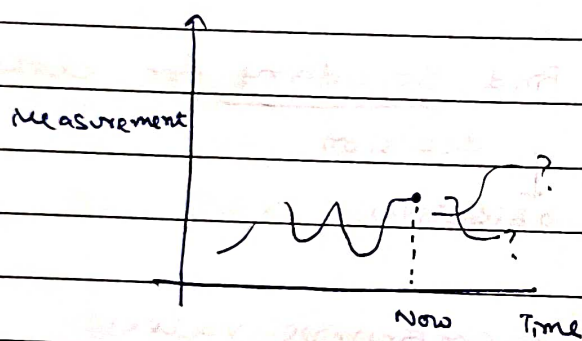\* **Time-Series Model**

    sequential data is given

      => (ordered data)

    Ex: stock Price Prediction

Given a sequence of data like years & Sales predict future samples (t+1)

OBJECTIVE: we want a ML model to understand sequences, not samples.

Popular technique: Recurrent neural networks



linear: $y = 3x_1 + 2x_2$

Polynomial: $y = 3x_1 + x_1^2 + x_2 + x_2^2 + x_1 x_2$

Model a sequence, Predict next

If values of y are 0 or 1 : discrete. classification

If values of y can be b/w 0 & 1: cont

\* **Nearest Neighbour classifier.**

Given dataset: Y values discrete ($\because$ 0 or 1)

      => classification

| $x_1$ | $x_2$ | Y |
|-------|-------|---|
| BMI | Age | Diabetic |

find nearest neighbour => will be same value

If label is there, can we know which class it belongs to

Feature vector       label       distance
        \             |            |
        |             |            |
        .             |            |

datapoint to predict is closer to which feature vector
=> Belongs to that class.

Given a test sample, predict the class.

New!
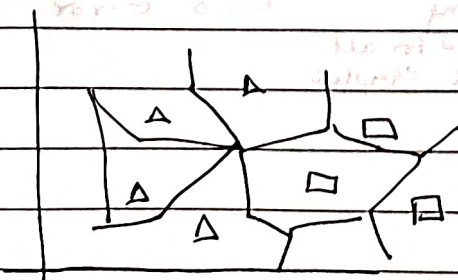→ Voronoi Tessellation : Given any space with a lot of
                                    datapoints

     Most effective way to classify points
     ( skin of Jackfruit)     Partitions.

                          all the region of one cell belong
                          to one category/class.

Discrete data: Categories / label / class

whenever we have two classes, we draw boundaries
to be able to distinguish.



we ignore boundaries
b/w samples of the
same class.

If a red point is surrounded by green points, our model must understand that theee red point is a noise. => we can ignore.

If there are more red points, ML model must re-define boundaries.

Hyper plane → high dim to low dm.

## Classification Algo :

Given :- $(x_i, y_i)$

  - A set of $m$ test samples

Find : - label $(x_{test})$ using similarity measure
  - return $y_{pred}$

find accuracy of the prediction : - Evaluation of a Classifier.

Time Complexity : $O(n)$

## Evaluating a classifier

```
+-----------+        +------------+ -----> Accuracy
| Test Data | -----> | Classifier |
+-----------+        +------------+ -----> Error Rate
                      Predicting
                      y pred for all
                      M test samples.
```