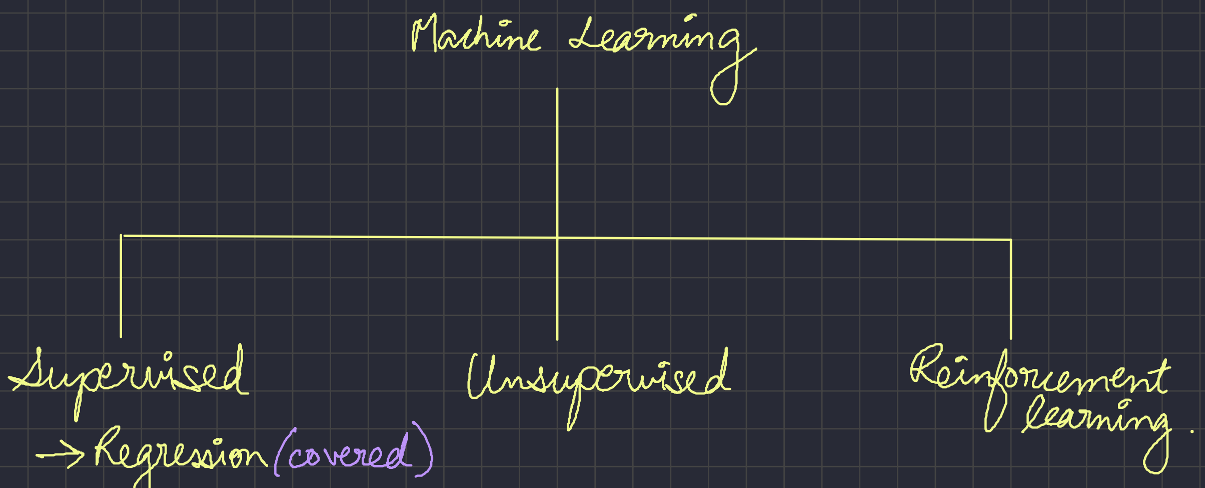
Classifiers and classification.

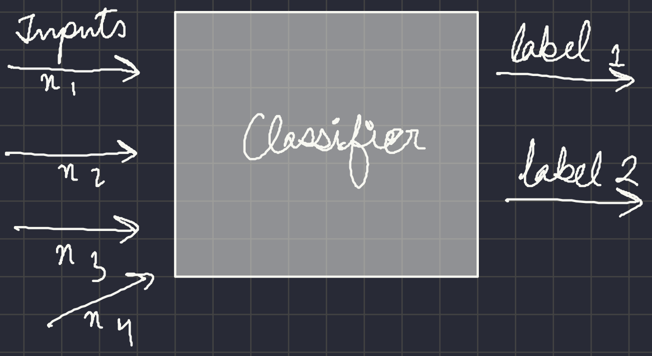
Tree for Machine Learning:



→ Regression (covered)

→ Classification

Q.> What is a Classifier (?)



Output

[set of classes].

labels.

$\{0, 1, 2\}$ ~~1.5~~ ~~2.5~~ ~~-1.0~~

{'tall', 'short'}

email classifier {'spam', 'ham'}

only discrete values allowed.

$$Output = \{x_1, x_2 : label\ 1 \\ x_3, x_4 : label\ 2\}$$

Difference b/w Classifier v/s classification model

Classifier → algorithm itself

methodology/set of rules to classify input data

classification model → Trained using a classifier algorithm.

gives output labels on input data with specific featureset.

Types of classification algorithms

Classifiers we will cover today.
- → Decision Tree
- → Naive Bayes Classifier
- → KNN (K-nearest neighbors)

Later portion
- → SVMs (Support Vector Machines)
- → Artificial Neural Network (ANN).

Example:

Algo. Binary search

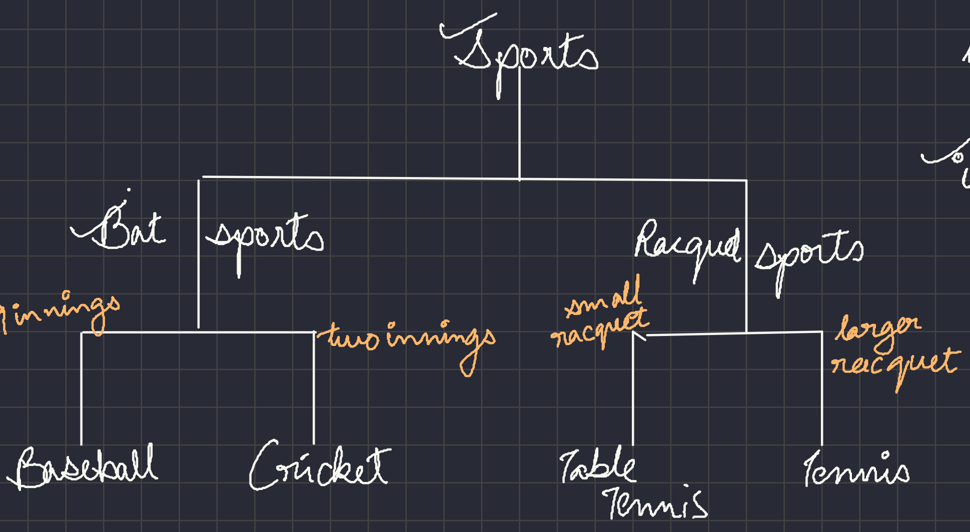
↓

special usecase

↓ becomes model.

Model: binary search for specific DSA problem.

Decision Trees



input

↓

if-then statement

↓

if-then statement

⋮

↓

specific label given as output

output labels = [Baseball, Cricket, Table Tennis, Tennis]

Naive Bayes Classifier

Uses probability [0 - 1]

Text classification (Sentiment analysis)

Comment 1 : great (+ve [0.8] 80% ✓ correct.)

comment 2 : not so great (✓ -ve [0.58] 58%)

c 3: bad, horrible (-ve [0.9] 90% correct.)

inputs

threshold

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

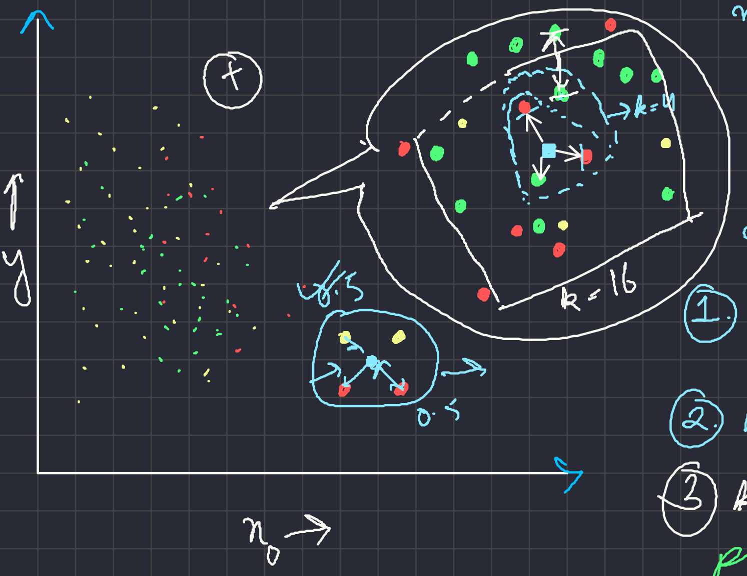
(+ve yes, -ve no, -ve)

Classifier that gives only two labels as output → Binary classifier = [Yes, No]

K-nearest neighbors

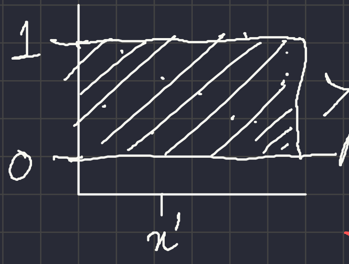
$k > 0$ labels = [● ● ●]

□ new data



Titanic = Regression dataset

new person data → determine probability of survival.



0.75% — [0 – 1]
> 0.5
∴ yes person will survive

k

① k = 3.
pred1 = red

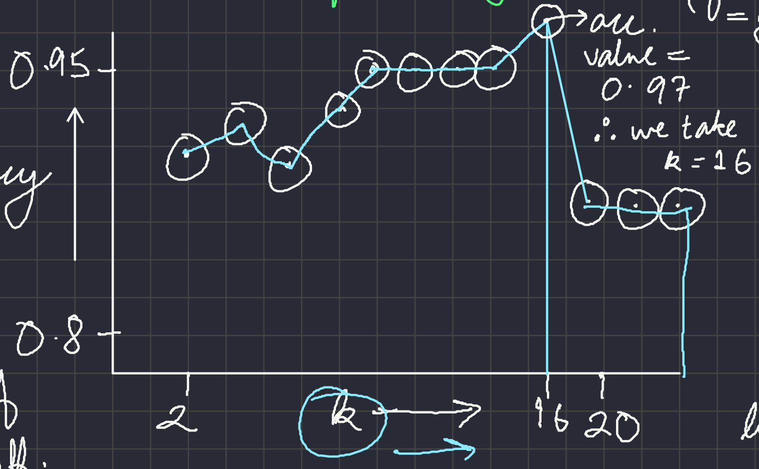
red more than green ∴ label = red

② k = 4.
0.5, 0.5 inaccurate

③ k = 16
pred3 = green

green more ∴ label (for k = 16) = green

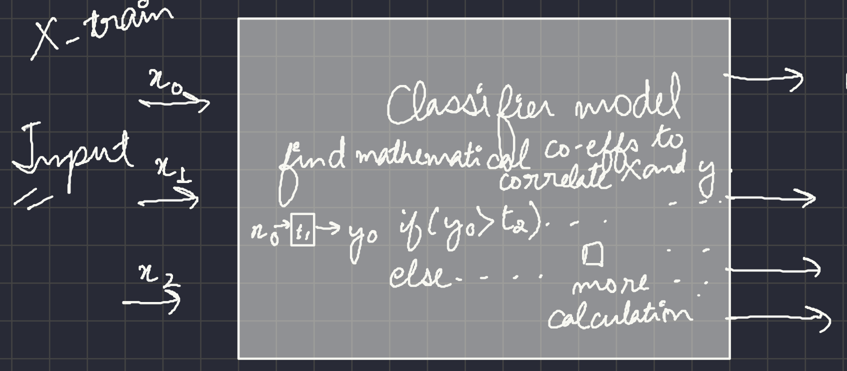
* regression output [0 – 1]
↓
classification output [yes or no based on threshold values for labels].



For KNNs, we measure accuracy over different values of k and map those on a graph. Then select the optimum value of k from the graph itself.

```
for i in range [2-20]:
    ~~ff~~ = KNN(k = i)
    fit( )
    predic( )
    list.append(accuracy( ))
```

∴ When $k = 16$, we get an accuracy of 0.97 or 97% and that k value is suitable for predicting on newer data.


X_train
Input
x_0
x_1
x_2
Classifier model
find mathematical co-effs to correlate X and y
$x_0 \to \boxed{t_1} \to y_0$ if $(y_0 > t_2)$. . .
else . . . .
□ more calculation

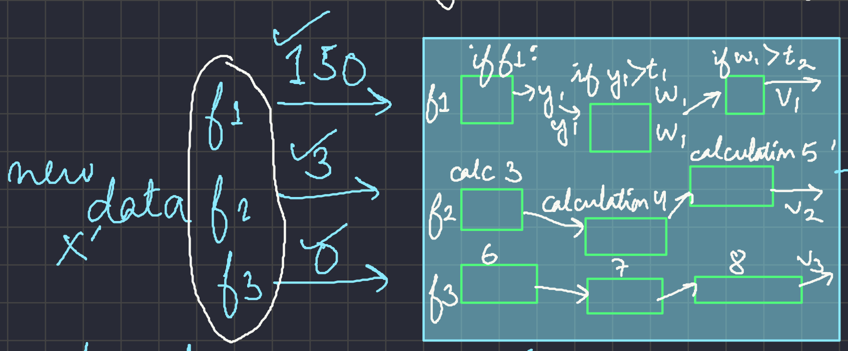
labels y_train given the dataset for associated inputs x_0, x_1, x_2 ⋮ x_n.

During training

X_train is input data for training classifier.

y_train is input data X_train's labels.

** After training: (Main portion of ML) (Make predictions on newer data).



Classification Model

label prediction for newer data X' which has features (f1 . f2 . f3) is {ham} (example).
∴ It is a valid email.

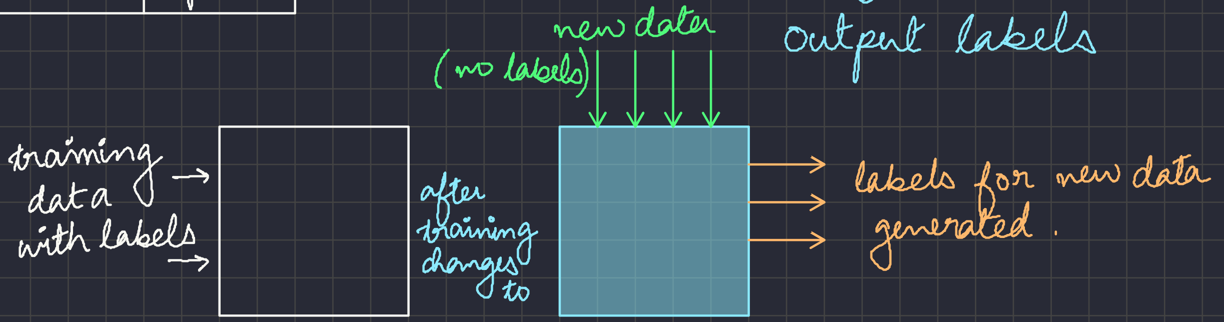
f1: no. of words
f2: no. of links
f3: exclamation point

Sale !! → spam emails may have exclamation marks.

f1	f2	f3	y
800	5	8	spam
80	1	0	ham
200	7	5	spam

already trained ⟹

new data
↓
Classification model
↓
output labels


training data with labels
after training changes to
new data
(no labels)
labels for new data generated.

Classifier

Classification Model trained on some specific data.

Points to remember:

1. Classification is a machine learning technique that predicts the class/category or targets of a given set of data points.
2. Output labels are discrete values.
3. Classifiers are algorithms or set of rules that are used by machines to classify input data.
4. Classification models are the end result of our classifier after it undergoes training on our dataset.