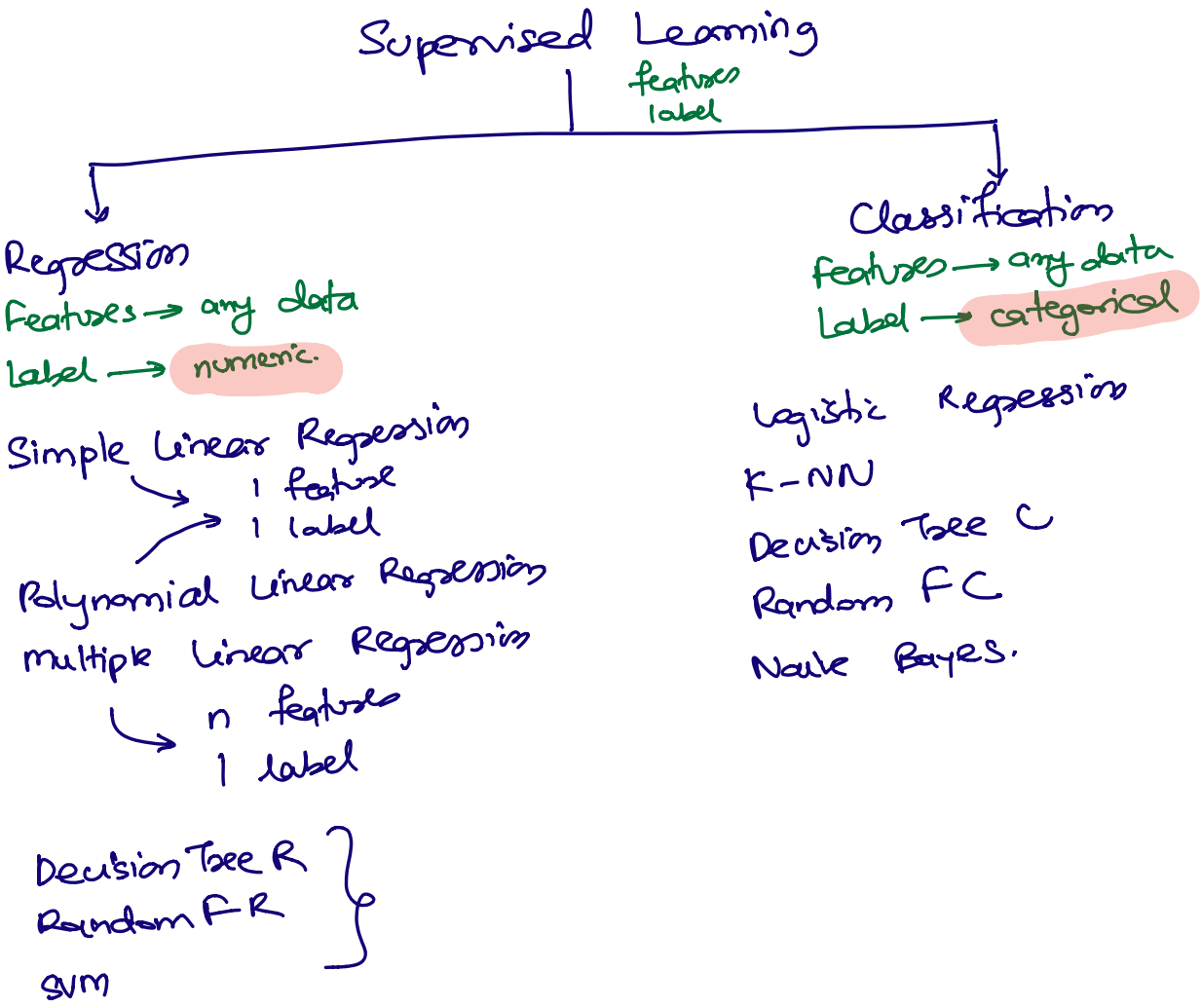
Classification Basics
Approaching a Classification Problem
Logistic Regression
K- Nearest Neighbourhood.



Classification
dealing with categorical labels
in terms of grouping the data

Binary classification
label will be having
two unique data points

Yes | No
0 | 1
spam | ham

Logistic Regression
K-NN
Naive Bayes
DTC
RFC
SVC

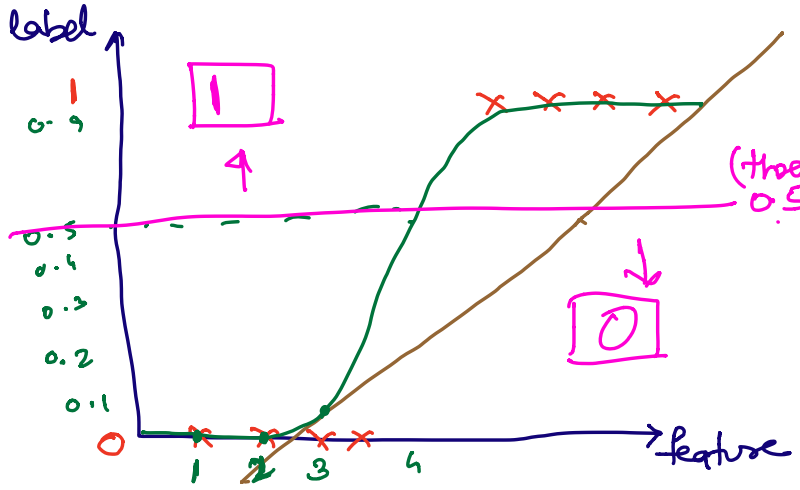
multi-class classification.
label has more than
two unique data points

setosa | versicolor | Virginica
Primary | Social | updates | Promotions

K-NN
Naive Bayes
DTC
RFC
SVC

Logistic Regression (Linear Regression & Sigmoid)

misleading term.



Binary classification

In the world of linear regression I always get a continous numeric value.

fn → sigmoid function -or- sigmoid curve

$$S = \frac{1}{1 + e^{-y}}$$

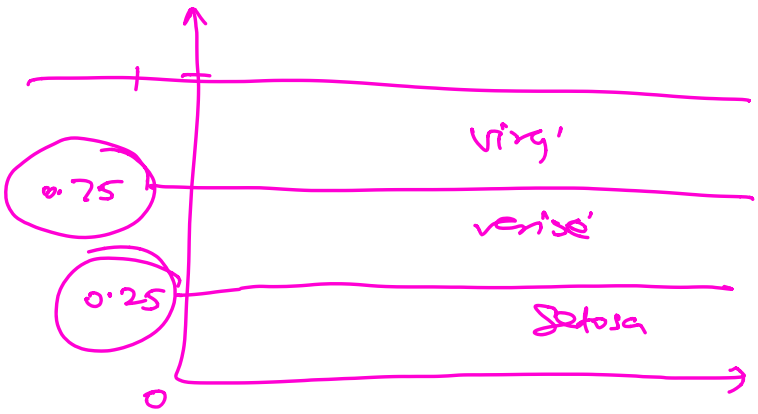
∴ Let y be the linear regression equation.

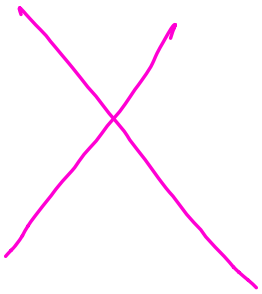
$$S = \frac{1}{1 + e^{-(b_0 + b_1(\text{features}))}}$$

↳ Logistic Regression

The output of sigmoid curve is always a probability value ranging from 0 to 1

for multi-class LR





K - Nearest Neighbourhood

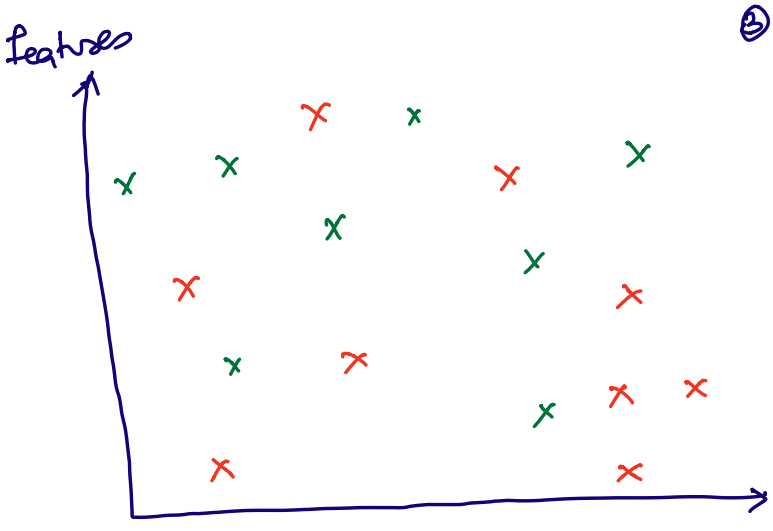
↳ Nearest Neighbourhood.

↳ Distance formula

① Euclidean Distance
② manhattan Distance
③ Minkowski Distance

↓

used Euclidean dist as baseline for distance calc.



x — red
x — green
} binary

k = 3

k represents number of neighbors

voting

Algo :-

① Define the value of K. (Choose only odd numbers)

② Get K nearest data points from the unknown point.

③ Perform voting and select the majority class.

④ Reassign majority class as final class output for the given feature.

EVALUATION METRICS IN SUPERVISED LEARNING

How we approach a regression problem?

① Label is numeric

② Features and label must be a 2D array

③ Ensure our model is a generalized model

$$\text{Score}_{(\text{train})} < \text{Score}_{(\text{test})}$$

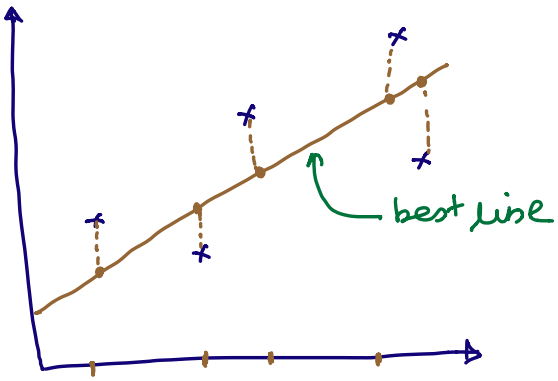
The model is generalized

④ Both features and label must be strictly numeric before initiating model training.

⑤ Evaluate your model.

Check MSE* / MAE / RMSE (Root mean Square error)

(mean squared error) (mean absolute error)

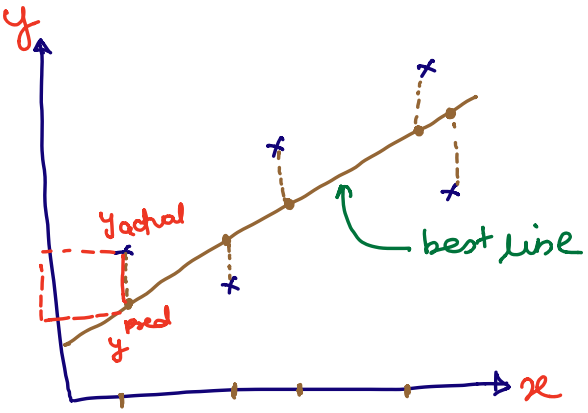


motto of linear regression model is to get the best line.

Which line is the best line?
Any line that has less or no error.

Calculate the error of best line?

MSE (mean squared Error) ↓



$$\text{error for each pts} = (y_{actual} - y_{pred})^2$$

$$MSE = \frac{\sum_{i=1}^{n} (y_{actual} - y_{pred})^2}{n}$$

n is no of data pts.

To get a line with minimal or no error.

⑥ Check score of generalized model. If convinced with score, perform deployment else improve.

How we approach a classification problem?

① Check whether the given problem is a binary class or multi-class problem.

eg. data.Purchased.unique()

② Check whether the given dataset is a balanced dataset or not. (which metric to follow for model evaluation)

by checking the value_counts of your label column.

eg: iris.csv

Setosa	50
Versicolor	50
Virginica	50

} Balanced dataset

③ Ensure features are 2d array and label is 1d array

EVALUATION METRICS FOR CLASSIFICATION.

1. Accuracy
2. Precision
3. Recall
4. F1 score

} Evaluate your classification model

Dataset

- Balanced dataset → Accuracy } w.r.t statistics
- Unbalanced dataset → Precision, Recall, F1 score } w.r.t domain knowledge

CONFUSION MATRIX

Binary classification → 0 & 1

Actual values \ Predicted values	0	1
0	Correct: 50 Predictions	Wrong: 20 Predictions
1	Wrong: 40 Prediction	Correct: 40 Prediction

$$\text{Accuracy (model)} = \frac{\text{Correct Pred}}{\text{Total Predns}} = \frac{50+40}{50+20+40+40} = 0.6$$

Precision → horizontal (each label)

$$\text{Precision}(0) = \frac{50}{50+20} = 0.71$$

$$\text{Precision}(1) = \frac{40}{40+40} = 0.5$$

Recall → vertical
(each label)

$$\text{Recall}_{(0)} = \frac{50}{50+40} = 0.55$$

$$\text{Recall}_{(1)} = \frac{40}{20+40} = 0.6$$

F1 score →

True \ Predicted	Setosa	Versi	Vir
Setosa	50	0	0
Versicolor	0	45	5
Virginica	0	1	49

$$\text{Acc (model)} = \frac{50+45+49}{50+45+5+49+1} = 0.96$$

Precision

$$P_{setosa} = \frac{50}{50} = 1$$

$$P_{versicolor} = \frac{45}{45+5} = 0.9$$

$$P_{virginica} = \frac{49}{49+1} = 0.98$$

Recall

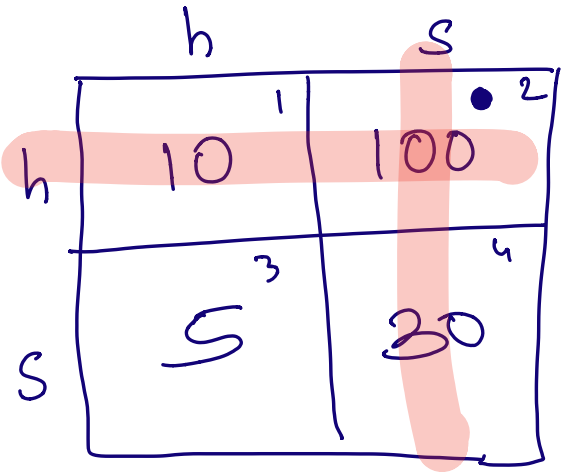
$$R_{setosa} = 1$$

$$R_{versi} = 0.97$$

$$R_{vir} = 0.90$$

Email Spam classification

h- ham
s- spam
dataset is unbalanced!



accuracy = 0.27

$Precision_{(h)} = 0.09$

$Precision_{(s)} = 0.85$

$Recall_{(h)} = 0.66$

$Recall_{(s)} = 0.23$

spam → inbox ✓

→ ham → spam ✗