__/__/____

Machine learning:

- It is subset of AI
- It gives machines the ability to learn by themselves & make predictions
- It uses many algorithms for building models and make predictions using historical data.
- It is used to find hidden patterns in the data.

Examples

- Google detects the faces from photos
- Smaurphone unlocking by face detection
- Amazon recommends the products based on browsing history
- LinkedIn recommending friends you might be likely to connect with

Supervised ML:

- It is a ML method in which we provide labelled data to the model and train the model according to it and then the model predicts the output.

- The goal of supervised learning is to map input data with output data.

Eg: Spam filtering
Based on labelled data we can determine if msg is spam or not

Supervised has 2 categories
- classification
- Regression.

Classification :-

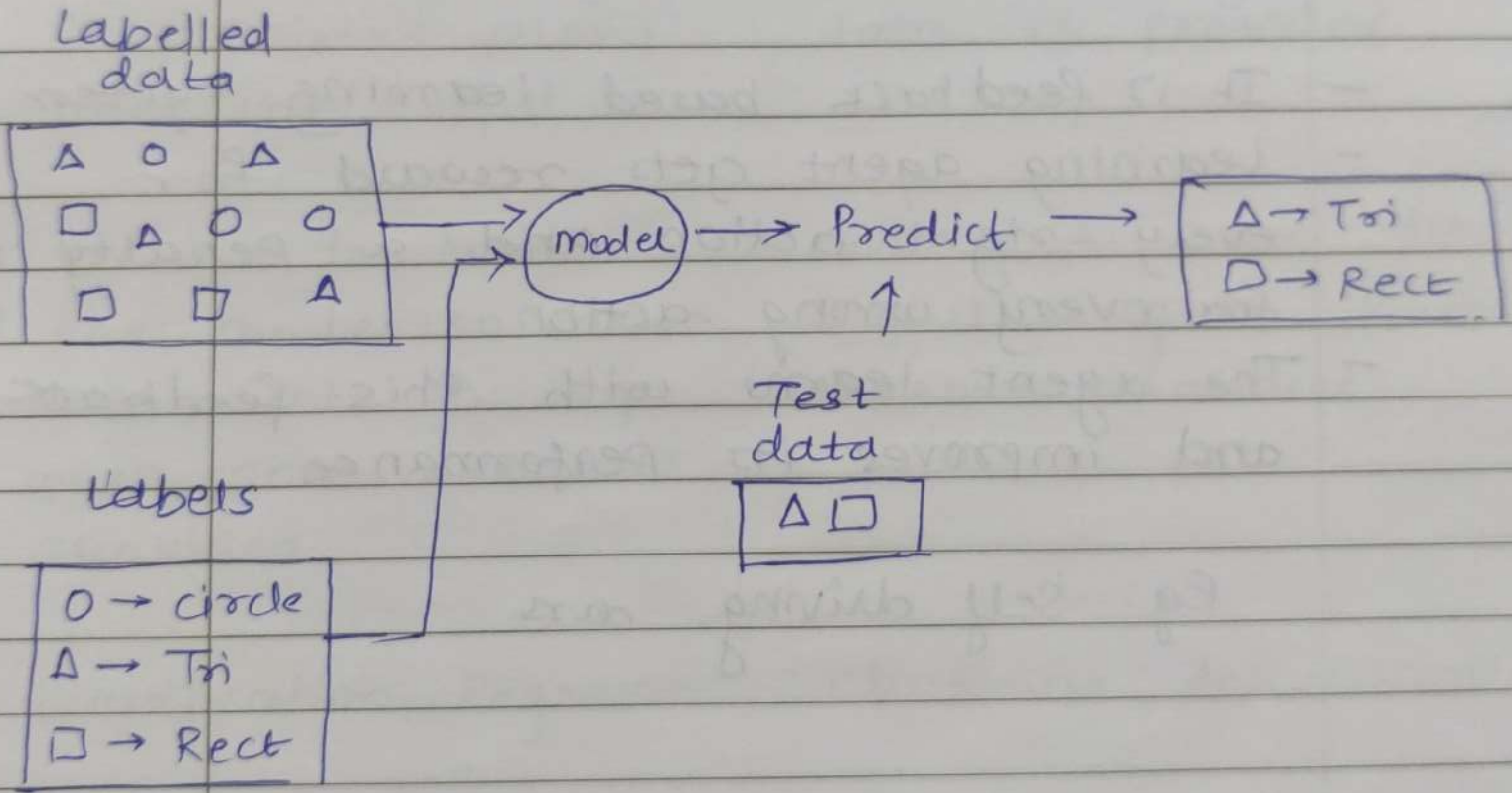
- we want to classify the output into different classes.
- Eg: Person is having diabetes or not
outcome will only be Yes / No

Regression:

- Used if there is a relation between input and output variable.
- We want to find out how the change in i/p affects the o/p.

Eg: Weather Forecasting, Stocks analysis

Working of supervised algorithm



__/__/____

Unsupervised learning:

- In this, model will identify the patterns in the data
- It deals with unlabelled data.
- This algo groups data into clusters and then the model can make conclusions

Reinforcement learning:

- It is feedback based learning system.
- Learning agent gets reward for every right action and a penalty for every wrong action.
- The agent learns with this feedback and improves its performance.

Eg: Self driving cars.

__/__/____

Supervised	Unsupervised
- Trained using labeled data.	- Trained using unlabbed. data.
- SL model predicts the output.	- UL model finds the hidden pattern in data
- In SL, input data is provided along with the output.	- In UL, only input data is provided.
- Goal is to train the model so that it predicts the o/p when new data is provided.	- Goal is to find the useful insights from unknown dataset.
- classification, Regression	- Clustering, Association.

* Linear Regression:

- used to predict relationship between two variables.
- It aims to find the best fit line.

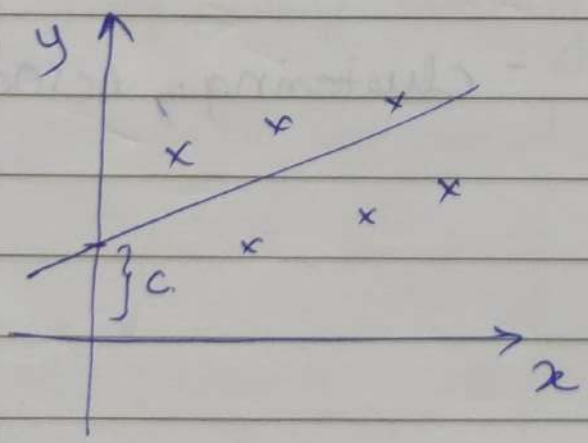
$$y = mx + c$$

$m \rightarrow$ slope

$c \rightarrow$ intercept / constant

Best fit line:

The best fit line is the line that has the least error. This means that the error between the actual and predicted values is minimum.



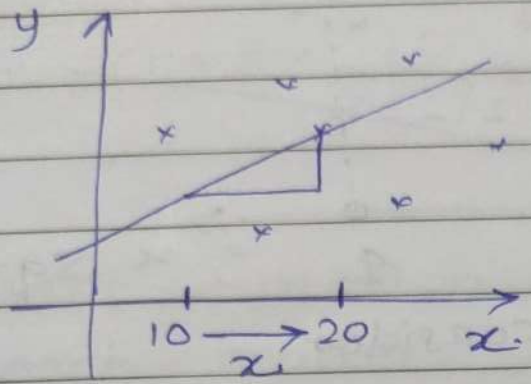
$y = mx + c$

when $x = 0$,

$y = c$

$\therefore$ c is the value on y-axis when x is 0.

Slope (m)

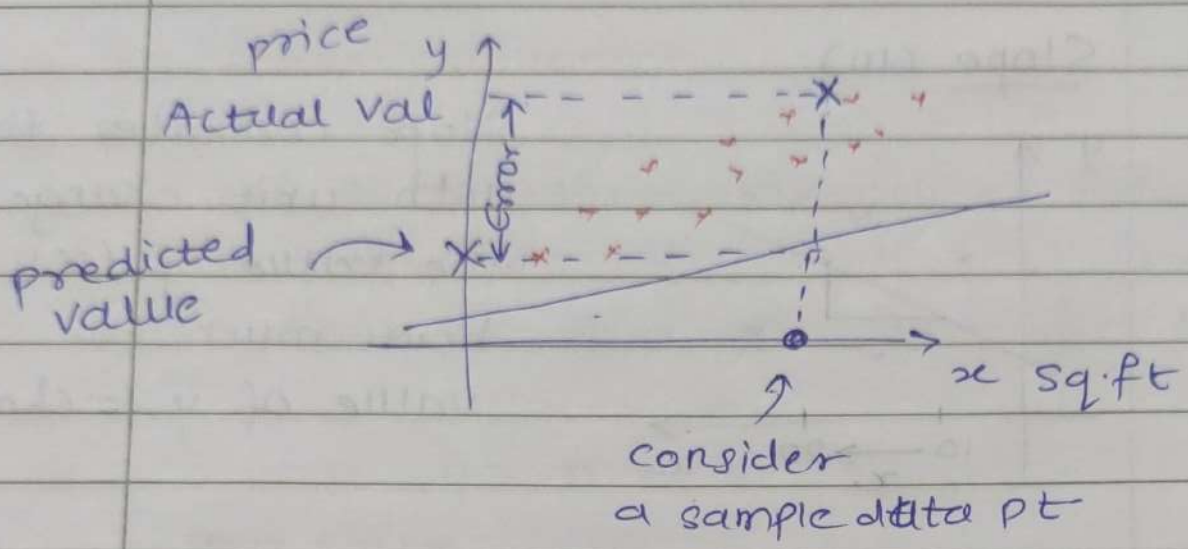


Slope indicates that with unit change in the value of (x), how much the value of y is changing

- We can draw multiple best fit lines to select the best one.
- Whichever best fit line gives least error that line can be selected and it will also give the slope (m), and c.

* Cost function.

- Represents the error between the actual values and predicted values
- The main goal is to minimize the cost function so as to find the line with least error.



For linear regression,

for m number of data points,

$$\text{Cost Funtion} = \frac{1}{2m} \sum_{i=1}^{m} (y_A - \hat{y})^2$$

$y_A \rightarrow$ actual value

$\hat{y} \rightarrow$ predicted values

We are calculation sum of all errors

We multiply by 1/2m just for averaging purpose.

Cost function can also be written as,

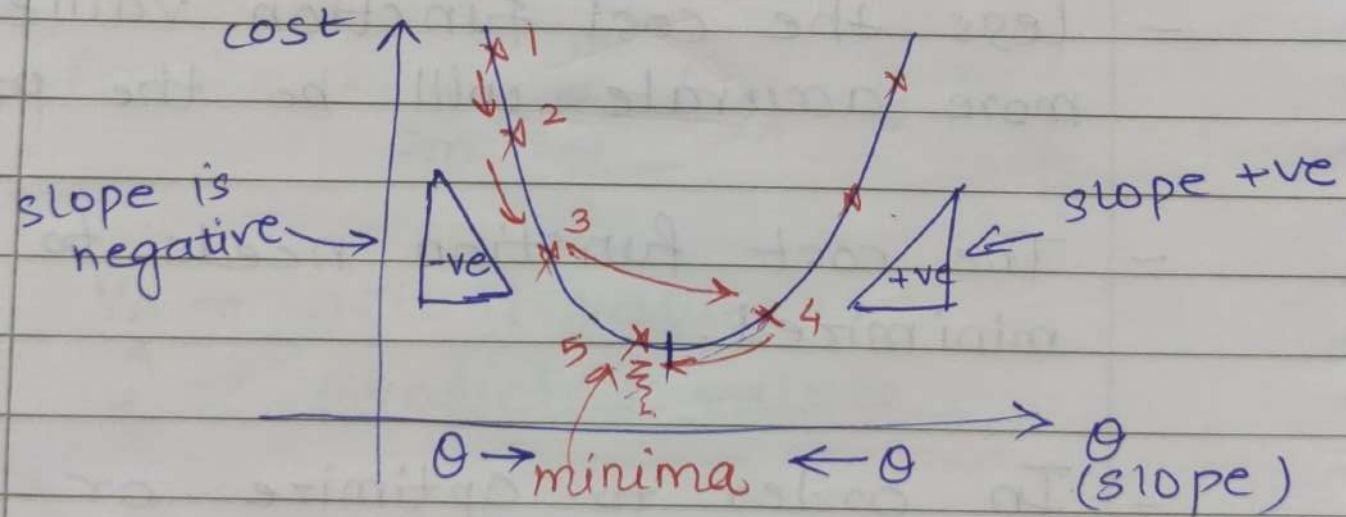
$$\text{cost} = \frac{1}{m} \sum_{i=1}^{m} (|y_A - \hat{y}|)$$

- Cost function will differ for each model.
- Cost function also determines how well our model is predicting for a given dataset.
- Less the cost function value, more accurate will be the prediction
- The cost function needs to be minimized.
- In order to optimize or minimize the cost function, we need to use gradient descent.

__/__/____

Gradient Descent:

- Used to minimize the cost function so that we get best fit line
- It is a repetitive process where the model eventually reaches near the minima.
- At this minima, the error is the least and cost function is optimized.



- on x-axis we have θ which is the slope
- on y axis we have the cost function.

- Suppose we are at pt 1 and we have to reach minima, then we take a step at a time.

- After taking a step we reach at point 2 and 3.

- At the left of minima, the slope is negative hence θ will increase. Hence, the point will come down to reach minima.

- Now we take step from 3 and then we reach step 4..

- As the step is too big it will overshoot the minima.

- The slope on right of minima is +ve. Hence, the θ will decrease and we will come close to minima.

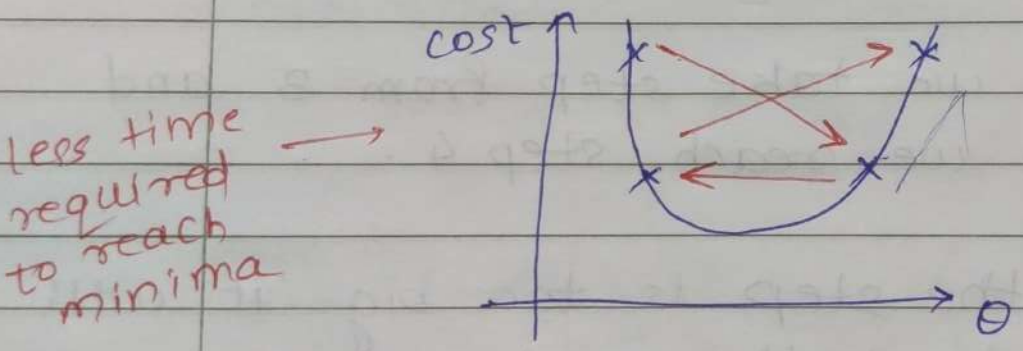
- In this way, continuous backtracking happens to reach the minima.

- The number of steps that are taken to reach the minima is called as learning rate.

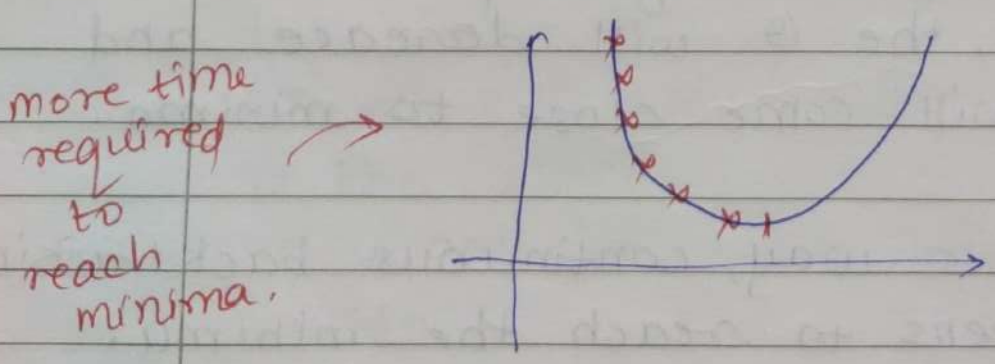
$$\theta = \theta - \alpha \left(\frac{\partial \text{cost}}{\partial \theta} \right)$$

$\alpha \rightarrow$ learning rate

- When learning rate is high, overshooting occurs,



- When learning rate is low.



* Residuals

- The difference between actual value and predicted value is called as residual. $(y_A - y_P)$

* RSS → Residual sum of squares.

- It is defined as the sum of squares of the residuals.

$$\sum_{i=0}^{m} (y_A - \hat{y})^2.$$

* ESS → Explained sum of squares.

$$\sum_{i=0}^{m} (y_{pred} - y_{mean})^2$$

* TSS → Total sum of squares.

$$\sum_{i=0}^{m} (y_{Actual} - y_{mean})^2$$

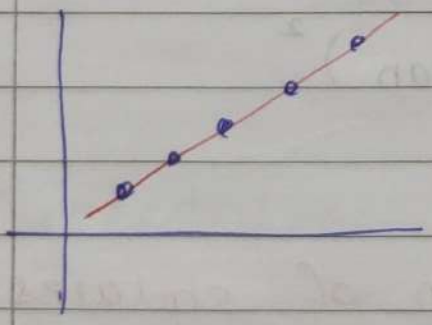
Linear Reg is evaluated by,

* R2 (R squared)

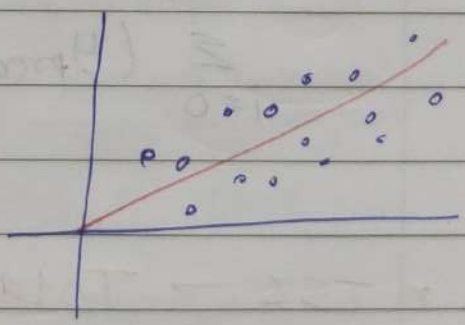
- It is a number that explains the amount of variation captured by the model.
- It ranges from 0 to 1
- The higher the val of R2, the higher is the prediction accuracy.

$$R^2 = 1 - \frac{RSS}{TSS}$$

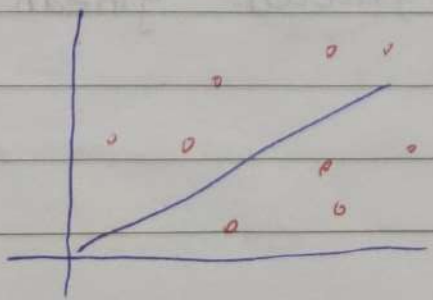
i) $R^2 = 1$



ii) $R^2 = 0.5$



iii) $R^2 = 0.05$



__/__/____

Linear regression is also evaluated by
- Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{RSS}{n.}}$$ ← total data pts.

* **Bias**

- Bias can be referred to as the gap between the actual & predicted values
- High bias means the gap between predicted & actual values is large.
- Low bias means the gap between predicted & actual value is very less.
- A model should have low bias.

* **Variance**

- Its a measure of how the predicted values are scattered from each other.

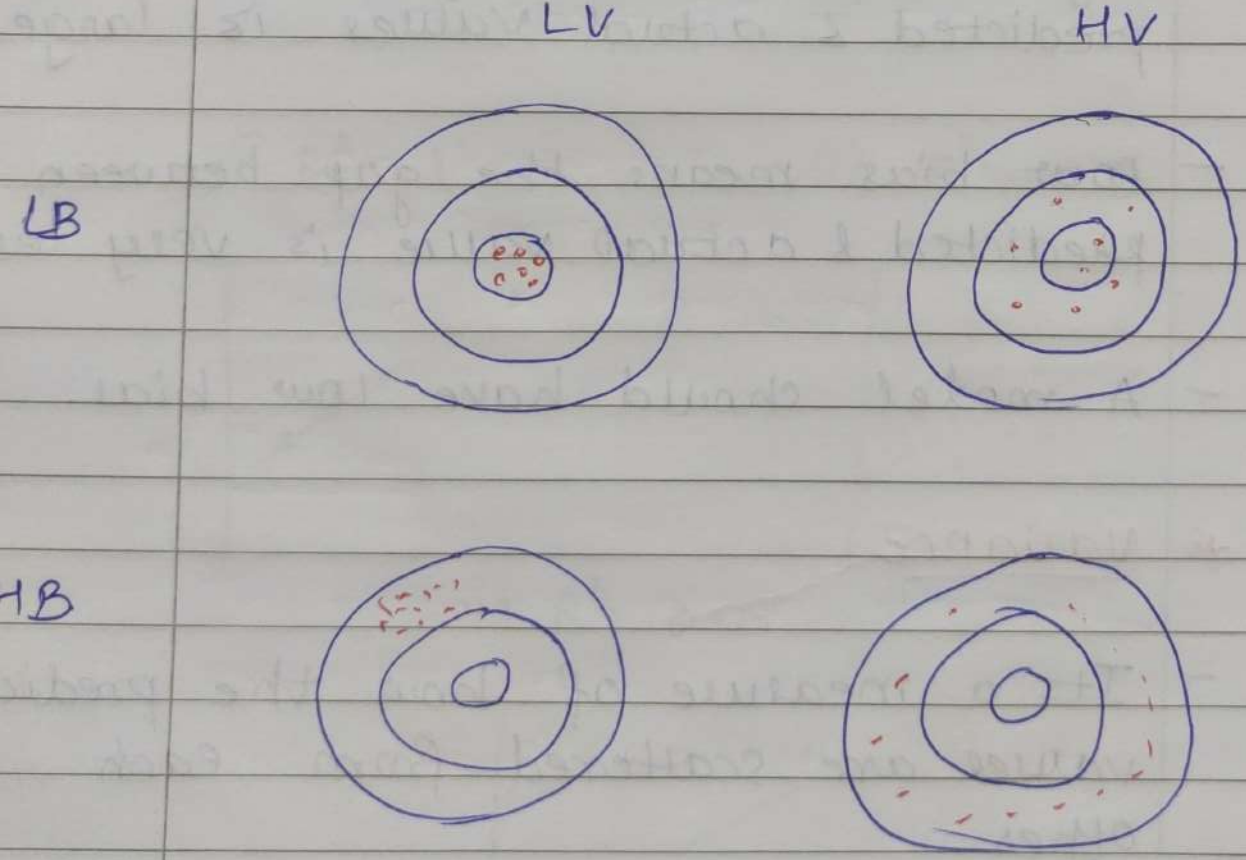
__/__/____

- Low variance means that predicted values are not that scattered from each other.

- High variance means predicted values are very much scattered.

- A model should have low variance.

Bias and variance are inversely proportional to each other



- To overcome overfitting, we need to do regularisation.

Regularisation

- Regularisation is a method to reduce the model complexity by reducing the coefficients.

$$y = 23x_1 + 45x_2 + c$$

After regularisation,

$$y = 1.3x_1 + 3.2x_2 + c$$

Ridge Regression
Lasso Regression
Elastic net Regression.

* Ridge

- In ridge regression, we add a penalty term which is equal to the sum of square of coefficients.

__/__/____

- Ridge expression can be given as,

$$\text{Ridge} = \text{Loss} + \alpha \|w\|^2$$

Loss → diff between actual and predicted values.

$\alpha \|w\|^2$ is penalty.
α is the constant
$\|w\|^2$ → sum of sq. of coefficients
$|w|^2 \rightarrow w_1^2 + w_2^2 + \cdots + w_n^2$.

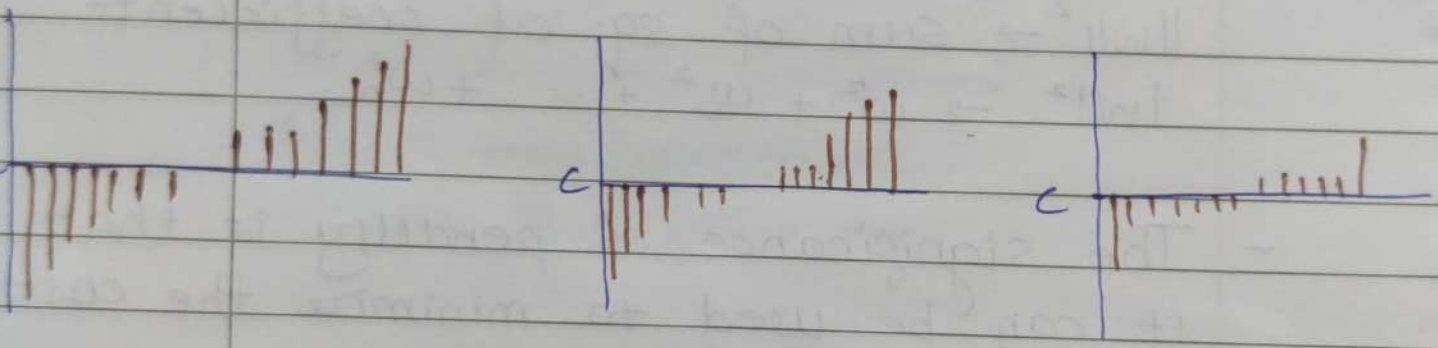
- The significance of penalty is that it can be used to minimize the cost and hence reduce co-efficients.

Eg: car accident happens and we suffer a loss.
If penalty is paid for accident, effect of loss will reduce.

- By changing the value of α, we are controlling the penalty.

- Higher the value of α, bigger is the penalty and therefore the magnitude of coefficients are reduced.
- By using ridge regression, the values of coefficients can come close to zero but not exactly zero.

$\alpha = 0.05$ $\alpha = 1$ $\alpha = 2$



c → co-efficients of x

Limitations of Ridge

- It cannot reduce the no. of variables as coefficient cannot go till zero. Hence it cannot be used for feature selection.

__/__/____

* Classification

- supervised learning algo.
- Response has multiple class labels
- We are given labelled training data and we have to predict our predictor falls in which class label.

Eg: classify email as spam / not spam.

4 types of classification.

1) Binary classification
2) multi-label classification
3) multi-class classification
4) Imbalenced classification

Binary classification:

- In this there are only 2 categories in the output.

Eg: Spam / Not spam.

__/__/__

Multi-class classification:

- There can be any no. of categories in the response.
  Eg: Types of crops.

Multi-label classification:

- There can be 2 or more categories in the response
- And the data which we want to classify can belong to many categories all at the same time.
  Eg: classify which traffic signs are present in image.

Imbalenced classification.

- In this, the dataset is imbalenced.
- The target column suppose has 90 Yes and 10 No's.
  Major part of dataset that will go for training would be 'Yes' response Hence predictions can be inaccurate

  Eg:

__/__/__

* Lazy Learning & Eager learning.

- bifurcated based on training style.

- Generally we give the training data & model is built on the data and then predictions are done.

Lazy learner

- In this model takes the input data but does not train it.
- The algorithm will just store the data.
- When the prediction has to be done, algorithm uses the stored data, perform all operations and then return the o/p.
- The training time is very less and prediction time is high as all operations are performed while doing predictions

Eg: KNN

__/__/____

Eages learning:

- It works like a traditional ML algo.
- The data is processed during the training phase.
- Hence, this increases training time.
- Predictions are very fast.

Eg: Linear Reg, Logistic Reg
Decision tree, Naive Bayes.

* Logistic Regression

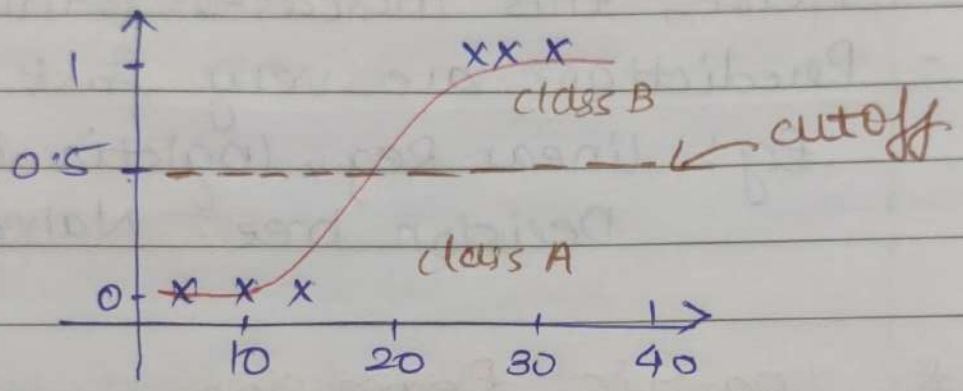
- It is a classification model.
- Used when the response has categori
mostly two categories.
- Eg: Student has passed exam or not → Yes / No.

$$y = \frac{1}{1+e^{-x}}$$ } Sigmoid fn.

$x \rightarrow$ independent variable
$y \rightarrow$ dependent variable
$e \rightarrow$ euler's constant 2.718

- Instead of fitting a regression line, we fit an 'S' shaped logistic function which predicts 2 values (0/1)

- Sigmoid function is simply toying to convert the independent variable into expression of probability that ranges from 0 to 1 with respect to dependent variable (y).



- If a data point falls below cutoff, then it belongs to class A.
- If a data point falls above cutoff, then it belongs to class B.

- If any data point falls on the cutoff, then it cannot be classified.

- The dataset needs to be free of null values if we want to use logistic regression.

Linear Regression	Logistic Regression
- Used to predict a continuous dependent variable with given set of independent variables.	- Used to predict a categorical dependent variable with given set of independent variable
- Used for solving regression problem	- Used for solving classification problem.
- In this, we find the best fit line	- In this we find the S-curve
- The dependet and independent variables should have a linear relationship	- Linear relationship is not required.
- There might be high correlation between independent variables.	- There should be no correlation between independent variable.

* Naive Bayes classifier

- It is a classification algorithm based on Bayes theorem.
- It goes with the assumption that all the independent vaidble should not be related to each other.
- It also assumes that all the attributes must equally contribute to the response.

Conditional Probability → $P(A|B)$

↓

probability of A happening given that B has occured.

likelihood →

$$\text{posterior} \rightarrow P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \leftarrow \text{prior}$$

$\leftarrow$ marginal

A is the hypothesis
B is the evidence

We have to find a hypothesis given that we have some evidence.

Likelihood → P(B|A)
Probability of the evidence given that the hypothesis is true.

Prior → P(A) → probability of hypothesis before considering the evidence.

Marginal → P(B) → probability of the evidence

posterior → P(A|B) → probability of hypothesis given we have some evidence

Eg: $P(\text{king} \mid \text{Face}) = \frac{P(\text{Face} \mid \text{king}) \times P(\text{king})}{P(\text{face})}$

$= \frac{1 * 4/52}{12/52}$

$= \frac{1}{3}$

__/__/____

	Outlook	Temp	Humidity	Windy	Play
0	Rainy	hot	high	False	No
1	Rainy	hot	high	True	No
2	Overcast	hot	high	False	Yes
3	Sumy	Mild	High	false	Yes
4	Sunny	cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	mild	High	False	No
8	Rainy	cool	Normal	False	yes
9	Sunny	mild	Normal	False	Yes
10	Rainy	mild	Normal	True	yes
11	Overcast	mild	high	True	yes
12	overcast	hot	Normal	False	Yes
13	sunny	mild	high	True	No

Features: outlook, temp, humidity, windy

Response: Play

Outlook.

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5		

Temperature

	Yes	No	P(yes)	P(No)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5		

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5		

__/__/____

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5		

Play		P(yes) / P(no)
Yes	9	9/14
No	5	5/14
Tot	14	

- We have done all the necessary computations hence the classifier is ready.

Testing the classifier:

today = (Sunny, Hot, Normal, False)

Sunny → Outlook, Hot → Temp, Normal → Humidity, False → Windy

Probability of playing golf is,

$$P(\text{Yes}|\text{today}) = \frac{P(\text{Sunny}|\text{Yes})\,P(\text{Hot}|\text{Yes})\,P(\text{Normal}|\text{Yes})\,P(\text{False}|\text{Yes})\,P(\text{Yes})}{P(\text{today})}$$

Probability of not playing golf is

$$P(\text{No}|\text{today}) = \frac{P(\text{Sunny}|\text{No})\,P(\text{Hot}|\text{no})\,P(\text{Normal}|\text{no})\,P(\text{False}|\text{no})\,P(\text{no})}{P(\text{today})}$$

- As P(today) is common in both probabilities, they can be ignored.

$$P(\text{Yes}|\text{today}) = \frac{2}{9} * \frac{2}{9} * \frac{6}{9} * \frac{6}{9} * \frac{9}{14}$$

$$\approx 0.0141$$

$$P(\text{No}|\text{today}) = \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * \frac{2}{5} * \frac{5}{14}$$

$$\approx 0.0068$$

P(Yes|today) + P(No|today) = 1

- The obtained nos can be converted into probability by making sum = 1

$$P(Yes|today) = \frac{0.0141}{0.0141 + 0.0068} = 0.67$$

$$P(No|today) = \frac{0.0068}{0.0141 + 0.0068} = 0.33$$

P(Yes|today) > P(No|today)

Hence it is predicted that with today = (Sunny, Hot, Normal, False)

↓

Golf will be Played.

* How do Naive Bayes work?

1) Convert dataset to freq table.
2) Find the probabilities table
3) Use naive bayes equation to calculate the posterior probability for each class label

(9) How does naive bayes treat numerical and categorical values.

- For categorical value we can calculate the posterior probability.

- When the values are numerical, it is assumed that they follow Gaussian / Normal distribution.

- It calc mean & std. for each class and then uses these values for classification.

* Multinominal Naive Bayes:

- It is commonly used for text and document classification.
- We can count how many times the word is occuring in the document.
- The features / predictors used by this classifier is the frequency of words.

* Bernoulli Naive Bayes:

- In this the predictors are boolean values.
- The parameters that we use to predict class variable only takes up the value Yes/No.
- Eg: We have a document of words, we can have 0 as word found in the document
  1 → word not found in document

* Gaussian Naive Bayes

- It is used in classifications where the features follow normal / Gaussian distribution.

Q) When to use Naive Bayes?

- When we have to perform text classificatn,
- When there are large no. of features.
- When we have to perform multi-class classification.
- When features are independent.
- It performs well even with limited training data.

__/__/____

* Applications of Naive Bayes.

- Can be used to make realtime predictions
- Can be used for multi-class predictions.
- It has high success rate hence can be used for spam filtering
- It can also be used for sentiment analysis to identify +ve/-ve reviews.
- Naive Bayes along with collaborative filter can be used in recommend^n sys.

NAIVE BAYES IS GENERATIVE ALGO

* K-N-N (k-nearest neighbours)

- It is supervised → classification → lazy algo
- It works for classification as well as regression.
- It makes decisions based on the similarity of a new data point with its k-nearest neighbours.

why is KNN lazy?

- During training, KNN just stores the entire dataset for reference
- When any prediction has to be done, it calculates the distance of the new data point with the training data
- on the basis of the distance it selects the neighbours & predicts the class label.

- When knn is used for regression, it predicts a continuous value as o/p.
- Average value of target value of knn is used to make predictions.
- If K = 5, it will take avg of target of 5 neighbours and assign that value to new data pt.

__/__/____

When to use KNN ?

1) Small - medium dataset
2) When there is non-lineal relationship between response & predictor
3) When we have noise in data. Noise can be dealth as there are multiple neighbors to make decisions.
4) Supports multi-class clasification

* When K is too low...

- When K is too low, algorithm becomes too sensitive to noise & outliers.
- This degrades algorithm peformance.
- Overfitting occurs when K is too low as only few pts will be selected for predictions.

* When K is too high

- It smooth out decision bounday too much.
- It causes underfitting.
- It will lead to biased clasification if there is imbalance in dataset.
- Using high K value will not capture variations.

Distance metrics in knn.

1) Euclidean distance

- cartesian dist between two points.

$$d(x,y) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

2) Manhattan distance:

- defined as sum of absolute distance between co-ordinates.

$$d(x,y) = |x_2-x_1| + |y_2-y_1|$$

3) Minkowski distance:

- Euclidien and manhattan are special cases of minkowski

$$d(x,y) = \sum_{i=1}^{n} \left[(x_i-y_i)^p\right]^{1/p}$$

* Default metric used by KNN is EUCLIDEAN DISTANCE