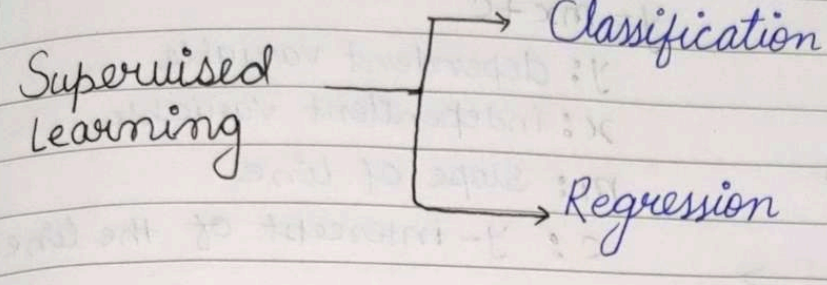
UNIT: 2.1

Dashrath Nandan

AIML

Date :



Classification is a type of supervised machine learning that involves training a model to identify which category or class an input data point belongs to

* Supervised learning is where the model is trained on a labelled dataset. The labelled dataset used in supervised learning consist of input features and corresponding output lables.

* Classification are used to solve the classification problem in which the output variable is categorical, such as Yes or No, Male or Female, etc.

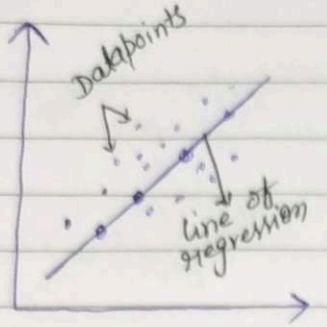
  Eg: Random Forest Algo, Decision Tree Algo, SVM algo,

* Regression is a supervised learning technique used for predicting continous or real-valued output based on input feature. It's aims to establish a relationship between the Independent Variable (Input) and dependent variable (Output) by using regression model to the training data.

Linear Regression

- Linear regression algorithm shows a linear relationship between a dependent variable and one or more independent variable.
- It is a statistical method that is used for predictive analysis.

Date :



$y = mx + c$

y: Dependent variable

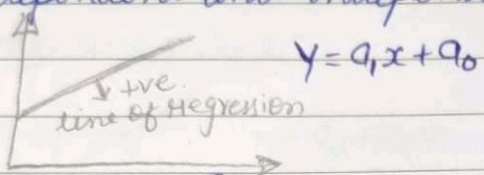
x: Independent variable

m: Slope of line

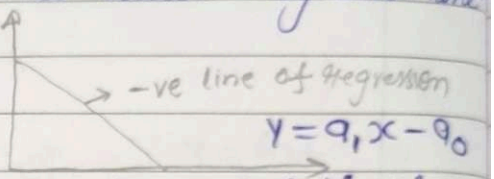
c: y-Intercept

Linear Regression line:

A linear line showing the relationship between the dependent and independent variable is called regression line.



Positive linear Relationship



Negative linear relationship

* Assumptions of linear Regression:

1) Linear relationship between the features and target.
2) Small or no multicollinearity (high-correlation) b/w features.
3) Homoscedasticity assumption: Homoscedascity is a situation when the error term is same for all value of indepen. variable.
4) No autocorrelations in error terms.

* Types of Linear Regression:

dependent variable — Independent variable

- I) Simple Linear Regression: 1 D.V. (interval or ratio), 1 IDV (interval or ration)
- II) Multiple Linear Reg.: 1 DV (interval ratio), 2+ IDV (interval or ratio or dichotomous)
- III) Logistic reg.: 1 DV (dichotomous), 2+ IDV (interval or ratio or dichotomous)
- IV) Ordinal reg.: 1 DV (ordinal), 1 + IDV (nominal or dichotomous)
- V) Multinomial reg.: 1 DV (nominal) + 1 + IDV (interval or ratio or dichotomous)

Date :

The major uses of regression analysis are
(1) determining the strength of predictors
(2) forcasting an effect (3) trend forcasting

Some popular applications of linear regression are:

Analyzing trends and sales estimates
Salary forecasting
Real estate prediction
Arriving at ETAs in traffic.

* Simple linear Regression

- It is a type of regression algorithm that models the relationship between a dependent variable and a single independent variable.

- The dependent variable must be a continuous/real value.

- $E(Y) = a_1 x + a_0 + \varepsilon$ Y: Y is the estimated value of y for a given value of x.

a_1: slope , ε: Error term.

- RSS (Residual Sum of Square) $= \sum (\text{actual output} - \text{predicted output})^2$
$= \sum_{i=1}^{n} (y_i - (mx_i + c))^2$

- Formula to calculate:

$$m(\text{slope}) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}, \quad c(\text{intercept}) = \frac{\sum y \sum x^2 - \sum x \sum xy}{\sum x^2 - (\sum x)^2}$$

n: number of samples.

* Multiple linear Regression:

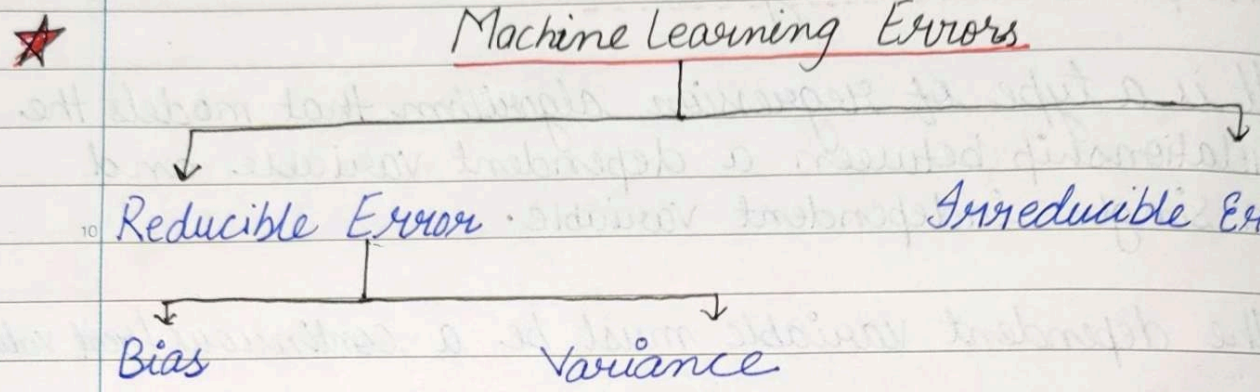
It is a type of regression algorithm that models the relationship between a single dependent variable and more than one independent variable.

$$y_i = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + \varepsilon$$

Assumptions: • A linear relationship b/w dependent and independent variable.
• The independent variable are not highly correlated.

Date :

Decision boundary seperates two or more classes from one another. It is used to predict which class a data belongs.



Bias: Bias is defined as the inability of model because of that there is some error occur between the model's predicted value and actual value.

- It is a systematic error that occurs due to wrong assumptions in ML process.

a) Low bias value means fewer assumptions are taken to build target function. Model will closely match Training dataset.

b) High bias value means more assumptions are taken to build target function.

- Ways to reduce High Bias in ML
  - Use a more complex model.
  - Increase the no. of features.
  - Increase the size of training dataset.

Bias-Variance Tradeoff: As $V\downarrow \rightarrow B\uparrow$
As $B\downarrow \rightarrow V\uparrow$

Date :

* **Variance:**
Variance is the measure of spread in data from it mean position when it is trained on different subsets of the training data.

a) Low variance means that the model is less sensitive to changes in the training data and produce consistent output.

b) High variance means that the model is very sensitive.....

- Ways to Reduce Variance in ML
  - Cross-Validation
  - Feature Selection
  - Regularization: Use L1 or L2 regularization
  - Early stopping

* **Validation:** Model validation is refferred to as the process where a trained model is evaluated with a testing dataset.

* **Cross Validation (CV):** It is one of the technique used to test the effectiveness of a ML models.

- **Steps**
  - Reserve some portion of sample data set.
  - Using the rest data-set train model.
  - Test the model using reserve portion....

- **Methods of CV**
  i) Train Test Split.
  ii) Hold out method.
  iii) K-flod cross validation
  iv) Leave P-out CV

- **Different combination of Bias-Variance:**

	Low Variance	High Variance
Low Bias	(i) → Ideal / Not practically possible	(ii) → Overfitting
High Bias	(iii) → underfitting	(iv) → Inaccurate / Inconsistent

Date :

Methods of Cross Validation:

1) Train test split :- Split the complete data set into training and test sets. (Train → Test)

2) K-Fold Cross Validation: The procedure has a single parameter called 'K' that refers to the number of groups that a given set is to be split into.

eg: K = 4,

If you are in i^{th} iteration, then select set i for test and remaining for training the data.

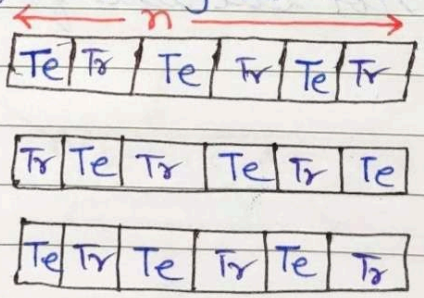
	set 1	set 2	set 3	set 4
iteration 1	Test	Train	Train	Train
itr^n. 2	Train	Test	Train	Train
itr^n. 3	Train	Train	Test	Train
itr^n. 4	Train	Train	Train	Test

3) Leave P out Cross Validation:
If there are n data points in the original sample then, n-P samples are used to train the model and p points are used to Test or validate set.

⇒ We select P example/set (randomly) for testing in each iteration.

eg: P = 3,

n: Total data Points
P: testing data
n-P: Training data.



Te = testing
Tr = training

4) leave One Out Cross Validation: A particular case of leave P out CV, when P = 1.

⇒ This process repeats for each datapoints. Hence for 'n' samples, we get n different training sets and n test set.

Date :

☆ Performance metrics : To evaluate the performance or quality of the model, different metrics are used, which are known as performance or evaluation metrics.

Some important aspects to access are — Robustness, Correctness, efficiency, Scalability, Bias and fairness, Real-world performance, etc.

★ Performance Metrics for Classification

(1) Accuracy : $$\text{Accuracy} = \frac{\text{No. of Correct Predictions}}{\text{Total no. of predictions}}$$

(2) Confusion Matrix : It is a tabular representation of prediction outcome of any binary classifier.

	Predicted : NO	Predicted : Yes
Actual : NO	True Negative (TN)	True Positive (TP)
Actual Yes	False Negative (FN)	False Positive (FP)

(3) Precision : It is used to overcome the limitations of Accuracy.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

(4) Recall or Sensitivity: It aims to calculate the proportion of actual positive that was identified incorrectly.

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

(5) F-Score: It is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class.

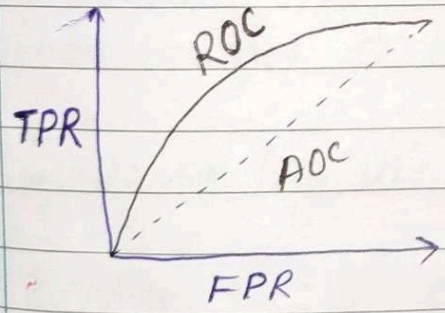
$$\text{F-score} = 2 * \left(\frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \right)$$

F-score ↓ check productivity or to maintain stability of system

↓ Harmonic mean

(6) AUC-ROC: "Area Under the Curve" of the "Receiver Operating Characteristics" curve.

It shows the performance of classification of model at different threshold level.



$$TPR = \frac{TP}{TP+FN} \quad , FPR = \frac{FP}{FP+TN}$$

Application: Classification, Healthcare, Binary classification.

★ Performance Metrics for Regression

1.) Mean Absolute Error (MAE): It measure the absolute difference between actual and predicted value.

$$MAE = \frac{1}{N} \sum \left| \text{Actual output} - \text{Predicted output} \right|$$

Date :

It measures the average of the Squared difference between predicted values and the actual value

2.) Mean Squared Error (MSE) $= \frac{1}{N} \sum (\text{Actual output} - \text{Predicted output})^2$

3.) $RMSE = \sqrt{MSE}$

R squared error is also known as Coefficient of Determination

4.) R-Squared Score: It compare our model with a constant baseline to determine performance.

5. Adjusted R-Squared $\Rightarrow$ $R^2 = 1 - \frac{MSE(Model)}{MSE(Baseline)}$

Error rate = $\frac{\text{No. of incorrect Prediction}}{\text{No. of Predictions Total}}$

Specificity = $\frac{\text{True Negative}}{TN + FP}$

UNIT: 2.2

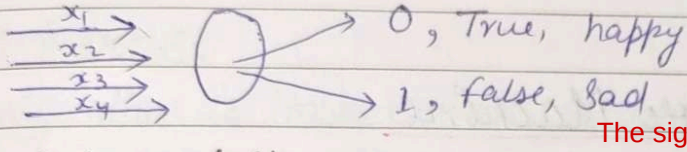
Dashrath Nandan

Date :

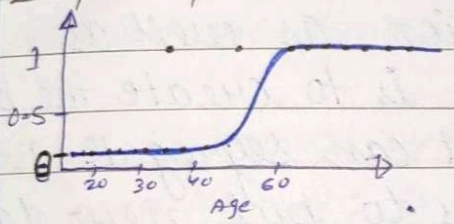
★ Logistic regression:

It is used when the dependent variable ( ) is catagorical.

- Dependent variable is binary : 1 (True, Success), 0 (False, Failure)
- Independent Variable can be continous or binary



Graph Representation:



$$y = \frac{1}{1+e^{-(mx+b)}}$$

$$Sigmoid(z) = \frac{1}{1+e^{-z}}$$

The sigmoid function is a mathematical function used to map the predicted values to probabilities

It maps any real value into another value within a range of 0 and 1.

Sigmoid function in logistic regression is simply trying to convert the I.DV(z) into a expression of probability that ranges 0 & 1 w.r.t. Dependent variable.

* Types of Logistic Regression:

(i) Binary L.R. : The categorical has only 2 possible outcomes.
Eg: Spam or not.

(ii) Ordinal L.R. : Three or more categorical with ordering.
Eg: Movie rating 1 to 5.

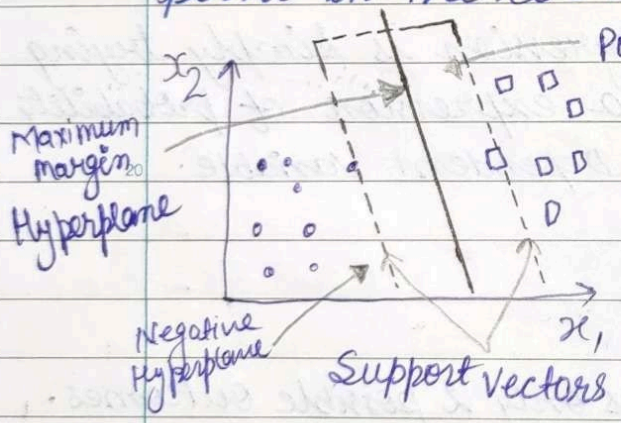
(iii) Multinomial L.R : Three or more without ordering categorical outcomes.
Eg: Which food reffered more (veg, non-veg, vegan)

Date :

* Linear Regression	Logistic Regression
• Predict continous dependent var.	Predict Categorical depdent. var.
• Used for solving Regression problems.	Used for solving classification problems.
• The output is Continuous value.	The output is categorical value, such as 0, 1, etc.

★ Support Vector Machine

SVM is used for classification as well as regression problems. It's goal is to create the best line or decision boundary that can segregate n-dimensional space into classes to put new data point in the correct category in future.

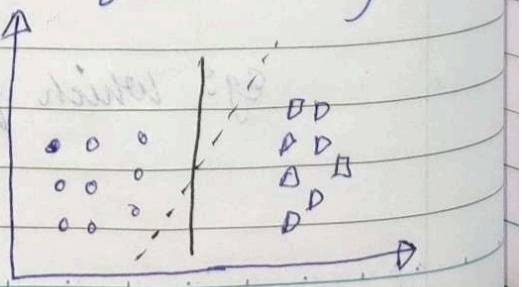


Hyperplane: The best fit line or boendary.

Margin: The distance between vectors and hyperplane.

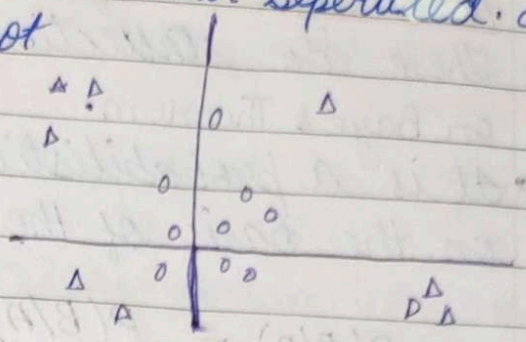
* Types

1) Linear SVM: Linear SVM is used for linearly seperable data, which means dataset can be classified into two classes by using a single straight line.



Date :

2.) Non-linear SVM: It is used for non-linear seperated data, which means dataset cannot be classified using a straight line.



* SVM Kernel

SVM algorithm is implemented with kernel that transforms non-seperable problems into seperable problems by adding more dimensions to it.

Types of Kernal used by SVM:

1) Linear Kernel: Dot product b/w any two observation.

$$K(x, x_i) = sum(x * x_i)$$

2) Polynomial Kernel: More generalized form of linear kernel.

$$K(x, x_i) = 1 + sum(x * x_i)^d$$ → degree of polynomial

3) Radial Basis Function (RBF): Mostly used in SVM classification.

$$K(x, x_i) = exp(-gamma * sum|x - x_i|^2$$

gamma ranges from 0 to 1.

* Pros: i) It is really effective in higher dimension.
ii) Effective where the no. of feature are more.
iii) Best algorithm when classes are seperable.

Cons: i) For large dataset, it requires a large amount of time.
ii) Doesn't perform well in case of overlapped classes.

Date :

★ Naive Bayes Classifier Algorithm:
These are collection of classification algorithm on Baye's Theorem.
• It is a probabilistic classifier i.e. it predicts on the basis of the probability of an object.

Likelihood prob.
Posterior Prob. ← $P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$

* Steps: i) Convert given dataset into freq. table.
ii) Generate likelihood table.
iii) Use Bayes theorem to calculate the posterior probability.

* Types:
i) Gaussian: It assumes that features follow a normal distribution.
ii) Multinomial: It is used when data is multinomial distributed.
iii) Bernoulli: Same as Multinomial, but the predictor variables are the independent boolean variables.

* Application: Credit Scoring, medical data classification, Text classification such as Spam filtering.

* Advantages: It is one of the fast and easy ML algo.
• Can be used for binary as well as multi-class classification.
• Most popular choice for text classification problem.

Disadvantages:
→ It assumes all features are independent, so it cannot learn the relationship features.

Date :

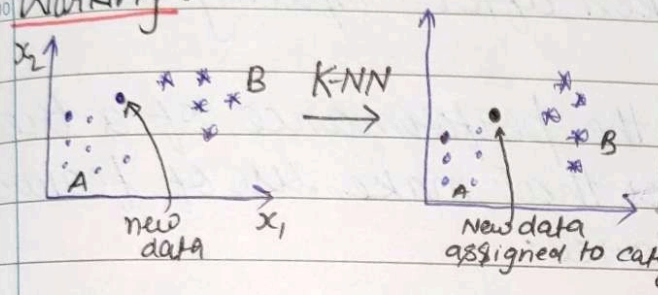
K-nearest Neighbor (KNN) Algorithm :

KNN uses 'Feature similarity' to predict the value of new datapoints.

Lazy learning algorithm because it dosn't have specialized training phase.

Non-parametric learning algorithm because it doesn't assume anything about the underlying data.

Working :



Step 1 : Select the no. of K of the neighbors.

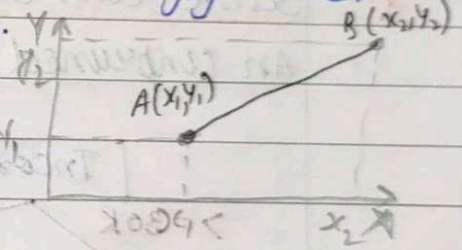
Step 2 : Calculate Euclidean distance

Step 3 : Take K nearest neighbor as per calculated Euclidean dist.

Step 4 : Count the no. of data point in each category.

Step 5 : Assign the new data point, to that category which has maximum neighbors.

Step 6 : Model is ready.



Euclidean dist. $= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

b/w A and B.

Application : Banking System, Calculating Credit Ratings, Politics.

Pros :

- It is simple to implement.
- It is robust to noisy training data.
- More effective if training data is large.

Cons :

- Always need to determine K, which may be complex.
- The computation cost is high.

Date :

★ Decision Trees

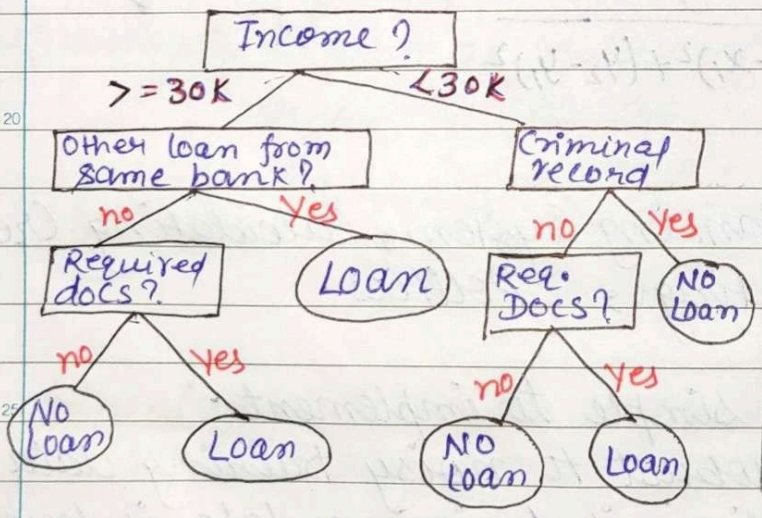
Also reffered as CART (Classification and Regression Tree).

- A decision tree is a flow-chart like structure in which each internal node represent a "test" on an attribute, each branch represent outcomes of the test, and each leafnode represent a class label (decision taken after computing all attributes).

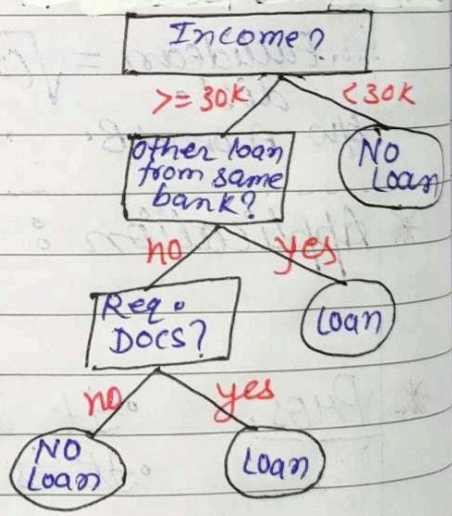
Pruning: It increases the performance of a tree by removing the branches that make use of features having low importance.

* Same can be used as an example of Decision Tree.

An unpruned Decision Tree



A pruned Decision Tree



* Types

1) Categorical Variable D.T.: D.T. which has categorical target variable. Eg. Yes or No types.
2) Continuous Variable D.T.: Decision Tree has continuous target variables.

Date :

* Assumptions:

- At the begning, the whole training dataset is considered as root.
- Feature values are preffered to be categorical.
- Records are distributed recursively on the basis of attribute value

Advantage of CART	Disadvantage of CART
i) Simple to understand, Interpret, Visualize.	→ It can create overcomplex tree.
ii) Can handle both categorical and numerical data.	→ It can cause overfitting.
iii) Non-linear relationship b/w parameters doesn't affect performance.	

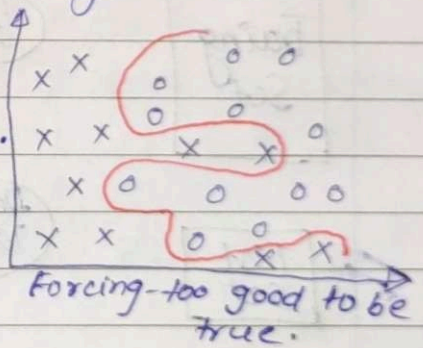
☆ **Overfitting**: Overfitting occurs when our ML model tries to cover all the data points or more than the required data points present in given dataset.

Sign: The model has high accuracy on training data, poor performance on Test data.

Techniques to reduce Overfitting:

i) Increase training data.

ii) Reduce mode complexity.



☆ **Underfitting**: Underfitting occurs when a model is too simple to capture underlying patterns in data.

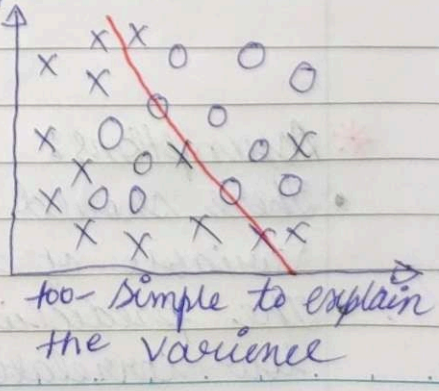
Sign: Model has low accuracy on both training and test data.

* Techniques to reduce Underfitting:

i) Increase model complexity.

ii) Increase no. of features.

iii) Remove noise from data.

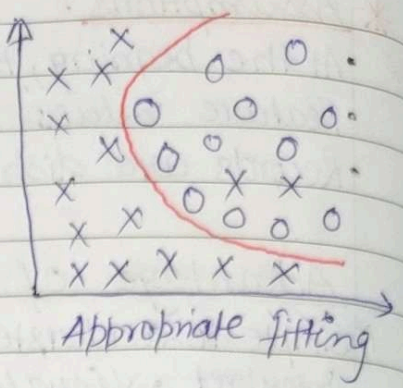


Date :

☆ Proper Fitting

It includes —

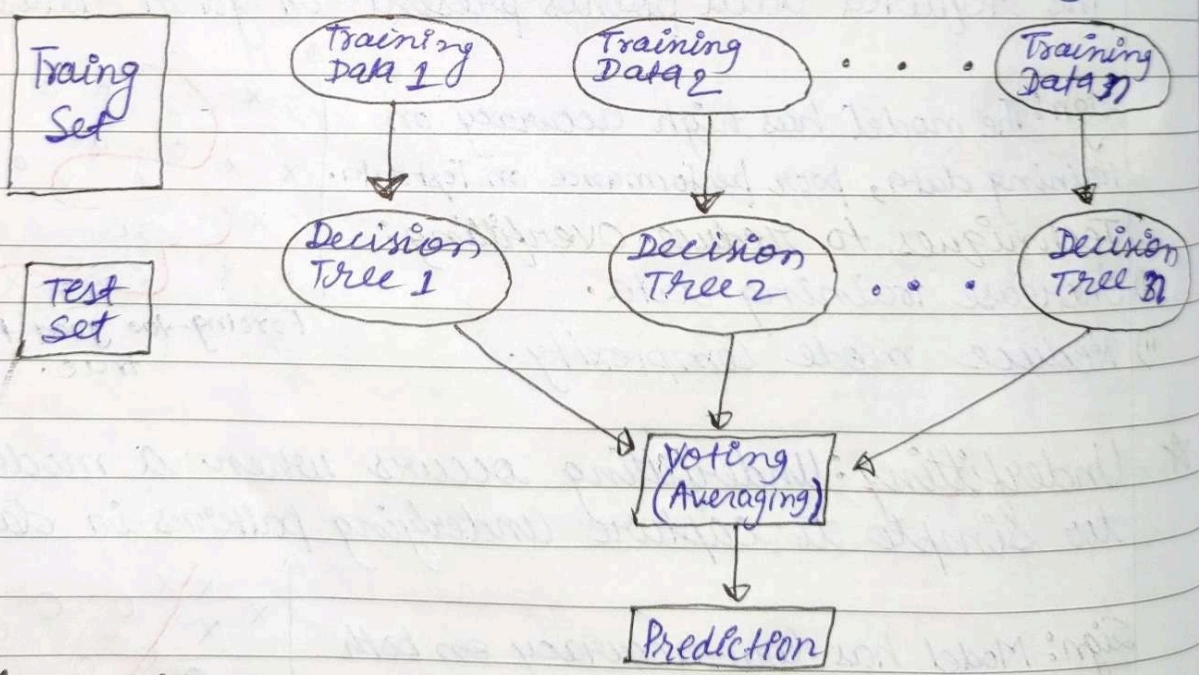
i) Seperate training and test data.
ii) Trying appropriate algorithms.
iii) Fitting model parameters.
iv) Proper performance metrics.



★ Random Forest

It is a combination of various decision trees which is used for both classification and Regression.

The greater number of trees in forest leads to higher accuracy and prevent overfitting.



✳ Assumptions:

- There should be some actual values in feature variable of dataset.
- The predictions from each tree must have low correlation.

Date :

* How Random Forest Work?

Step1 : Select random K datapoints from training set.
Step2 : Build decision tree associated with selected datapoints.
Step3 : Choose no. of for decision trees that you want to build.
Step4 : Repeate Step 1 and 2.
Step5 : for new datapoints, find the prediction of each decision tree, and put it in that wins majority votes.

* Applications

- Banking : For identifying loan risk.
- Medicine : disease trends & risk can be identified.
- Land Use : Can identify areas of similar use.
- Marketing : Marketing trends can be identified.

* Advantages :

i) Random forest is capable of both classification & Regression.
ii) It is capable of handling large datasets.
iii) It enhances the accuracy and prevents overfitting issues.

* Disadvantages :

It is not suitable for Regression tasks.

Date :

☆ Numericals:

1) On Linear Regression:

Ques:

X	2	4	6	8
Y	3	7	5	10

Sol :-

xy	x^2	x	y
6	4	2	3
28	16	4	7
30	36	6	5
80	64	8	10
Σ 144	120	20	25

$$m = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$n = 4$

$$= \frac{4 \times 144 - (20)(25)}{4 \times 120 - (20)^2} = \frac{576 - 500}{480 - 400} = \frac{76}{80} = 0.95$$

$$c = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{\Sigma x^2 - (\Sigma x)^2} = \frac{25 \times 120 - 20 \times 144}{4 \times 120 - 400} = 1.5$$

⇒ formula for Linear Regression is given by —

$$y = mx + c$$

$$\boxed{y = 0.95x + 1.5}$$

, for RSS $= \sum_{i=1}^{n} (y_i - (mx_i + c))^2$

Residual Sum of Squares $= 8.7$

2) On Baye's Rule —

Ques) A doctor knows that meningitis causes stiff neck 50% of time.

→ prior probablity of any patient having meningitis is 1/50000.

→ Prior probability of any patient having stiff neck is 1/20.

If a patient has stiff neck, what's probability He/She has meningitis?

Date :

Sol:- meningitis

$$P(M/S) = \frac{P(S/M)\,P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20}$$

Stiff neck

$$= 0.002 \quad \text{Ans}$$

Q2.) Naive Bayesian classifier:

Weather	Football
Rainy(R)	Yes
Sunny(S)	Y
Overcast (O)	Y
O	Y
S	No
R	Y
S	Y
O	Y
R	N
S	N
S	Y
R	N
O	Y
O	Y

Sol:

Weather	Y	N	Total	Prob.
Overcast	5	0	5	5/14
Rainy	2	2	4	4/14
Sunny	3	2	5	5/14
Total	10	4	14	
Prob.	10/14	4/14		

$$\Rightarrow P(Y/\text{Sunny}) = \frac{P(S/Y)\cdot P(Y)}{P(\text{Sunny})} = \frac{(3/10)\cdot(10/14)}{(5/14)} = 0.6$$

$$\Rightarrow P(\text{No}/\text{Sunny}) = \frac{P(S/N)\cdot P(N)}{P(S)} = \frac{(2/4)\cdot(4/14)}{(5/14)} = 0.4$$

Ques:
Weather → S
Football → Y/N?

$$\Rightarrow P(Y/O) = \frac{P(O/Y)\cdot P(Y)}{P(O)} = \frac{(5/10)\cdot(10/14)}{(5/14)} = 0.1$$

Date :

* On KNN - Classification:

	Math	CS	Result
S_1	4	3	fail
S_2	6	7	Pass
S_3	7	8	pass
S_4	8	8	pass
S_5	5	5	fail

Query ⇒
x (Math = 6, CS = 8)
whether it will Pass or fail?
for $\boxed{K=3}$

Sol:-
Euclidean distance:-
$d = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$
Let → ∄ $x_1 = 6, y_1 = 8$ (new data)

$S_1 : \sqrt{(6-4)^2+(8-3)^2} = 5.38$
$S_2 : \sqrt{(6-6)^2+(8-7)^2} = 1$
$S_3 : \sqrt{(6-7)^2+(8-8)^2} = 1$
$S_4 : \sqrt{(6-8)^2+(8-8)^2} = 2$
$S_5 : \sqrt{(6-5)^2+(8-5)^2} = 3.16$

New data is near to S_2, S_3, S_4 then 3(K) neighbour.
↓ ↓ ↓
Pass Pass pass = 3P ∵ 3 > 0
fail = 0 ∴ x (new data) is declared Pass.

* On Confusion Matrix :

	Predicted NO	Predicted Yes
Actual NO	TN = 45	FP = 5
Actual Yes	FN = 5	TP = 95

(I) Accuracy = $\frac{TN+TP}{Total\ Pred.} = \frac{45+95}{150}$
$= 93.33\%$.

(II) True Positive rate or Recall or Sensitivity $= \frac{TP}{FN+TP} = \frac{95}{100} = 95\%$.

Accuracy, Recall, Precison?
....

(III) False Positive Rate (FPR) $= \frac{FP}{Actual\ NO\ (TN+FP)} = \frac{5}{50} = 10\%$.

(IV) True Negative rate or Specificity $= \frac{TN}{Actual\ NO} = \frac{45}{50} = 90\%$.