Experiment No. 1

Aim: Select appropriate dataset & perform the following preprocessing steps

1) Imputation of is the process of replacing missing values in the dataset there are several techniques to handle missing values
   a) Mean/Median/mode: Imputation replace missing values with the mean, median, mode of the respective column.
   b) Forward and Backward fill: Fill missing values using the previous or next observation.
   c) Interpolation use linear or polynomial interpolation to estimate missing values.
   d) K-nearest: Neighbours Use the KNN algorithm to impute missing values based on the values of their nearest neighbours

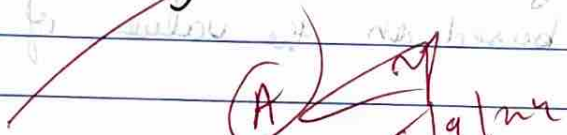
2) Anomaly Detection: It involves identifying or abnormal data points in the datasets.
   a) Statistical Methods
   b) Machine learning models
   c) time series methods

3) Standardization :- Transforms data to have a mean of standard deviation of one useful for algorithms that assume normally distributed data

4] Normalization :- Transform data to have a mean of zero & standard deviations of one useful for algorithms that assume normally, which is useful for algorithms that rely on dist measurement.

5] Encoding : used to convert categorical data into numerical format

a) Label encoding
b) One-hot encoding
c) Ordinal encoding
d) Freqency encoding



Experiment No. 2

Aim: Implement Gradient Descent Algorithm to minimize give obj function

Theory: Gradient Descent is an optimization algorithm used to minimize a given objective function particularly in machine learning (optimization problems) it is an iterative approach that seeks to find the minimum of a function by updating parameters in the direction

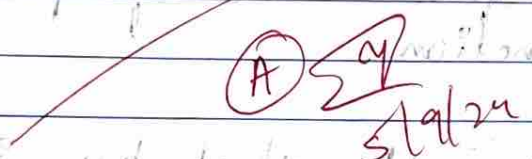
1] Objective function:-
- the Objective function, also known as the cost function or loss function measures how good our had a set of parenthesis is the goal of descent is to find the set of parameters that minimizes this function

2] Initialization :- You start by initializing the parameters of the model with some initial values

3] Gradient Calculation:- Calculate the gradient of the objective function with some initial values the gradient is a vector that points in a direction of the increases of the function. In the context of optimization you need to move in the opposite direction to really the minimum.

4) Update parameter : Update the parameters by moving them a small step in the direction of the negative gradient this step in the direction of the negative gradient this step size is determined by a parameter called the learning

5) Iteration : Repeat step 3 & 4 until a stopping criterion is met. The criterion would be maximum number of iterations the change in the objective function very small or some other measure indicating convergence.



Experiment No. 3

Aim:- Implement Simple Linear Regression

Theory:-

Linear Regression model studies the relationship between a single dependent variable 'Y' & one or more independent variable 'X' using a best fit straight line function. A line can be characterised by the slope intercept with one of the axis. Here a given data point goal is to find a line which indicates finding slope & intercept

$$Y = B_0 + B_1 X$$

Where ϵ - random error associated with data point so we assume underlying function from which the data is generated $Y_0 = B_0 + B_1 X + \epsilon$. For given data point, goal is to find out B_0 & B_1

$$Y = B_0 + B_1 x_1 + \epsilon$$

where B_0 = population of y - intercept
B_1 = Population slope
ϵ = random error

eg

height	weight	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
191	6.3	2.8	-2.3	-6.4	7.84
144	81	20.2	20.2	317.14	407.04
138	56	-15.8	-15.8	146.96	249.67
186	91	32.2	32.2	827	1036.8
128	47	-25.8	-28.8	472	656.6
136	57	-17.8	-17.8	-147	116
149	76	25.2	25.2	61	84.64
163	72	9.2	9.2	84	8464
152	62	-1.8	-1.8	44	3.24
131	48	-22.8	-17.8	264	3127.6
153	65				

formula $b_1 = \frac{2649.6}{3927.6}$

$= 0.6746$

(3) $b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \bar{y} - b_1 \bar{x}$

$= 65 - 3(0.6749 \times 153.8)$

$= -38.4551$

$= 0.6742 - 38.4551$

$y = 0.6742 - 38455$

A+

Experiment No. 4

Aim :- Implement Logistic Regression using machine learning method without using sklearn

Theory :- Logistic Regression estimates the probability of an event occurring, such as 'voted or didn't vote', based on a given dataset of independent variables. Since the outcome is a prob the dependent variable is bounded between 0 & 1.

1] Binomial :- In this there can be 3 or more possible unordered types of the dependent variables such as 0 or 1.

2] Multinomial : In this there can be 3 or more possible unordered type of the dependent variables. such as 'cat, dogs, sheeps, etc'.

3] Ordinal: In ordinal logistic regression there can be 3 or more possible ordered types of dependent variable such as 'low, medium or higher'.

1] Independent Observations : Each observations is depend of the others meaning there is no correlation btw any input variables.

2] Binary dependent variable It takes the assumption that the dependent variable must be binary or dichotomous

3) Large sample size :- the sample size is large.

4) Linearity relationship b/w independent variables & log odds:
Let the independent input matrix

$$x = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

& the dependent variable 'y' having only binary value:

$y = 0$ if class 1
$= 1$ if class 2

Probability $P(y=1) = \sigma(z)$
$P(y=0) = 1 - \sigma(z)$

∴ final logistic regression

$$P(x, b, w) = \frac{1}{1+e^{wx+b}}$$

(A) Sigmoid

Experiment No. 5

Aim:- Implement Decision Tree classification Algorithm

Theory: A decision tree is a tree like structure that represents a set of decisions and their possible consequences. Each node in the tree represents a decision. The leaves of the tree represent the final decisions. This leaves of the tree represent the final decisions.

Consider the dataset

Age	Competition	Types	Profit
old	yes	software	down
old	yes	software	down
old	No	Hardware	down
Mid	yes	software	down
Mid	yes	Hardware	down
Mid	No	Hardware	up
Mid	No	software	up
New	Yes	software	up
New	No	Hardware	up
New	No	software	up

$$\text{gini}(T) = \left[1 - \left[\left(\frac{5}{10}\right)^2 + \left(\frac{5}{10}\right)^2\right]\right] = 0.5$$

Age:

split = $\frac{3}{10}$ gini old + $\frac{4}{10}$ gini (mid) + $\frac{3}{10}$ gini (new)

$= \frac{3}{10}\left[1 - \left[\left(\frac{0}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]\right] + \frac{4}{10}\left[1 - \left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right) + \frac{3}{10}\left[1 - \left(\frac{3}{3}\right)\right.$

$= \underline{0.2}$

* Competition

split = $\frac{4}{10}$ gini yes + $\frac{6}{10}$ gini no

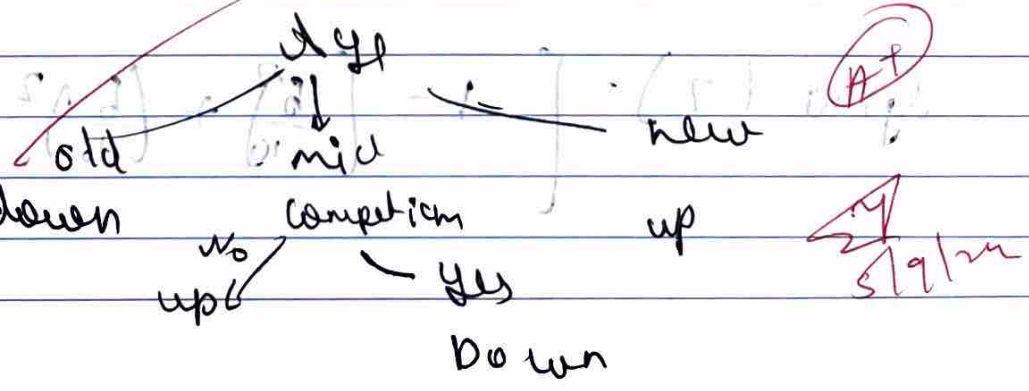
$= 0.42$

* type

split = $\frac{6}{10}$ gini (software) + $\frac{4}{10}$ gini (hardware)

$= \underline{0.5}$

∴ Split values of age is the smallest age - root node



Experiment 6

Aim: Implement SVM for classification.

Theory: SVM is a powerful machine learning algorithm used for linear or non linear classification regression and even outlier detection tasks. SVM can be used for a variety of task, such as text, such classification image classification, spam detection, handwriting identification gene extension analysis, face detection tasks. SVMs are adapted & efficient in a variety, efficient in a variety of application because they can manage high-dimensional data & non linear relationships. SVM algorithms are very effective as we try to find the maximum seperating hyperplane between the different classes accesible in the target feature

SVM terminology :-

Hyperplane : Hyperplane is the decision boundary that is used to seperate the data points of different classes in a feature space.

Support Vectors :- Support vectors are the closest data points to the hyperplane, which makes a critical role in deciding the hyperplane & margin

Margin :- Margin is the distance b/w the support vector & hyperplane. Hard margin : the hard margin hyperplane is a hyperplane that properly seperates the data points of

of different categories without any mis calculat

soft margin: When the data is not perfectly separable or contains outliers, SVM permits a soft margin technique. Each data point has a slack variable introduced by the soft margin SVM formulation, which softens the hard margin sqeuen & permits certain misclassification.

Conclusion:

We successfully implemented SVM for classificatio

A

11/10/24

Experiment 7

Aim :- Implement ensemble methods to combine different model.

Theory :

Ensemble methods are powerful techniques in ML that combine the predictions of multiple individual models to improve overall prediction performance.

Advantages :

1. Reduces Overfitting
2. Increases Model Robustness
3. Enhance prediction accuracy

1. Voting ensemble :

Hard voting :- Multiple base models make predictions & final decision is the one with majority vote.

Soft voting : Similar to hard voting, but instead of selecting the class with most votes, it calculates weighted avg of class probabilities predicted by the base models.

2. Bagging :- Bagging aims to reduce the variance of an individual base model by training multiple base models on random subsets of the training data with replacement. The final prediction is often made by averaging or majority voting

(3) Boosting :-

They focus on improving the accuracy of weak base models sequentially. Each base model is trained to correct the mistakes made by the previous model

(4) Stacking :-

It combines multiple base models by training a meta model on their predictions. The base models output serve as features for the meta model It is known for its high predictive performance

(5) Ensemble pruning :-

These techniques aim to reduce the complexity of an ensemble by selecting a subset of base models that contribute the most to the ensembles performance Reduces computational overhead

Other methods : Stacking with cross validation
weighted Average ensembles
adaptive ensembles

Conclusion :
Ensemble method implemented

14/10/24

Experiment 8

Aim :- Implement Principal Component Analysis technique

Theory :-

PCA is a dimensionality reduction technique commonly used to transform high-dimensional data into a lower dimensional representation while preserving the most important information

PCA . this by finding orthogonal axes along which the data exhibits the maximum variance

These principal components are linear combination of the original features, allowing you to reduce the dimensionality of the data

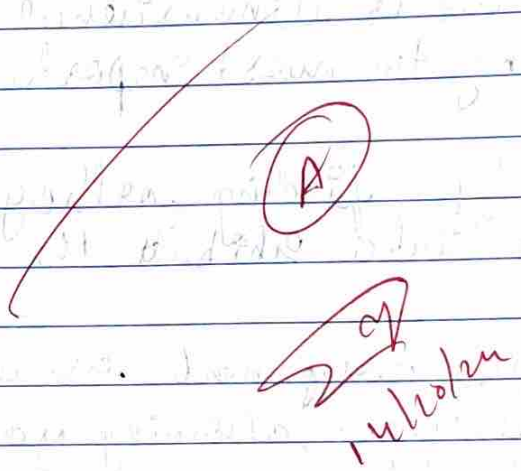
The steps & concepts of PCA are :-

1. Centering the data : By subtracting the means of each feature from the corresponding data point
2. Covariance matrix :
3. Eigen value decomposition
4. Eigen value & Eigen vectors

(5) Selecting principal components

(6) Projection :

Conclusion: Principal component Analysis implemented

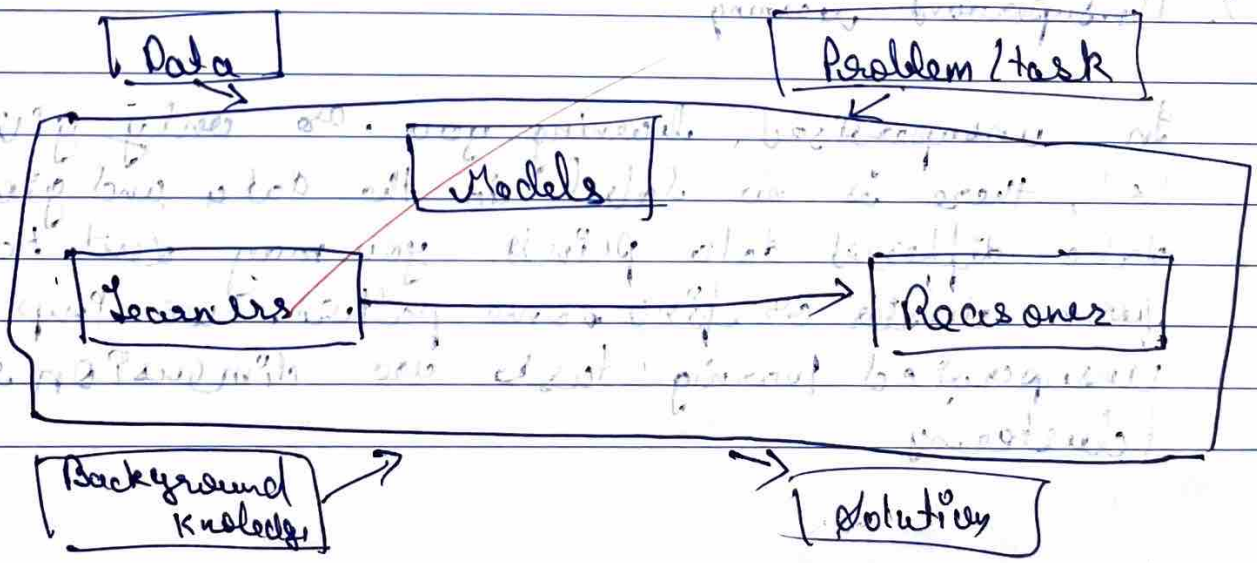


Assignment 1

Q1] What is learning ? Explain different types of learning with example.

⇒ Machine learning is a category of Artificial Intelligence. In Machine learning computer have the ability to learn themselves. Explicit programming is not required. Machine learning focuses on the study and development of algorithm that can learn themselves, explicit programming is not required. Machine learning focuses on the study and development of algorithm that can learn from data and also make predictions on data

Machine learning mainly focuses on the design and development of computer programs that can teach themselves to grow and change when exposed to new data. Using machine learning we can collect information from a dataset by asking the computer to make some sense from data. M L is turning data into information

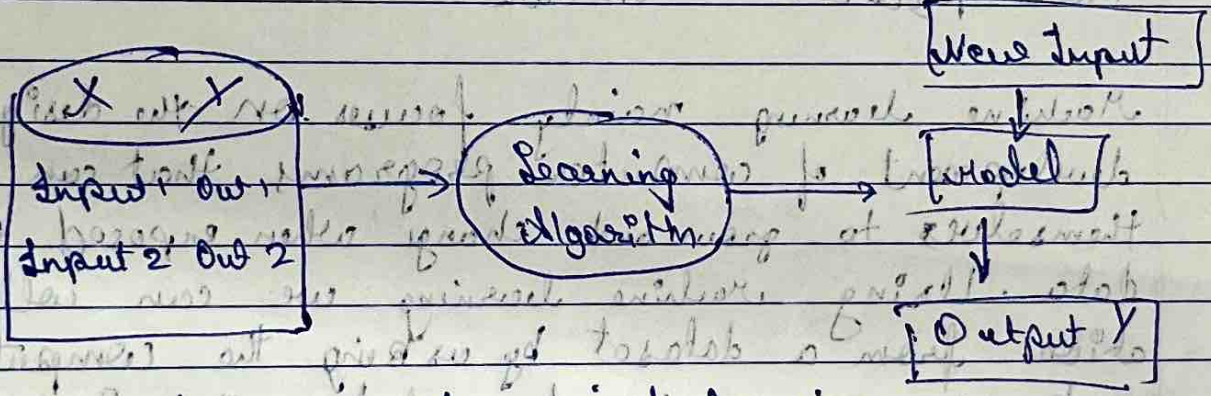


types of machine learning are:

1. Supervised learning

In this type of learning we use data which is comprises of input and corresponding output. In supervised learning training data is labelled with the correct answers eg. 'spam' or 'ham'

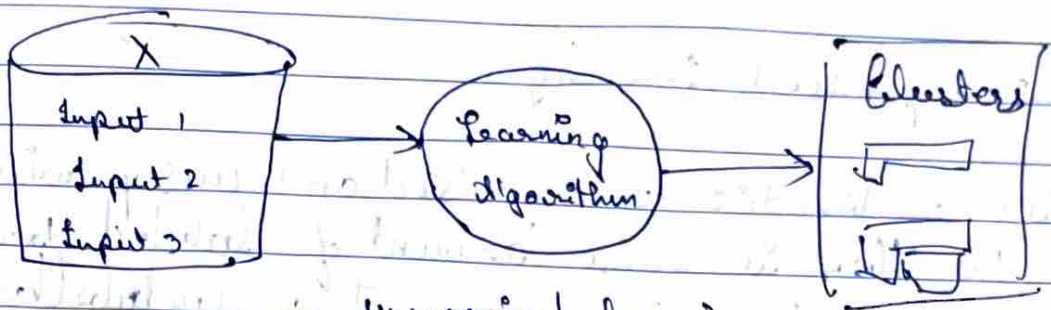
two most important types of supervised learning are classification (where the output are discrete labels, as in spam filtering) and regression (where the output are real valued).



→ Supervised learning

2. Unsupervised learning

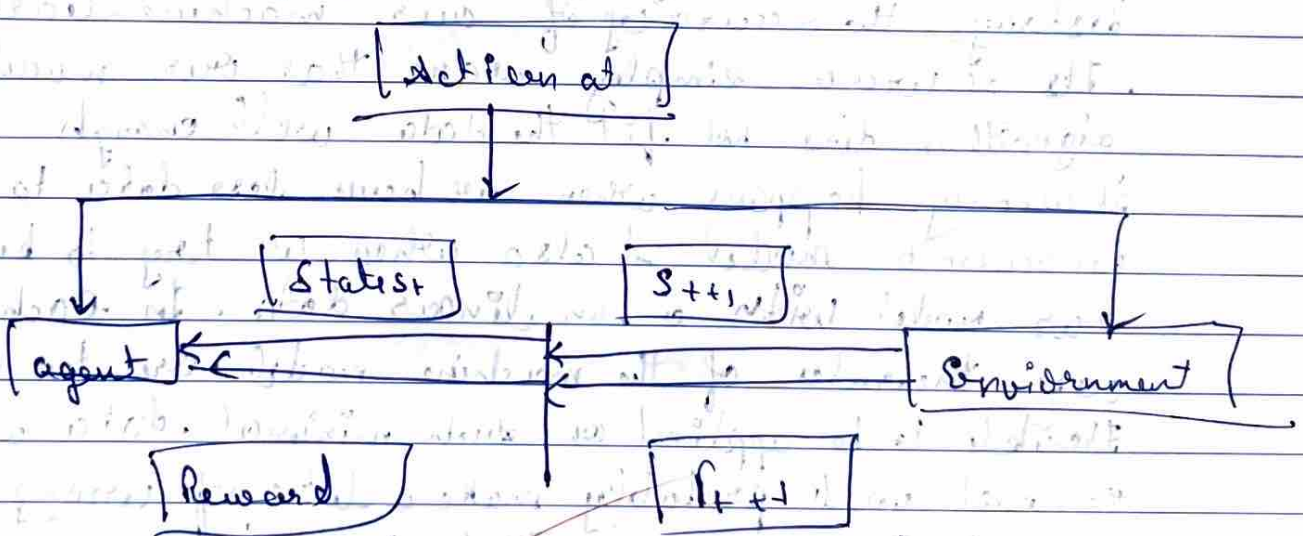
In unsupervised learning you are only given input 'x', there is no label to the data and given the data different data points, you may want to form cluster or find some pattern. Two imp unsupervised learning tasks are dimension reduction & clustering



Unsupervised learning

3. Reinforcement Learning

In reinforcement learning you have an agent who is acting in an enviornment and you want to find out what action the ajent must take based on the reward or penalty that the agent gets. In this an agent, Example a robot or controller, learn to learn the optimal actions based on the outcomes of past actions



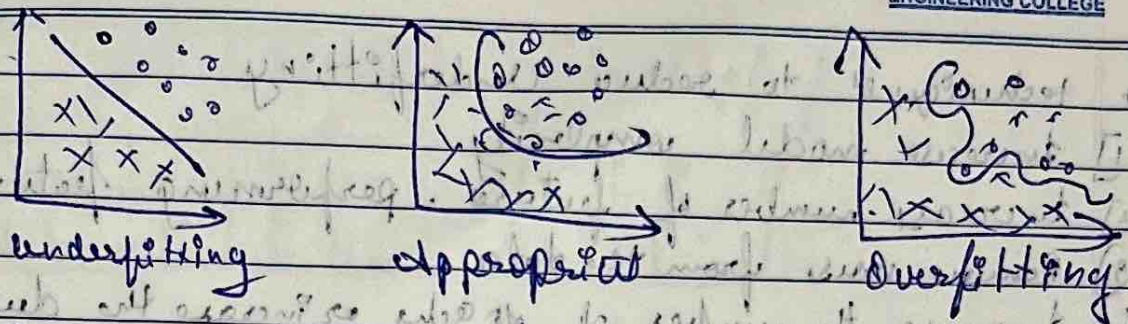
Reinforcement learning

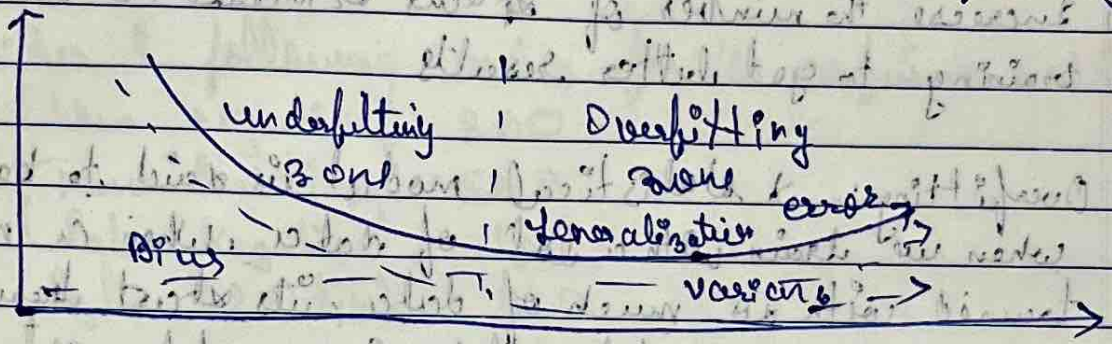
4) Semi-Supervised Learning :

It is a combination of supervised and unsupervised learning. In this there is some amount of labeled train data and also you have large amount of unlabelled data and you try to come up with some learning algorithm that converts or even when training data is not labelled eg: text classification image classification, where small amount of labelled

Q2) Define Overfitting & Underfitting.

⇒ Underfitting : A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough. It usually happens when we have less data to build an accurate model & also when we try to build a linear model with a non-linear data. In each such case the rules of the machine model are too easy & flexible to be applied on such minimal data and therefore the model will probably make a lot of wrong predictions. Underfitting can be avoided by using more data & also reducing the features by feature selection. Underfitting can be refined as high bias and low var





a7] Illustrate process of learning with the gradient descent for a univariate linear regression using a bell shaped error function. Explain how stepping is modulated on every iteration

⇒ Gradient descent is an iterative optimization algorithm that tries to find the optimum value (minimal maximum) of an objective function. It is one of the most used optimization techniques in machine learning projects for updating the parameters of a model in order to minimize a cost function.

The main aim of gradient descent is to find the best parameter of a model which gives the highest accuracy of training as well as testing dataset. In gradient descent the gradient is a vector that points in the direction of the steepest increase of the function at a specific point. Moving in the opposite direction of the gradient allows the algorithm to gradually descend towards lower values of the function and eventually reaching to the minimum of the function.

* techniques to reduce Underfitting

1) Increase model complexity
2) Increase number of features, performing feature engineering
3) Remove noise from the data.
4) Increase the number of epochs or increase the duration of training to get better results

Overfitting :- A statistical model is said to be overfitted when we train with a lot of data, when a model gets trained with so much of data, its start learning from the noise and inaccurate data entries in our data set. Then the model does not categorize the data correctly because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the data set and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like maximal depth if we are using decision trees.
Overfitting It can be defined as begin and low bias

techniques to reduce overfitting.

1. Increase training data.
2) Reduce model complexity
3) Early stoping during the training phase
4) Ridge Regularization and Lasso
5) Use dropout for neural network to tackle over fitting

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Consider the following values for the confusion matrix
TN (True negative) = 300
TP (True Positive) = 500
FN (False Negative) = 150
FP (False Positive) = 50

a] Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is defined as the ratio of the number of correct predictions and the total number of predictions. It lies between (0, 1). higher accuracy means a better model (TP & TN). higher accuracy means better model. (TP & TN) must be high

Accuracy is not a useful matrix in case of an imbalanced datasets, eg 960 patients without cancer & 40 patient with cancer

Accuracy using above values : $\frac{(500+300)}{(500+50+150+300)} = \frac{800}{1000}$ %

- Precision

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

Precision is useful when we want to reduce the number positives. Eg. a system that predicts if the email received is spam or not. taking spam as a positive, then we not want our system to predict non spam emails

Precision using above values

$$= \frac{500}{500+50} = \frac{500}{550} = 90.90\%$$

3] Recall

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

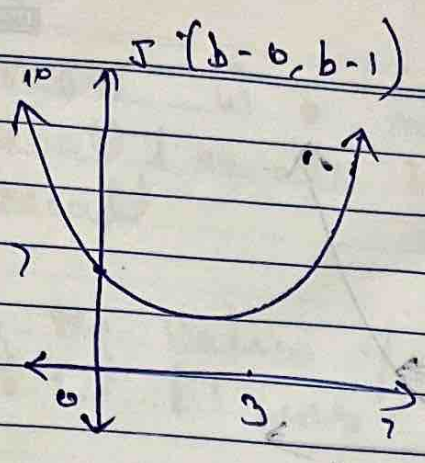
Recall is useful in case of cancer detection, where we want to minimize the number of false negative for any practical use since we dont want to mark person with cancer as safe.

Recall using above values

$$= \frac{500}{500+150} = \frac{500}{650} = 76.92\%$$

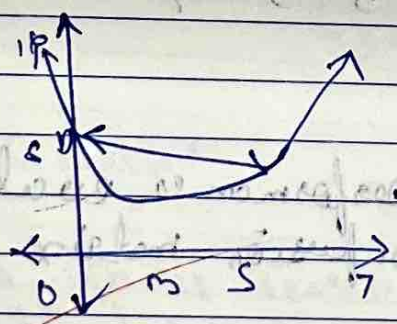
4] F-1 score

$$f1 = 2 \times \frac{\text{Precision} \cdot \text{recall}}{\text{Precision} + \text{Recall}}$$

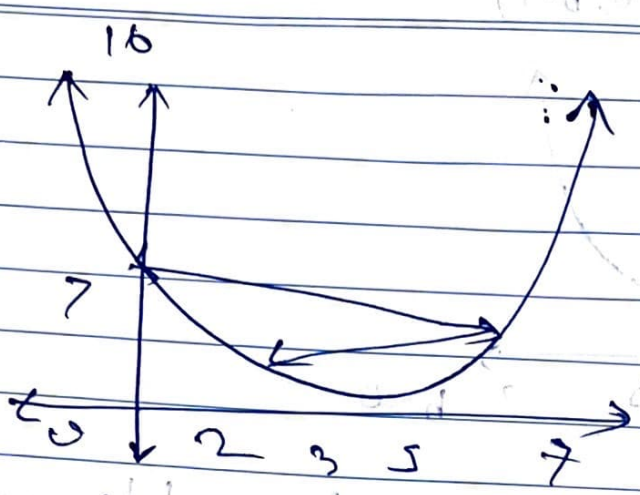


As we can see we have a simple parabola with a minimum at b-0 = 3 the model will use the gradient descent algorithm to iteratively tune the values of b_0 & b_1
to start this process, our model will initialize all parameters to 0 this means our initial cost is 7

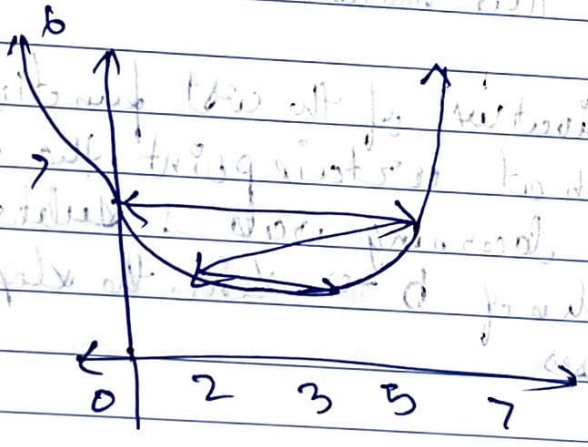
It takes the derivative of the cost function & returns the slope of the graph at certain point. the model multiplies the slope by the learning rate & subtracts the product from current value of b-0. Since the slope is negative, the value of b_0 increases



the value of b-0 becomes the algorithm will now take the derative of cost function. when b-0 is 5 & carry out the process above



this process will continue until the derivative returns a slope that is very close to 0, signifying that model has reached a near optimal value for the parameter being tuned.



Q4) Explain the following performance evaluation parameters with the help of confusion matrix. Illustrate with appropriate example

⇒ A confusion matrix is a table that is often used to describe the performance of a classification model (or classifier) on a set of test data for which the true values are known

F1 score is a metrics that combines precision and recall & equals to the harmonic mean of precision and recall

Using the values of precision & Recall (0.9040 & 0.7692) F1 score = 0.8333 = 83.33%.

e] Specificity

Specificity is defined as the ratio of true Negatives & True negatives + False Positives. We want the value of specificity to be high

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative + False Positive}}$$

Thus specificity using above values will be

$$300/(300+50) = \frac{300}{350} = 85.71\%$$

f] AUC - ROC curve

AUC (Area Under Curve) - ROC (Receiver Operating Characteristic) curve is one of the most important evaluation metrices for checking any classification models performance.

It is a plotted between FPR (X-axis) & TPR (Y-axis) if the value is less than 0.5 than the model is even worse than random

guessing model

$$TPR = \frac{TP}{FP+FN}$$

$$FR = \frac{FP}{FP+TN}$$

THADOMAL SHAHANI TSEC ENGINEERING COLLEGE

Assignment 2

Q) Discuss different ensemble learning techniques

=> Ensemble techniques combine individual models together, to improve the stability and predictive power of the model.

Individual models, also known as base models can have a high bias, high variance or both, which in turn is not good. Such models are known as weak models. When weak models are combined they can produce strong models having low bias & low variance.

A diverse set of models in comparison to single models can make better predictions. This diversification in ML is achieved by a technique called ensemble learning.

Ensemble methods → required

Some ensemble learning techniques are

a) Max Voting: This method is generally used for classification problems. In this technique multiple models are used to make prediction by each model are considered as a vote. The prediction we get from the majority of the models is used as a final prediction.

b) Averaging: Multiple predictions are made for each data point and the average of all predictions is the final predictions.

c) Weighted averaging: This is an extension of

of averaging where all models are assigned weights

Boosting :- It trains multiple models sequentially with each new model focusing on the mistakes made by the previous model. The idea is to correct the errors earlier models. The idea is to correct the errors earlier models & improve the overall prediction accuracy. The final model is usually a weighted average of all the models is usually weighted average of all the models predictions. Boosting can significantly improve model performance and reduce both bias & variance. It is particularly useful for complex datasets

Q2) Explain the following
a) Weak learner & strong learner
→ A weak learner is a model that performs slightly better than random guessing. It is typically a simple model with limited predictive power, often making predictions that are only marginally better than random chance. In decision tree algorithms, a shallow tree with a few nodes and a simple rule-based classifier can be considered a weak learner & strong learner is a model that performs significantly better than random guessing. It has a high predictive accuracy and it can capture complex patterns in data
eg :- Complex models like deep neural networks

strong learner are often the good ensemble methods. they are built by combining weaker models or using sophisticated algo to achieve high prediction performance

b) Meta learning :
It refers to the process of learning how to learn. It involves developing models that can learn from various learning algorithms or datasets and adapt to new tasks more efficiently. The idea is to improve the learning process itself by using prior experience or data. eg Meta learning can involve learning the best hyperparameters for different algorithms selecting appropriate models for various tasks or even adapting to new types of data with additional training

c) Random forest :
It is an ensemble learning method specifically designed for classification & regression tasks. It builds a number of decision trees & combines their predictions to make a final decision tree
Working Bagging - Random forest uses the bagging technique to train each decision tree on a different random subset of training data sampling. During the training of each decision tree for regression tasks, the final prediction is the average of the prediction of all trees

Q3] Comp are Bagging, Boosting & Stacking

	Bagging	Boosting	Stacking
1]	Parallel training (independent models)	Sequential training (Correct errors)	Sequential training combines predictions
2]	Average or majority voting	Weighted average based on models	Meta learner combines base model predictions
3]	Reduces variance	Reduces both bias and variance	Reduces depending on models
4]	Less prone to overfitting	It prone to everything	Less prone to overfitting
5]	Less computational	can be computational	can be computational
6]	Simple to implement and understand	More complex sequential train	Complex, involves multiple model types
7]	Eg:- Random forest	eg:- Adaboost	Eg:- model stacking with decision tree, logistic regression

Q4) Explain Adaboost algorithm

⇒ Adaboost is a boosting algorithm that also works on the principle of stage wise addition method where multiple weak learner models are used to get strong learner models. The value of alpha parameters in this case will be indirectly proportional to the errors of weak learner

- Working:

1) Initialization :- All training samples are given equal weight.
2) Training the Algorithm :- Iteratively trains weak learners
3) Weighting : the influence of each weak learner is determined by its accuracy
4) Updating :- Weights of misclassified samples are increased for the next iteration
5) Final model : the final prediction is a weighted vote of all weak learners

- This algorithm enhances performance by focusing on outliers & it is generally more resistant to overfitting than other models

- Advantages
  - simple & effective
  - no overfitting
  - can be used with various types of weak learners
- Disadvantages
  - sensitive to noisy data
  - It is computationally expensive

14/10/24