21/04/2025

Training Day-46

Decision Tree in Machine Learning:-

A decision tree in machine learning is a versatile, interpretable algorithm usedfor predictive modelling. It structures decisions based on input data, making it suitable for both classification and regression tasks. This article delves into thecomponents, terminologies, construction, and advantages of decision trees, exploring their applications and learning algorithms.

Decision Tree in Machine Learning:-

A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction. The decision tree algorithm falls under the category of supervised learning. They can be used to solve both **regression** and **classification problems**.

Decision Tree Terminologies:-

There are specialized terms associated with decision trees that denote variouscomponents and facets of the tree structure and decision-making procedure. :

- **Root Node:** A decision tree's root node, which represents the originalchoice or feature from which the tree branches, is the highest node.
- **Internal Nodes (Decision Nodes)**: Nodes in the tree whose choices are determined by the values of particular attributes. There are branches on these nodes that go to other nodes.
- **Leaf Nodes (Terminal Nodes)**: The branches' termini, when choices orforecasts are decided upon. There are no more branches on leaf nodes.
- **Branches (Edges)**: Links between nodes that show how decisions aremade in response to particular circumstances.
- **Splitting**: The process of dividing a node into two or more sub-nodesbased on a decision criterion. It involves selecting a feature and a threshold to create subsets of data.
- **Parent Node**: A node that is split into child nodes. The original nodefrom which a split originates.
- **Child Node**: Nodes created as a result of a split from a parent node.

- **Decision Criterion**: The rule or condition used to determine how the data should be split at a decision node. It involves comparing feature values against a threshold.
- **Pruning**: The process of removing branches or nodes from a decision tree to improve its generalization and prevent overfitting.

22/04/2025

Training Day-47

fit(), transform() and fit_transform() Methods in Python:-

It's safe to say that scikit-learn, sometimes known as sklearn, is one of Python's mostinfluential and popular Machine Learning packages. It includes a complete collection of algorithms and modeling techniques that are ready to be trained, including utilities for pre-processing, training, and grading models.

One of the most elementary classes in Sklearn is the transformer, which implementsthree different methods: fit(), transform(),and fit transform().

- **fit() Method:-**
  In the fit() method, we apply the necessary formula to the feature of the input data we want to change and compute the result before fitting the resultto the transformer. We must use the .fit() method after the transformer
  object.
  If the StandardScaler object sc is created, then applying the .fit() method willcalculate the mean (μ) and the standard deviation (σ) of the particular feature
  F. We can use these parameters later for analysis.
  Let's use the pre-processing transformer known as StandardScaler as an
  example and assume that we have to scale the features of self-created data. The example dataset in the code below is created using the arrange method and then divided into the training and testing datasets. After that, we create aStandardScaler instance and fit the feature of the training data to it to
  determine the mean and standard deviation to be utilized for scaling in thefuture.
  The significance of separating the dataset into the train and test datasetsbefore using any pre-processing process, such as scaling, must be
  emphasized. Test data points represent real-world data. Therefore, we mustonly execute fit() to the training feature to prevent future data to our model.

- **transform() Method:-**
  To change the data, we most likely use the transform() function, where weperform the calculations from fit() to each value in feature F. We transform

the fit computations. Hence we must use .transform() after we have appliedthe fit object.

When we make an object using the fit method, we utilize the example fromthe section above and place the object in front of the.

The scale of the data points is transformed using the transform and fit transform method, and the output we receive is always a sparse matrix orarray.

- **fit_transform() Method:-**

  The training data is scaled, and its scaling parameters are determined by applying a fit_transform() to the training data. The model we created, in thiscase, will discover the mean and variance of the characteristics in the training set.

  The mean and variance of every feature reported in our data are calculatedusing the fit approach. The transform method transforms all features usingthe corresponding means and variances.

  We wish scaling to be implemented in our testing data, but we also don'twant our model to be biased. We expect our test set of data to be entirely fresh and unexpected for our model. In this situation, the convert approach isuseful.

23/04/2025

Training Day-48

Data Preprocessing in Python:-

In order to derive knowledge and insights from data, the area of data science integrates statistical analysis, machine learning, and computer
programming. It entails gathering, purifying, and converting unstructured data intoa form that can be analysed and visualised. Data scientists process and analyse datausing a number of methods and tools, such as statistical models, machine learning algorithms, and data visualisation software. Data science seeks to uncover patterns in data that can help with decision-making, process improvement, and the creation of new opportunities. Business, engineering, and the social sciences are all
included in this interdisciplinary field.

Data Preprocessing:-

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data preprocessing is a technique that is used to convert the rawdata into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Steps in Data Preprocessing:-

Step 1: Import the necessary libraries . Step

2: Load the dataset

Step 3: Statistical AnalysisStep

4: Check the outliers:

Step 5: Correlation

Step 6: Separate independent features and Target VariablesStep 7:

Normalization or Standardization

24/04/2025

Training Day-49

R-squared in Regression Analysis in Machine Learning:-

The most important thing we do after making any model is evaluating the model. We have different evaluation matrices for evaluating the model. However, the choice of evaluation matrix to use for evaluating the model depends upon the typeof problem we are solving whether it's a regression, classification, or any other type of problem. In this article, we will explain R-Square for regression analysis problems.

What is R-Squared?

R-squared is a statistical measure that represents the goodness of fit of a regressionmodel. The value of R-square lies between 0 to 1. Where we get R-square equals 1when the model perfectly fits the data and there is no difference between the predicted value and actual value. However, we get R-square equals 0 when the model does not predict any variability in the model and it does not learn any relationship between the dependent and independent variables.

We calculate R-Square in the following steps:-

1. First, calculate the mean of the target/dependent variable y and wedenote it by $\bar{y}$
2. Calculate the total sum of squares by subtracting each observation y_i from $\bar{y}$, then squaring it and summing these square differences acrossall the values. It is denoted by

   $$SStot = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
3. We estimate the model parameter using a suitable regression model suchas Linear Regression or SVM Regressor
4. We calculate the Sum of squares due to regression which is denoted by SSR. This is calculated by subtracting each predicted value of y denotedby y_pred_i from y_i squaring these differences and then summing all the nterms. SSR = \sum_{i=1}^{n} (\hat{ypred}_i – \bar{y})^2
5. We calculate the sum of squares *(SS_{res})*. It explains unaccounted variability in the dependent y after predicting these values from an

independent variable in the model. SSres = \sum_{i=1}^n (y_i – {ypred}_i)^2

6. we can then use either

$$R^2 = \frac{SSR}{SStot} or R^2 = 1 - \frac{SSres}{SStot}$$

25/04/2025

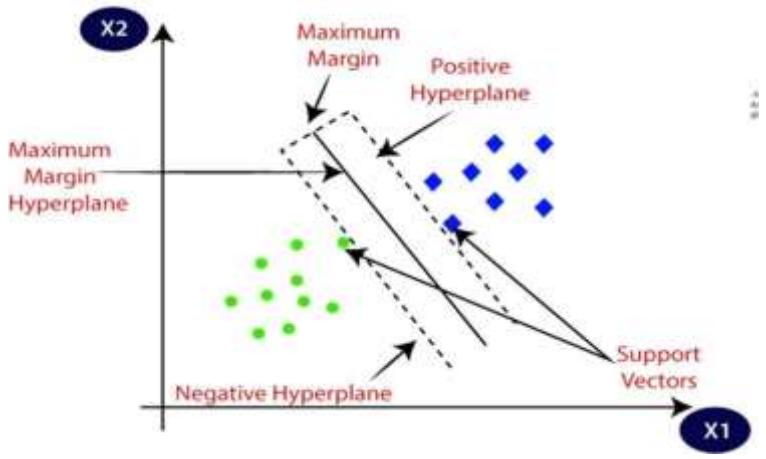
Training Day-50

Support Vector Machine (SVM) Algorithm:-

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new datapoint in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. Theseextreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Types of SVM:-

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which meansif a dataset can be classified into two classes by using a single straight line,

then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non- linear SVM classifier.