Assignment 1

Title :- Uber Price Prediction

Problem Statement :- Predict the price of the Uber ride from a given pickup point of the agreed drop-off location. Perform following tasks :-

1. Pre-Process the dataset.
2. Identify outliers
3. Check the correlation
4. Implement linear regression and random forest regression models.
5. Evaluate the models and compare their respective scores like R2, RMSE etc.

Objective :- To develop uber price prediction system.

Theory

The Assignment is about on world's largest taxi company Uber inc. In this assignment, we're looking to predict the fare for their future transactional cases. Uber delivers service to lakhs of customers daily. Now, it becomes really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

Data Preprocessing :

Data Preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing data.
- Encoding Categorical Data.
- Splitting dataset into training and test set
- Feature Scaling.

Outliers :

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution and arise due to inconsistent data entry, or erroneous observation. To ensure that the trained model generalizes well to the valid range of test inputs, its important to detect and remove outliers

Correlation :

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variables. Correlation. Statistical technique which determines how one variables moves/ changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. Its a bi- variate analysis measure which describes the association between different variables. In most of the business its useful to express one subject in terms of its relationship with others..

Positive Correlation

Two features (Variables) can be positively correlated with each other. It means when the value of one variable increase then the value of the other variable(s) also increase.

Negative correlation

Two features (Variables) can be negatively correlated each other. It means that the when the value of one variable increase then the value of the other variable decreases

No correlation

Two features (Variables) are not correlated with each other. It means that when

the value of one variable increase or decrease then the value of the other variable (s) doesn't increase or decreases.

Linear Regression

Linear Regression is one of the easiest and most popular Machine learning algorithm. It is statistical method that is used for predict Analysis. linear Regression algorithm shows a linear Relationship algorithm ~~the~~ between a dependent(y) and one or more independent (y) variable, hence called as linear Regression. Since linear Regression shows the linear relationship which means it finds how the value of the dependent Variable is changing according to the value of the independent Variable

Mathematically, we can represent a linear Regression as: $y = a0 + a1x + \varepsilon$

Here, Y = dependent Variable (Target Variable)
X = independent Variable (predictor Variable)
a_0 = intercept of the line (gives an additional degree of freedom) a1 = linear Regression co-efficient (scale factor to each input value). ε = random error.

The Value for x and y variables are training datasets for linear Regression model representation

Random Forest :-

Random Forest is an ensembles technique capable of performing both regression and

Classification tasks with the use of multiple decision trees and technique called Bootstrap and Aggregation, commonly known as bagging The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision tree. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

Conclusion :-

Studied and Successfully implemented Uber price predictor System.

Assignment No. 2

Title :- Email Spam Classification.

Problem Statement :- Classify the email using the binary classification Method. Email Spam detection has two states :- a) Normal State - Not Spam b) Abnormal State - Spam. Use K-Nearest Neighbours and Support Vector Machine for Classification. Analyze their performance.

Objective :-

- To classify email using the K-Nearest Neighbors algorithm.
- To classify email using the SVM algorithm.
- To analyse the performance of the Model.

Theory :

Data Preprocessing :-

Data Preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. When creating a machine learning project, it is not always the case that we come across clean and formatted way. So for this, we use a data

preprocessing task.

Why do we need Data preprocessing? Real-World data generally contains noises and missing values, and may be in an unusable format that cannot be directly used for machine learning models. Data preprocessing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:-

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing data
- Encoding Categorical Data
- Splitting dataset into training and test set.
- Feature scaling

K-Nearest Neighbors:-

- K-Nearest Neighbour is one of the simplest machine learning algorithm based on supervised learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.
- K-NN algorithm can be used for Regression as well

as for classification but mostly it is used for the classification problem.

- K-NN is a non-parametric algorithm, which means it does not make any assumptions on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and at the time of classification, it performs an action on the dataset.
- and when it gets new data, then it classifies that data into a category that much is similar to the new data.
- Example: Suppose, we have an image of a creature that looks similar to cat and dog but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features of the it will put it in either cat or dog category.

Why do we need a K-NN Algorithm?

Suppose there are two categories i.e Category A and category B and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm, as it work on a similarity Measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most features with the help of KNN we can easily identify the category or class of a particular dataset.

Support Vector Machine

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 data point 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. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Example:-

SVM can be understand with example that we have used in the KNN Classifier. Suppose we see a strange cat that also some features of dogs. So if we want a model that can accurately identify whether it is cat or dog so sucha model can be created by using the SVM algorithm. We will first train our models with lots of images of cats and dogs so that it can learn about different features of cats and dogs and then we test it with this strange creature. So as support vector creates a decision boundary between these two data and choose extreme cases, it will see the extreme case of cat and dog. On the basis of the Support vectors, it will classify it as a cat.

Conclusion:-

Successfully implemented email classification model.

Assignment No:- 3

Title :- Implement Gradient Descent Algorithm

Problem Statement :- Implement Gradient Descent Algorithm to find the local minima of a function. For example, find the local minima of the function $y = (x+3)^2$ Starting from the point $x = 2$.

Objective :- To find local minima of a System by using Gradient Descent Algorithm.

Theory

Introduction : Gradient Descent Algorithm

Gradient descent (GD) is an iterative First-Order Optimization used to find a local minimum / maximum of a given function. This method is commonly used in machine learning (ML) and deep learning (DL) to minimize a cost/loss function (eg:- in linear regression). Due to its importance and ease of implimentation, this algorithm is usually taught at the beginning of almost all machine learning courses.
However, it use is not limited to ML/DL only, its being widely used also in areas like :

- Control engineering (robotics, chemical, etc)

- Computer games
- Mechanical engineering

2. Function Requirements

Gradient descent algorithm does not work for all functions. There are two specific requirements. A function has to be:-

- differentiable
- Convex

First, what does it mean it has to be differentiable? If a function is differentiable it has a derivative for each point in its domain. - not all functions meet these criteria.

Next requirement - function has to be convex. For a univariate function, this means that the line segment connecting two function's points layer on or above its curve (it does not cross it). If it does it means that it has a local minimum which is not a global one.

Mathematically, for two points, x_1, x_2 laying on the function's curve this condition is expressed as:-

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

where λ denotes a points a location on a section line and its value has to be between 0 (left point) and 1 (right point). eg: $\lambda = 0.5$ means a location in the middle.

Another way to check mathematically if a univariate function is convex is calculate the

Second derivative and check if its value is always bigger than 0

$$\frac{d^2 f(x)}{dx^2}$$

Gradient Method's Step are :-

1. Choose a starting point (initialisation)
2. Calculate gradient at this point
3. make a scaled step in the opposite direction to the gradient (objective: minimise)
4. repeat points 2 and 3 until one of the criteria is met:
5. maximum number of iterations reached
6. Step size is smaller than the tolerate (due to scaling on a small gradient).

Conclusion :-

Successfully implemented Gradient Descent Algorithms.

Assignment no:- 4

Title:- Implement K-Nearest Neighbours Algorithm

Problem Statement:- Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Objective:-
- To implement K-Nearest Neighbors algorithm
- To calculate confusion matrix, accuracy, error rate, precision, and recall for the given model.

Theory
K-Nearest Neighbors
- K-Nearest Neighbors is one of the simplest Machine learning algorithm based on Supervised learning technique.
- K-NN algorithm assumes the similarity between the new case/data point based on the similarity. This means when data appears then it can easily classified into a well suite category by using K-NN algorithm.
- K-NN algorithm can be used for Regression as well as for classification but mostly it is used for the classification problems.
- K-NN is a non-parametric model, which means

it does not make any assumption on underlying data.

- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it perform an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example:

Suppose, we have an image of a creatures that looks similar to cat and dog. So, for this identification we can use the KNN algorithm, as it work on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs image and based on the most similar features it will put it in either cat or dog category.

Compute Confusion Matrix.

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aims to predict a categorical label for each input instance. The matrix displays the number of true positive (TP) true negative (TN) false positive (FP) and false negative (FN) produced by the model on the test data.

For binary Classification, the matrix will be of a 2x2 table. For Mult-class Classification, the matrix shape will be equal to the number of classes i.e for n classes it will be n x n.

A 2x2 confusion matrix is shown below for the image recognition having a Dog image or not a dog image.

		Actual	
		Dog	Not Dog
Predicted	Dog	True positive	False Positive
	Not Dog	False negative	True Negative

True Positive (TP):- It is total counts having both predicted and actual values are Dog.

True Negative (TN):- It is total counts having both predicted and actual values are Not Dog

False Positive (FP):- It is total counts having prediction is Dog while actually Not Dog.

False Negative (FN):- It is the total counts having predictions is Not Dog while actually, it is Dog.

From the confusion matrix, we find the following matrices.

Accuracy: Accuracy is used to measure the performance of the model. It is the ratio of Total correct instances to the total instances

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

Precision: Precision is a measure of how accurate a model's positive predictions are. It is defined as the ratio of true positive predictions to the total number of positive predictions made by the models.

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

Recall: Recall measures the effectiveness of a classification model in identifying all relevant instance from a dataset. It is the ratio of the number of the True positive (TP) instance to the sum of true positive and false Negative (TN) instances.

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

Conclusion: Successfully implemented K-Nearest Neighbors algorithm and also find out confusion Matrix, accuracy, precision and recall.

Assignment No:- 5

Title:- K-Means Clustering / hierarchical Clustering

Problem Statement:- Implement K-Means clustering / hierarchical clustering on sales_data_samples.csv dataset. Determine the number of clusters using the elbow method.

Objective:-
- To implement K-Means Clustering / hierarchical Clustering
- To identify the number of clusters.

Theory

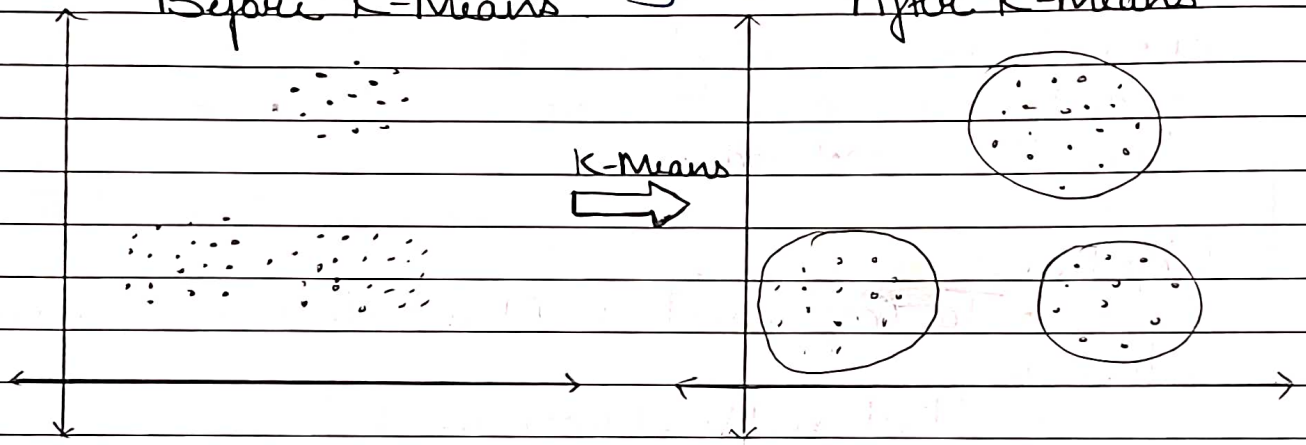
K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the Clustering problems in machine learning or data Science. It groups the unlabeled datasets into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters and for K=3, there will be three clusters and so on.

It allow us to cluster the data into different groups and a convenient way to discover the categories of group in the unlabelled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The below diagram explains the working of the K-means clustering Algorithm:-



The working of the K-Means Algorithm is explained in the below steps:-

Step1: Select the number K to decide the number of clusters

Step2: Select random K points or centroids.

Step3: Assign each data point to their closest Centroid, which will form the predefined K Clusters.

Step4: Calculate the variance and place a new centroid of each other cluster.

Step5: Repeat the third steps, which means reassign each datapoint to the new closest Centroid of each cluster.

Step6 :- If any reassignment occurs, then go to Step-4 else go to Finish.

Step7 :- The model is ready

Hierarchical Clustering :-

It is another unsupervised machine learning Algorithm, which is used to group the unlabelled dataset into a cluster and also known as hierarchical cluster analysis or HCA. In this algorithm, we develop the hierarchy of Clusters in the form of a tree and this tree-Shaped Structure is known as the dendrogram.

Sometimes the result of K-Means Clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means Algorithm.

The hierarchical Clustering Technique has two approaches :

1) Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with the taking all data points as single clusters and merging them until one cluster is left.

2) Divisive : The divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

Elbow Method.

The Elbow Method is the one of the most popular ways to find the optimal numbers of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS is given below :-

$$WCSS = \sum P_i \text{ in cluster1 distance}(P_i C_1)^2 + \sum P_i \text{ in cluster2 distance}(P_i C_2)^2 + \sum P_i \text{ in cluster3 distance}(P_i C_3)^2$$

In the above formula of WCSS :-

$\sum P_i$ Cluster1 distance $(P_i C_1)^2$: It is the sum of the square of the distance between each data point and its centroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and centroids, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal Value of clusters, the Elbow method follows the following below steps :-

- It executes the K-means Clustering on a given dataset for different K values (ranges from 1-10)
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated wcss values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Conclusion :-
Successfully implemented K-Means Clustering.