Problem Statement :

Predict the price of the uber ride from a given pickup point to the agreed drop off location.

perform following tasks:

* Pre-process the dataset
* Identify outliers
* Check the correlation
* Implement linear regression and random forest regression model.
* Evaluate the model and compare their respective score like R2, RMSE etc.

Objective:

The objective of a Linear regression model is to find a relationship betn one or more features and a continous target variable.

Theory:

Linear regression is one of the easist and most popular Machine Learning algorithm. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price etc.

Linear regression algorithm shows a linear relationship betn a dependant and one or more independant variables hence called as linear regression.

FOR EDUCATIONAL USE

Conclusion:

We learn linear regression and random forest regression models and linear regression makes predictions for continuous /real or numeric variables.

Engaging

FOR EDUCATIONAL USE

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 Neighbors and support vector machine for classification. Analyze their performance

Objective:

1> To detect and filter out spam and phishing emails with about 99.9 percent accuracy.

2> The implication of this is that one out of a thousand messages succeed in evading their email spam filter.

Theory:

a) Normal State - Not spam:

Understanding the problem is a crucial first step in solving any machine learning problem. In this practical we will explore and understand the process of classifying emails as spam or not spam. This is called spam detection. and it is a binary classification problem.

b) Abnormal State - spam use KNN and SVM:

In supervised learning a set of input variable such as blood metabolite or gene expression levels, are used to predict a quantitative response variable like harmone level or a qualitative one such as healthy versus diseased individuals.

Both have been successfully applied to challenging

Sundaram
FOR EDUCATIONAL USE

pattern-recognition problems in biology and medicine. SVM and KNN examplify several important trade off in machine learning (ML). SVM is less computationally demanding than KNN and easier to interpret but can identify only a limited set of patterns.

Conclusion

In this way we studied about Email classification using binary classification method and email spam detection.

Essentials

Problem statement:

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

Objective:

To use a database in which the data points are separated into several classes to predict the classification of a new sample point.

Procedure -

K-Nearest Neighbor (KNN) algorithm.

- K-Nearest Neighbor is one of the simplest ML algorithm based on supervised learning technique.
- KNN algorithm assumes the similarity betn the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- KNN 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 well suite category by using KNN algorithm.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because

Sundaram

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.

By calculating the Euclidean distance we got the nearest neighbors, as three nearetneighbors in category A and two nearest neighbors in category B.

Euclidean Distance betn A_1 and B_2

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

Confusion matrix

It is a performance measurement of ML classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

Example

Predicted value	Actual Values: Positive(1)	Actual Values: Negative(0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Conclusion:

In this way we find the neighbor data using a K-Nearest neighbor algorithm and by using confusion matrix we find accuracy, error, precision and recall.

Sundaram
FOR EDUCATIONAL USE

Problem Statement:

Implement k-means clustering / hierachical clustering on sales-data-sample.csv dataset. Determine the number of clusters using theelbow method.

Objective.

To minimize the sum of distances betn the points and their respective cluster centroid

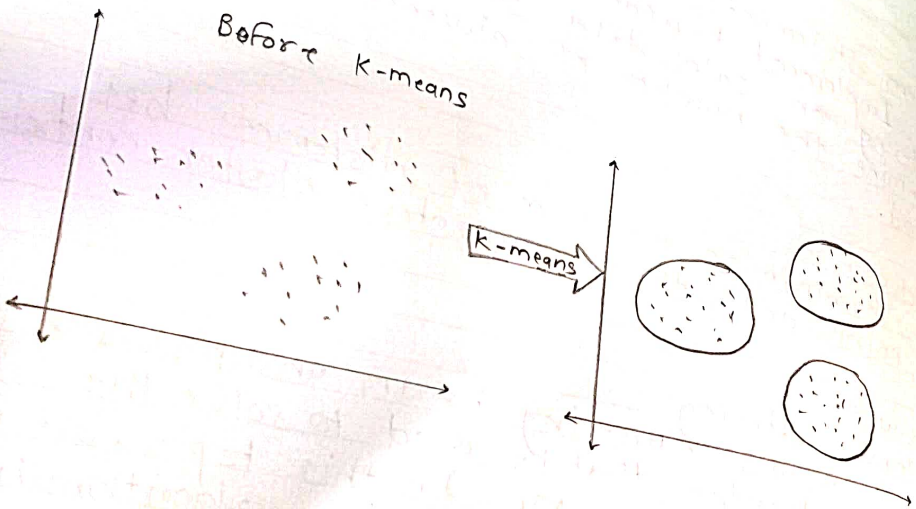
Procedure.

K-Means Clustering:

K-means clustering is an unsupervised learning algorithm that is used to solve the clustering problems in ML. In this topic, we will learn what is k-means clustering algorithm, how the algorithm works, along with the python implementation of k means clustering.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for k center point of centroids by aniterative process.
- Assigns each data point to its closest k-center. Those data points which are near to to particular k-~~n~~ center, create a cluster. Hence each cluster has data points with some commonlities, and it is away from other clusters.



Elbow Method

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

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

where, $\sum_{P_i \text{ in cluster1}} \text{distance}(P_i C_1)^2$. This sum of the square of the distance betn each data point and its centroid within a cluster 1.

Conclusion:

In this way we find the similar group of cluster using K-means clustering and find the number of groups elbow method.

Sundaram
FOR EDUCATIONAL USE