Assignment 2

Rajdhani
DATE / /

Aim:-

Classify the email wing binary classification method email spam detection has two states: Normal state - not spam, Abnormal state - spam. Use K-nearest neighbours and support vector machine for classification, analyze their performance.

Requirements:-

Jupyter notebook, python libraries: pandas, numpy, matplotlib, sklearn

Theory:-

K-nearest neighbours:

K-nearest neighbour is one of the simplest machine learning algorithms 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 classification, but mostly

Rajdhani

DATE / /

it is used for the classification problems.

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 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 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 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 based on the most similar features it will put it in either cat or dog category

How does KNN work ?

Step 1: Select the number 'K' of the neighbours

Step 2: Calculate the Euclidean distance of

Rajdhani
DATE / /

K number of neighbours.

Step 3: Take the K nearest neighbours as per the calculated Euclidean distance.

Step 4: Among these K neighbours, count the number of the data points in each category.

Step 5: Assign the new data points to that category for which the number of the neighbour is maximum.

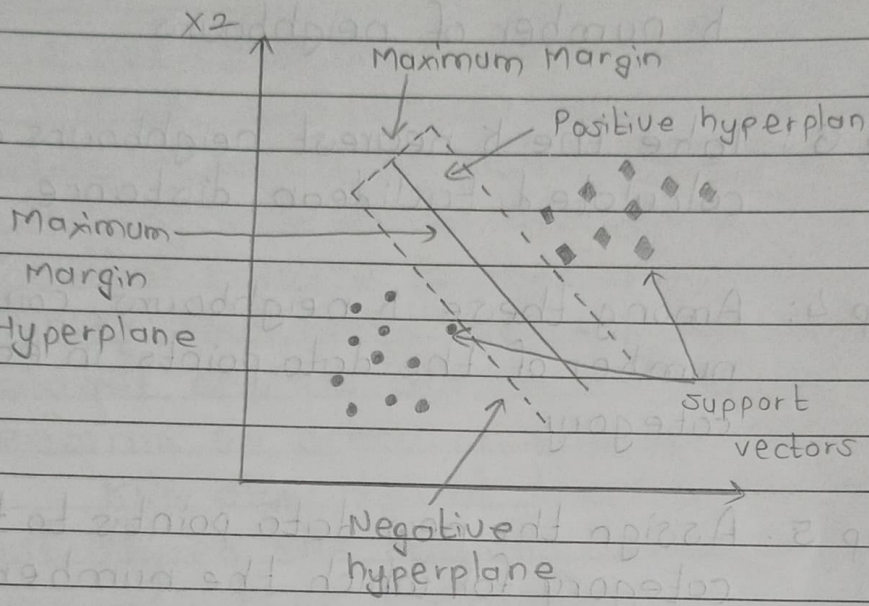
Step 6: Our model is ready.

Support Vector Machine 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 data point in the correct category in the future. This best decision boundary is called a hyperplane.

Rajdhani
DATE / /



Types of SVM:-

1) Linear SVM.
2) Non-linear SVM.

Linear SVM:-

Linear SVM is used for linearly separable data, which means if 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,

Rajdhani
DATE / /

then such data is termed as non-linear data and classifier used is called as non-linear SVM classifier.

Conclusion:-

Hence, we have used KNN and SVM for email spam classifications.