Assignment 2

Aim: 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. Dataset link: The emails.csv dataset on the Kaggle

Software Requirements:

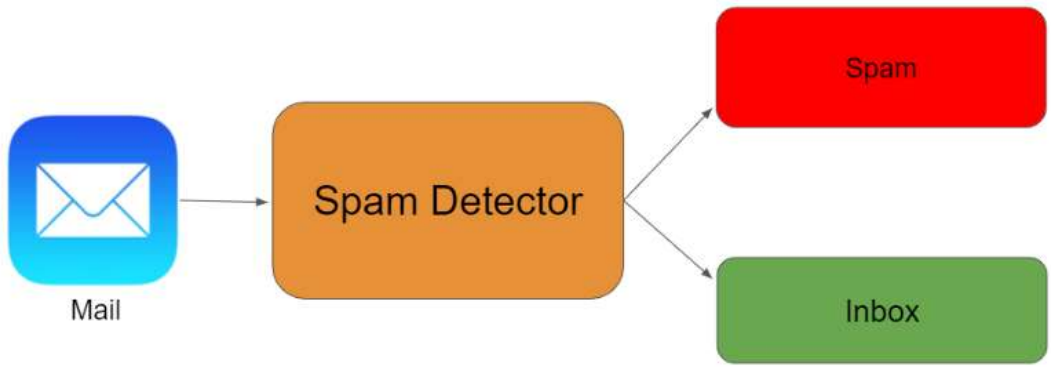
- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory-3

Email Spam Detection

The idea of this post is to understand step by step working of the spam filter and how it helps in making everyone life easier. Also, next time when you see a "You have won a lottery" email rather than ignoring it, you might prefer to report it as a spam. The email spam definition is ambiguous since everybody has their views on it. At present, email spam is getting the attention of everyone. Email spam ordinarily includes particular spontaneous messages sent in mass by individuals you do not know.

Understanding the problem is a crucial first step in solving any machine learning problem. In this article, 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.The reason to do this is simple: by detecting unsolicited and unwanted emails, we can prevent spam messages from creeping into the user's inbox, thereby improving user experience.



The K-nearest neighbors (KNN)

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning KNNalgorithms. is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors ○ **Step-2:** Calculate the Euclidean distance of **K number of neighbors** ○ **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance. ○ **Step-4:** Among these k neighbors, 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 neighbor is maximum. ○ **Step-6:** Our model is ready.

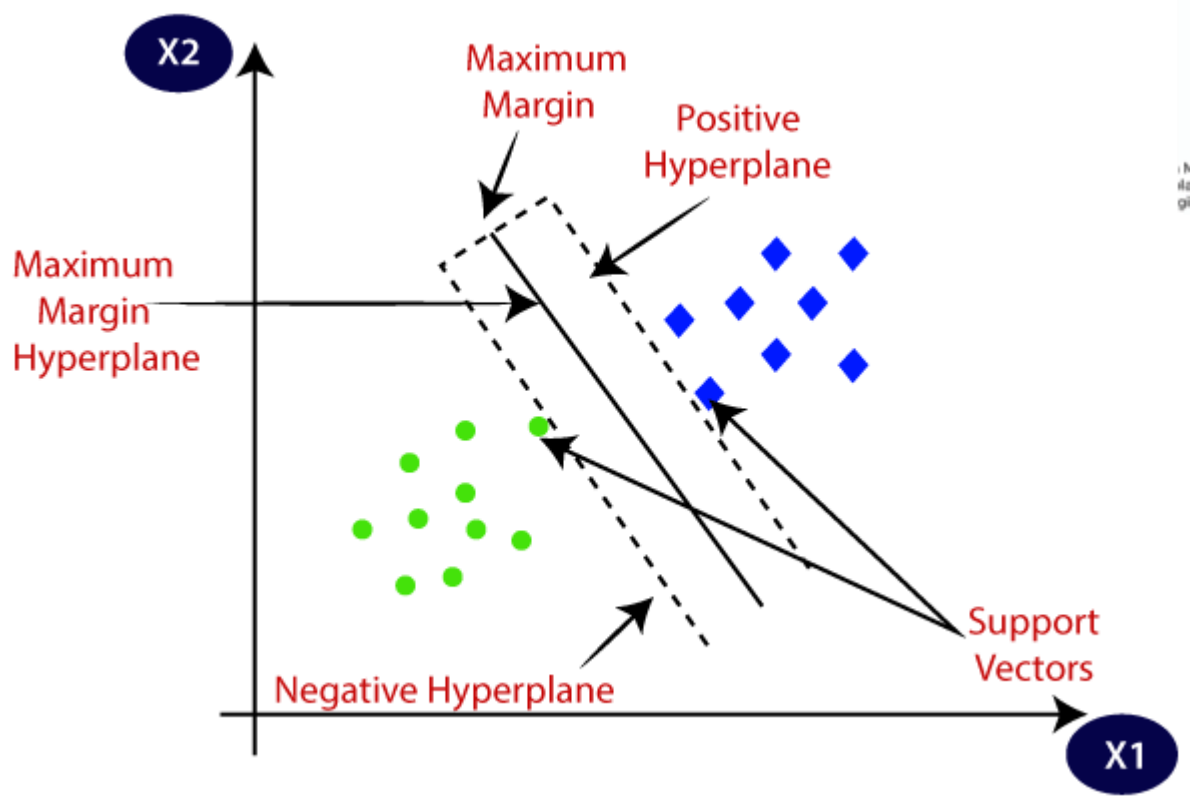
Support Vector Machine Algorithm

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.

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

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

Support Vectors:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

Conclusion:

KNN is a simple yet powerful classification algorithm. It requires no training for making predictions, which is typically one of the most difficult parts of a machine learning algorithm.

SVM is a supervised learning algorithm which separates the data into different classes through the use of a hyperplane.