ICFAI FOUNDATION FOR HIGHER EDUCATION

IcfaiTech
Faculty of Science & Technology (FST)

MACHINE LEARNING ASSIGNMENT

NAME : INDRANI INAPAKOLLA
ENO : 17STUCHH010056
FACULTY : MR. BRAHMA NAIDU

TOPIC : SUPPORT VECTOR MACHINE ALGORITHM (SVM)

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 the 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 :

1. **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.
2. **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.

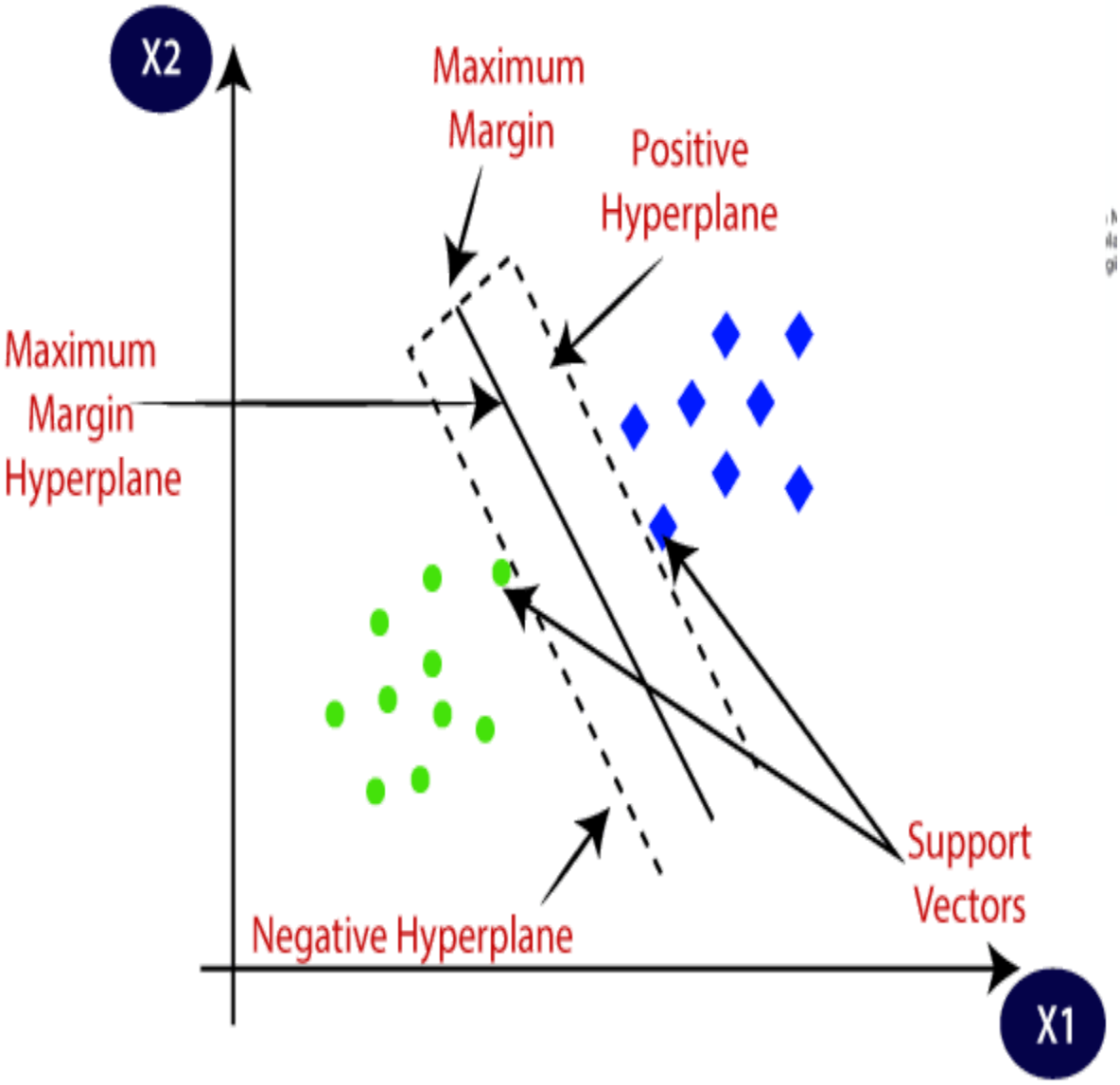
X2
Maximum
Margin
Positive
Hyperplane
Maximum
Margin
Hyperplane
Support
Vectors
Negative Hyperplane
X1

FIG (1)

OVERVIEW OF SUPPORT VECTOR MACHINES

SVM is a new machine learning method based on statistical learning theory and it is classified as one of the computational approaches developed by Vapnik .. Based on the structural risk minimization (SRM) principle, SVM can get decision-making rules and achieve small error for independent tests set and hence can solve the learning problems efficiently . Recently SVM has been applied to solve the problems such as nonlinear, local minimum and high dimension. In many practical applications, SVM can ensure higher accuracy for a long-term prediction compared to other computational approaches.

SVM is based on the concept of decision planes that define decision boundaries. SVM creates a hyperplane by using a linear model to implement nonlinear class boundaries through some nonlinear mapping input vectors into a high-dimensional feature space . In SVM, there is some unknown and nonlinear dependency for example in mapping of function = () between some high-dimensional input vector x and scalar output (or the vector output y as in the case of multiclass SVM). No information regarding the underlying joint probability functions and one must contribute a distribution-free learning. Training data set D= {(xi, yi) X x Y }, I = 1, l where l stands for training data pairs and it is the same to the size of training data set D. Frequently yi is stated as di, where d stands for desired target value. So, SVM is a part of supervised learning techniques.

There are three major advantages of SVM, they are 1) Only two parameters to be chosen, upperbound and the kernel parameter, 2) unique, optimal and global for solving a linearly constrained quadratic problem, the solution of, 3) good generalization performance due to the implementation of SRM principal.

APPLICATIONS OF SUPPORT VECTOR MACHINES

As we have seen, SVMs depend on supervised learning algorithms. The aim of using SVM is to correctly classify unseen data. SVMs have a number of applications in several fields.

Some common applications of SVM are-

1.Face detection – SVMc classify parts of the image as a face and non-face and create a square boundary around the face.

2.Text and hypertext categorization – SVMs allow Text and hypertext categorization for both inductive and transductive models. They use training data to classify documents into different categories. It categorizes on the basis of the score generated and then compares with the threshold value.

3.Classification of images – Use of SVMs provides better search accuracy for image classification. It provides better accuracy in comparison to the traditional query-based searching techniques.

4.Bioinformatics – It includes protein classification and cancer classification. We use SVM for identifying the classification of genes, patients on the basis of genes and other biological problems.

5.Protein fold and remote homology detection – Apply SVM algorithms for protein remote homology detection.

6.Handwriting recognition – We use SVMs to recognize handwritten characters used widely.

7.Generalized predictive control(GPC) – Use SVM based GPC to control chaotic dynamics with useful parameters.

OBSERVATION

DATA VISUALIZATION

1. We notice, first of all, the time doesn't impact the frequency of frauds. Moreover, the majority of frauds are small amounts.
2. This dataset is unbalanced which means using the data as it is might result in unwanted behaviour from a supervised classifier. To make it easy to understand if a classifier were to train with this data set trying to achieve the best accuracy possible it would most likely label every transaction as a non-fraud

CORRELATION OF FEATURES

1. As we can notice, most of the features are not correlated with each other. This corroborates the fact that a PCA was previously performed on the data.
2. What can generally be done on a massive dataset is a dimension reduction. By picking th emost important dimensions, there is a possiblity of explaining most of the problem, thus gaining a considerable amount of time while preventing the accuracy to drop too much.
3. However in this case given the fact that a PCA was previously performed, if the dimension reduction is effective then the PCA wasn't computed in the most effective way. Another way to put it is that no dimension reduction should be computed on a dataset on which a PCA was computed correctly.

DATA SELECTION

OVERSAMPLING : One way to do oversampling is to replicate the under-represented class tuples until we attain a correct proportion between the class

UNDERSAMPLING : However as we haven't infinite time nor the patience, we are going to run the classifier with the undersampled training data (for those using the undersampling principle if results are really bad just rerun the training dataset definition)

MODEL SELECTION

1. we'll use a SVM model classifier, with the scikit-learn library.
2. In this case we are gonna try to minimize the number of errors in our prediction results. Errors are on the anti-diagonal of the confusion matrix. But we can infer that being wrong about an actual fraud is far worse than being wrong about a non-fraud transaction.
3. That is why using the accuracy as only classification criterion could be considered unthoughtful. During the remaining part of this study our criterion will consider precision on the real fraud 4 times more important than the general accuracy. Even though the final tested result is accuracy.

ANALYSIS

1. In this previously used SVM model, the weight of each class was the same, which means that missing a fraud is as bad as misjudging a non-fraud. The objective, for a bank, is to maximize the number of detected frauds! Even if it means considering more non-fraud tuples as fraudulent operations. So, we need to minimize the False positives : the number of no detected frauds.
2. Indeed, by modifying the class_weight parameter, we can choose which class to give more importance during the training phase. In this case, the class_1 which describes the fraudulent operations will be considered more important than the class_0 (non-fraud operation). However, in this case we will give more importance to the class_0 due to the large number of misclassified non-fraud operations. Of course the goal is to lose as little effective fraud as possible in the process.