SVM Report

Donal Loitam

July 20, 2022

Contents

1 Introduction

- A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes
- SVMs are based on the idea of finding a hyperplane that best divides a dataset into two 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**
- Consider the below image to get a clear picture od **hyperplane** and **support vector**

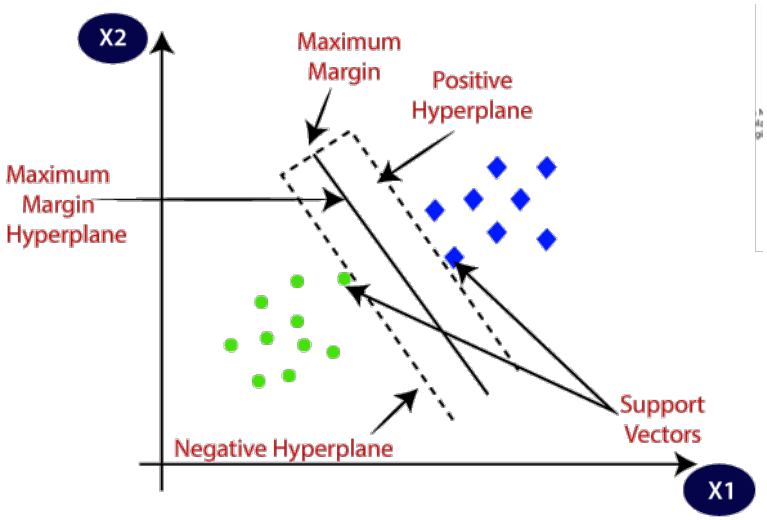
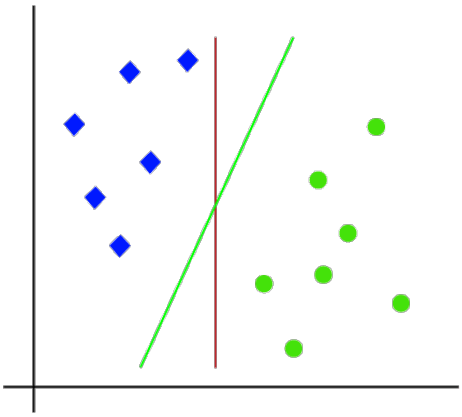


Figure 1: Illustration of support vectors and hyperplane

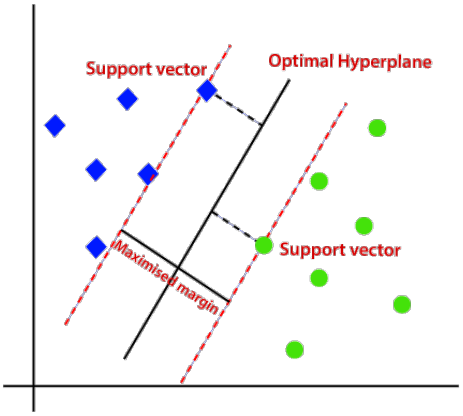
Support Vectors: Support vectors are the data points nearest to the hyperplane, the points of a data set that, if removed, would alter the position of the dividing hyperplane.
Hyperplane: The best decision boundary that helps to classify the data points in n-dimensional space. The dimensions of the hyperplane depend on the features present in the dataset
2 features (as shown in image) $\implies$ hyperplane is a straight line.
And 3 features $\implies$ hyperplane is a 2-dimension plane and so on .....

2 Key Points of the Algorithm

The fundamental concept behind SVM is to find the right hyperplane
Goal of the algorihtm: The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a greater chance of new data being classified correctly.
The algorithm: With a linearly seperable example -
Suppose we have a dataset that has two tags (green and blue), and the dataset has two features X1 and X2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. There may be multiple hyperplanes(here a line) that can perform the separation. Consider the image:



The algorithm finds the closest point of the lines from both the classes(Support vectors). The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. It was found that the **green hyperplane** is the **optimal** hyperplane



NON-LINEAR SVM - In the scenario below, we can't have linear hyper-plane between the two classes, so how does SVM classify these two classes?

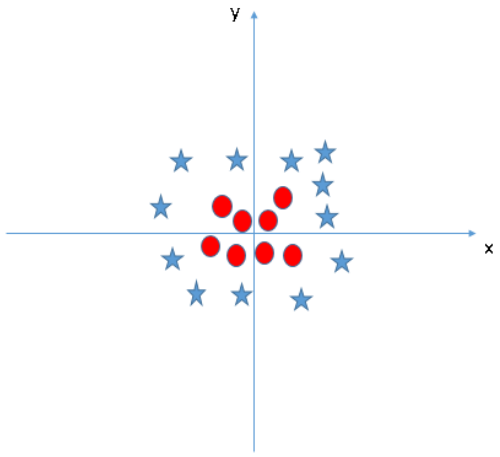


Figure 2: Non-linear SVM

SVM can solve this problem. It solves this problem by introducing additional feature. Here, we will add a new feature $z = x^2 + y^2$. By adding the third dimension, the sample space will become as below image.Now, let's plot the data points on axis x and z:

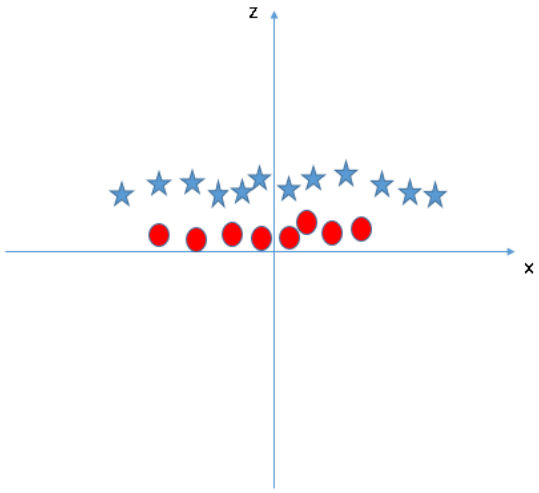


Figure 3: SVM

In above plot, points to consider are:

- All values for z would be positive always because z is the squared sum of both x and y
- In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to higher value of z.

But how do we know which particular feature to add such that the data becomes linearly sepearble. the SVM algorithm has a technique called the **kernel trick**.

3 Some Questions

1. Does any other vector apart form the supportvector has an influence oon decision boundary ?
Ans: Any instance that is not a support vector (not on the margin boundaries) has no influence whatsoever; you could remove them or add more instances, or move them around, and as long as they stay off the margin they won't affect the decision boundary.
2. Can an SVM classifier outputs a confidence score when it classifies an instance
Ans.
3. What are hard margin and soft Margin SVM's ?
Ans
4. What do you mean by Hinge loss
Ans.
5. What is a slack variable ?
Ans