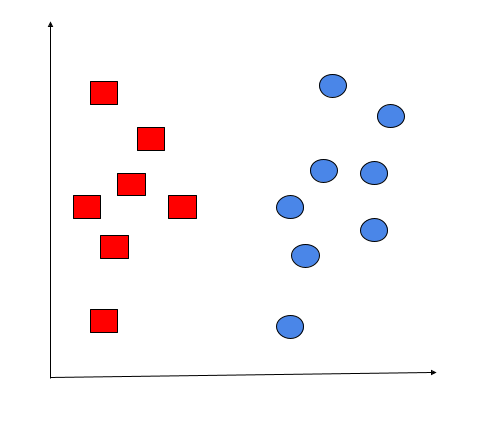
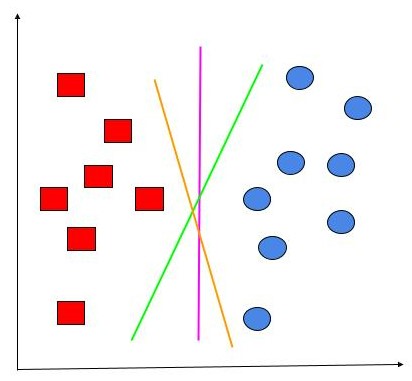
# Support Vector Machines

**Support Vector Machine (SVM)** is a supervised learning algorithm used for regression, classification, and outlier detection. The main objective of SVM is to create a line or a hyperplane which separates the data into classes.

Let’s understand this with an example. Suppose you have data points shown below and you need to classify the points into red rectangles and blue circles. So the key idea is to find a line that separates both these classes.



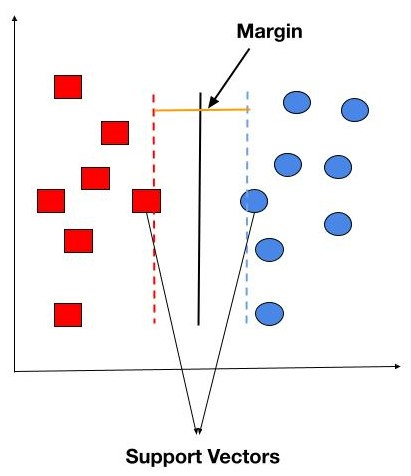
It is not very difficult to draw a line to separate both these classes, however, you could draw multiple lines that do this. So how do we know which one of these many lines is the best answer to our problem?



From the lines shown above, we can say the magenta line is the best one because it’s relatively further apart from either side of the data points, in comparison to the other lines. The green or orange lines may have been able to classify the training data above in a good enough manner, but we cannot guarantee that they would perform well on new, unseen data points.

Obviously, we cannot always just visualize and select the lines that will come up with the best classification. We need a more standardized procedure to do this.

Let’s understand how Support Vector Machines accomplish this task. An SVM finds the points closest to the line from both classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. Our goal is to maximize the margin. The hyperplane for which the margin is maximum is the optimal hyperplane for the SVM.



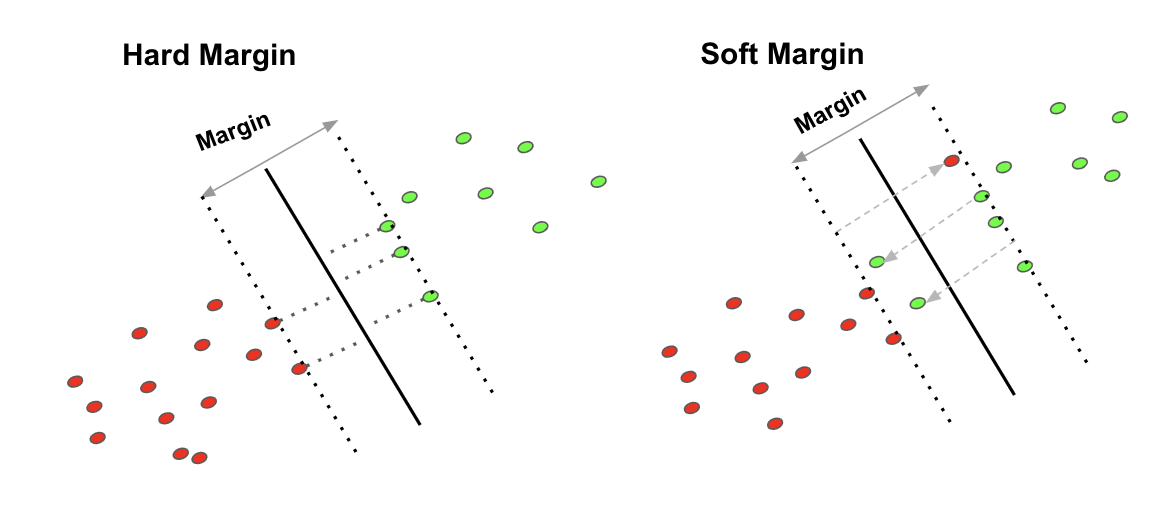
Thus, an SVM tries to make a decision boundary in such a way that the separation or margin between the two classes (the street) is as wide as possible.

We have two kinds of margins

1. Hard Margin
2. Soft Margin

**Hard Margin :**

In the Hard Margin method, the decision boundary makes sure that all the data points are classified correctly. This means that the classifier is not going to make any room for error in the training data. It may also reduce the size of the margin to accomplish this, defeating the whole purpose of using an SVM. This method is also sensitive to outliers.



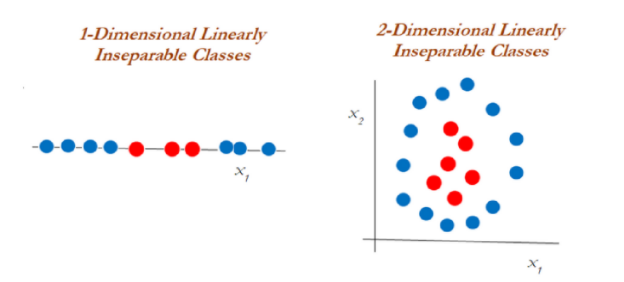
**Soft Margin :**

As most real-world data is not fully linearly separable, we should allow some margin violation or misclassification of the samples to occur in our margin; this is called the Soft Margin method of classification. The idea is to keep the margin as wide as possible. It tries to balance the trade-off between finding a line that maximizes the margin, as well as minimizing the misclassification. The Soft Margin method is robust to outliers and generalizes well on unseen data.

**Non-linear SVM:**

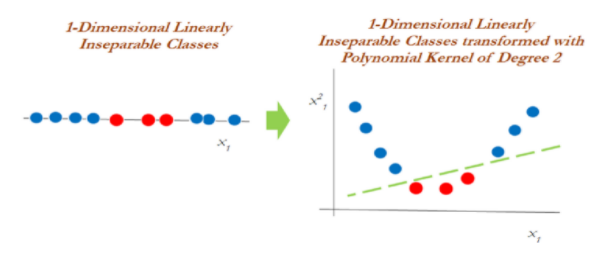
So far, the data we have seen has been linearly separable. But what if it’s not?

In that case, we cannot separate the data by a simple straight line.



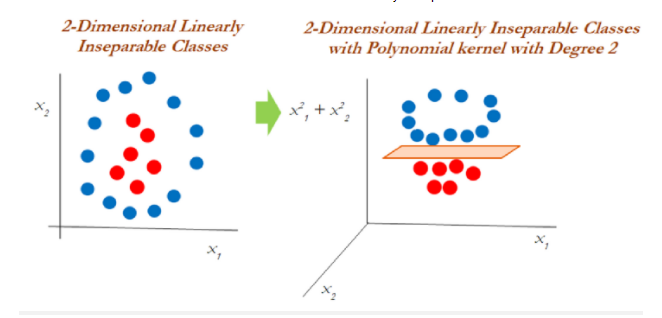
The second diagram above illustrates inseparable classes in a one-dimensional and two-dimensional space.

When it is difficult to separate non-linear classes, we can apply a technique called the **kernel trick** that helps handle the data.



In the above diagram, the data that was inseparable in one dimension got separated once it was transformed into two dimensions and after applying a polynomial kernel of the second degree.

Now let us see how to similarly handle the two-dimensional linearly inseparable data,



In two-dimensional data, the polynomial kernel of the second degree is applied by using a linear plane after transforming it to into higher dimensions.

**Kernel Functions:**

Kernel functions can be regarded as the tuning parameters in an SVM model. They are responsible for removing the computational requirement to achieve the higher dimensional vector space and deal with non-linearly separable data. Let us discuss two of the most widely used kernel functions:

1. Polynomial kernel
2. Radial Basis Function kernel

**1. Polynomial kernel**

A polynomial function is used with a degree k to separate the non-linear data by transforming it into higher dimensions. Take a look at the following equation:

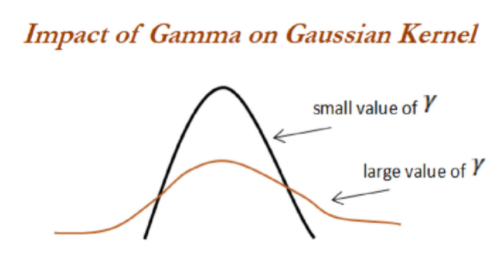
nonlinear svm4.png

**2. Radial Basis Function kernel**

This kernel function is also known as the Gaussian kernel function. It is capable of producing an infinite number of dimensions to separate the non-linear data. It depends on a hyperparameter ‘γ'(gamma) which needs to be scaled while normalizing the data. The smaller the value of the hyperparameter, the smaller the bias and the higher the variance it gives. On the other hand, a higher value of hyperparameter gives a higher bias and lower variance solutions.  It is explained with the help of the following equation:

nonlinear svm5.png

Let us understand the impact of gamma with an example:



Here, as we can see, when gamma is low, the ‘curve’ of the decision boundary is very low and thus the decision region is very broad. When gamma is high, the ‘curve’ of the decision boundary is high, which creates islands of decision boundaries around data points.

**Advantages and disadvantages:**

Some of the advantages of SVMs are:

1. They are flexible with respect to unstructured, structured, and semi-structured data.
2. The Kernel function eases the complexities in almost any data type.
3. Overfitting is not observed as often as in other models.

Despite these advantages, it also holds certain disadvantages which are:

1. Training time is more while computing large datasets.
2. Overall interpretation is difficult because of some black-box approaches.