Support Vector Machine

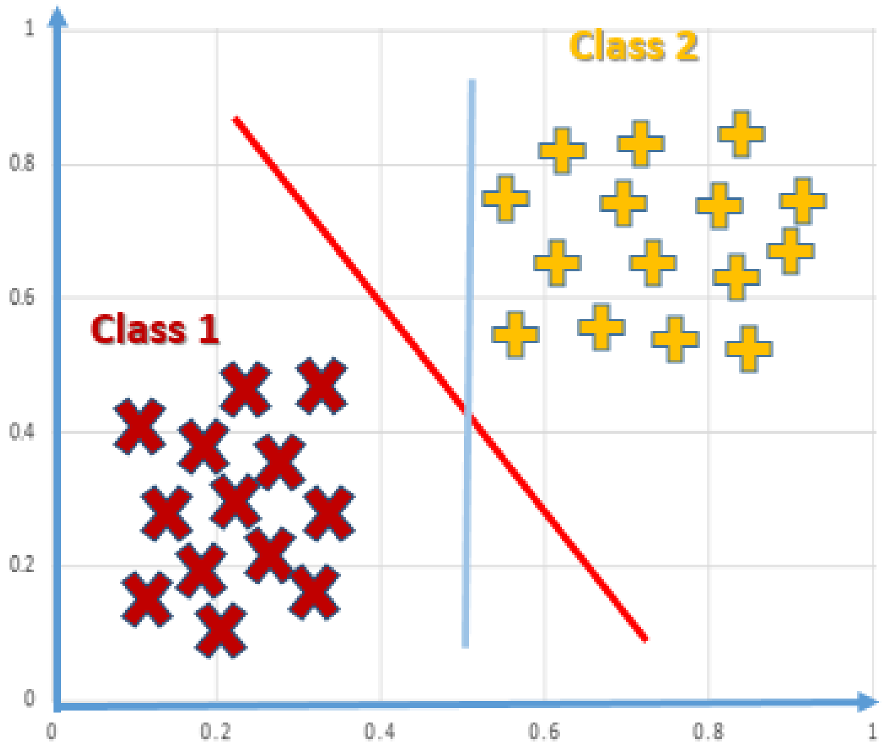
Presented by: Muhammad Zaqeem

Overview:

- Support Vector Machine
- Type of Support Vector Machine
- Soft Margin vs Hard Margin
- Multi class classification
- How it Works
- Kernel in SVM
- Advantage of SVM
- SVM vs Logistic Regression

Support Vector Machine:

- SVM was first developed by Vladimir Vapnik and Alexey Chervonenkis in the 1964
- "Support Vector Machine" (SVM) is a supervised learning machine learning algorithm that can be used for both classification or regression challenges.
- However, it is mostly used in classification problems, such as text classification
- SVM is particularly useful when there is a limited amount of label data
- the objective is to find a decision boundary (also called a hyperplane) that separates the data points belonging to different classes



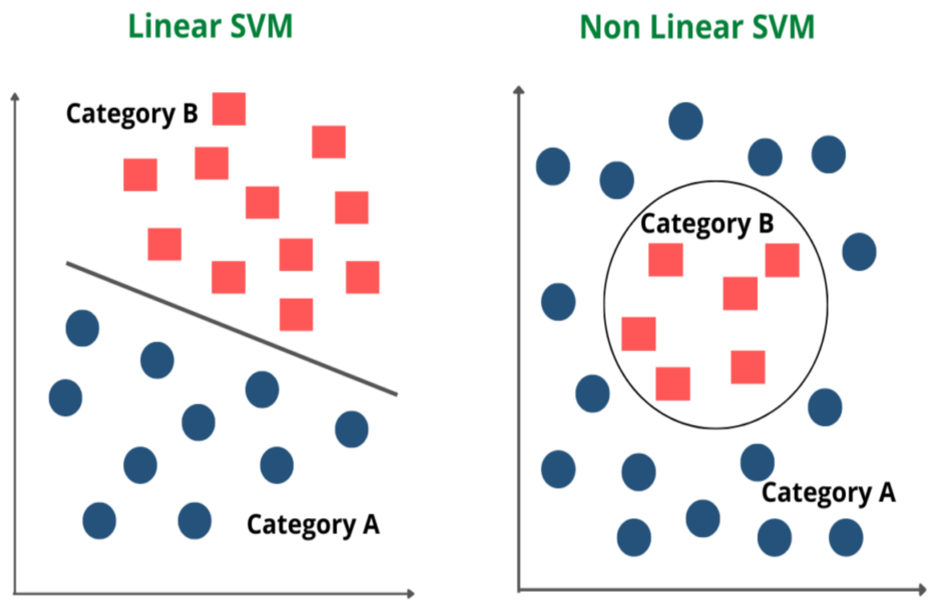
Type of SVM:

Linear SVM:

- It is used when the data is linearly separable.
- It means that the classes can be separated with a straight line (in 2D) or a flat plane (in 3D).
- The SVM algorithm finds the hyperplane that best divides the data into classes.

Non-Linear SVM:

- It is used when the data is not linearly separable
- SVM employs kernel functions to transform the data into a higher-dimensional space
- The algorithm then finds the optimal hyperplane in this new space.



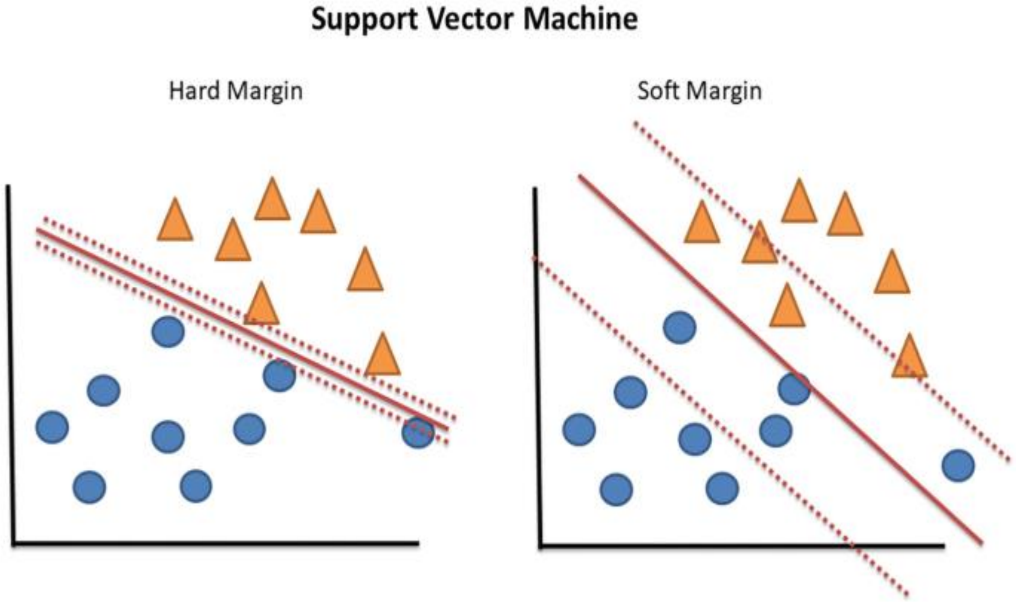
Soft Margin vs Hard Margin:

Soft Margin:

- Soft Margin SVM is used when the data is **not perfectly separable**.
- It allows some misclassification or points to be within the margin
- Soft Margin SVM is more robust to noise and outliers because it doesn't insist on perfect separation.

Hard Margin:

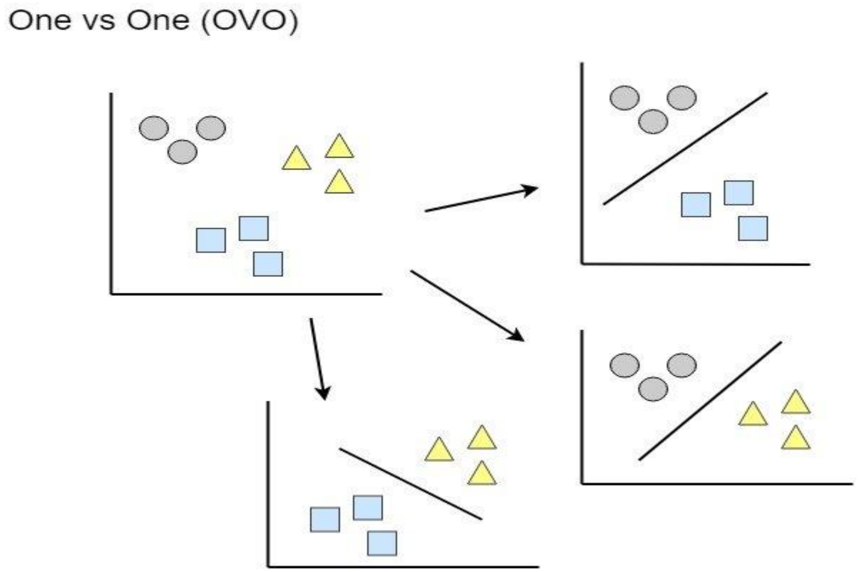
- Hard Margin SVM is used when the data is perfectly separable.
- No data points are allowed to be on the wrong side of the margin.
- It is not robust to **outliers**. One outlier can cause the model to perform poorly because it tries to strictly separate all points



Hard Margin: It doesn't accept data inside margin. It could result overfitting.
Soft Margin: Accept that data is inside margin

Multiclass Classification

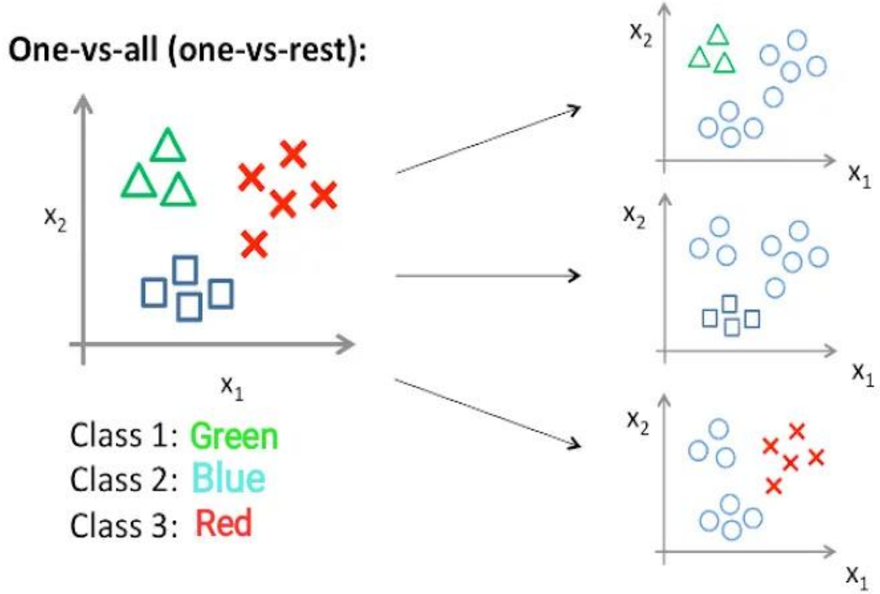
➢ The process of categorizing data into more than two distinct classes, where each input belongs to exactly one of several possible categories.



One vs One:

➢ **OvO** is a multiclass classification technique that creates a separate binary classifier for each pair of classes.

➢ Each classifier decides between two classes, and all classifiers "vote" on the final class prediction



One-vs-Rest (OvR):

➢ Creates one classifier for each class, treating the class as positive and all others as negative.

How it Works:

Data preparation: The first step is to collect and prepare the data. This involves selecting the features that will be used to represent the data and possibly scaling or transforming the data to ensure that all features are on the same scale.

Hyperplane: A hyperplane is a line that separates different classes in the feature space.

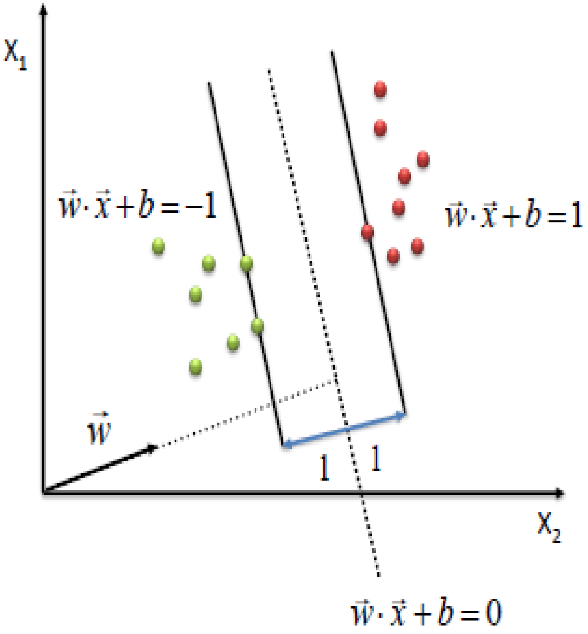
$$w \cdot x + b = 0$$

Support vectors: support vector are the data points closest to the hyperplane.

Margin: margin is the space or gap that exists between the hyperplane and the support vectors, which are the data points closest to the hyperplane.

$$\boxed{\max \frac{2}{\|w\|}}$$

Objective: The goal of SVM is to maximize this margin. The larger the margin, the better the generalization of the classifier to unseen data, as it reduces the risk of overfitting.



$$\boxed{\max \frac{2}{\|w\|}}$$

$s.t.$

$(w \cdot x + b) \geq 1, \forall x \text{ of class } 1$

$(w \cdot x + b) \leq -1, \forall x \text{ of class } 2$

Mathematical Intution:

- We want to maximize: $\text{Margin} = \dfrac{2}{\|\vec{w}\|}$
  - Which is equivalent to minimizing: $L(w) = \dfrac{\|\vec{w}\|^2}{2}$
  - But subjected to the following constraints:

$$y_i = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_i + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_i + \mathrm{b} \leq -1 \end{cases}$$

or

$$y_i(\mathbf{w} \bullet \mathbf{x}_i + b) \geq 1, \quad i = 1,2,\ldots,N$$

Hinge Loss:

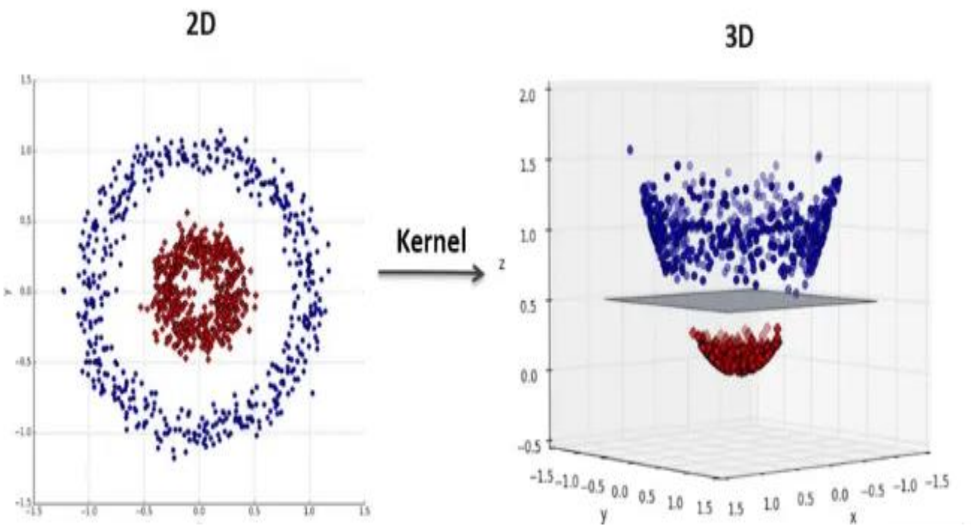
- Hinge loss is used to measure the performance of a classification model, especially for "maximum-margin" classifiers like . It ensures that the predicted classes are not just correct but also far away from the decision boundarySVM.

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

Kernel:

- **kernel** is a function that transforms or measures relationships between data points in a higher-dimensional space, enabling SVM to handle complex, non-linear data separations.

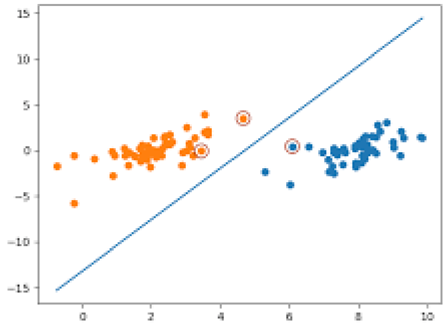
- Here are some common types of kernels in support vector machine algorithms:



Linear Kernel:

- The linear kernel is the simplest and is used when the data is linearly separable.

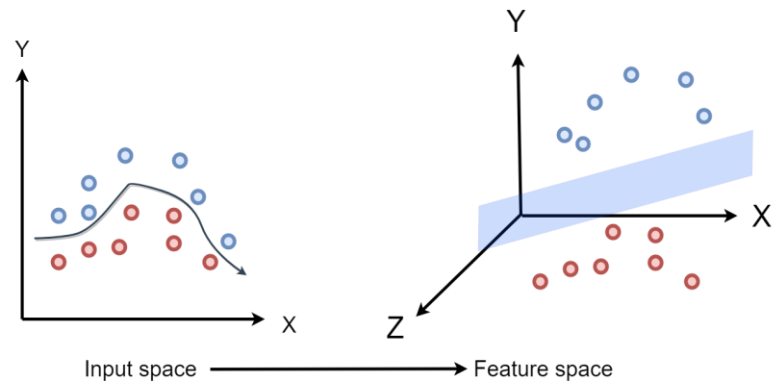
- It calculates the dot product between the feature vectors.



Polynomial Kernel:

- The polynomial kernel is effective for non-linear data.

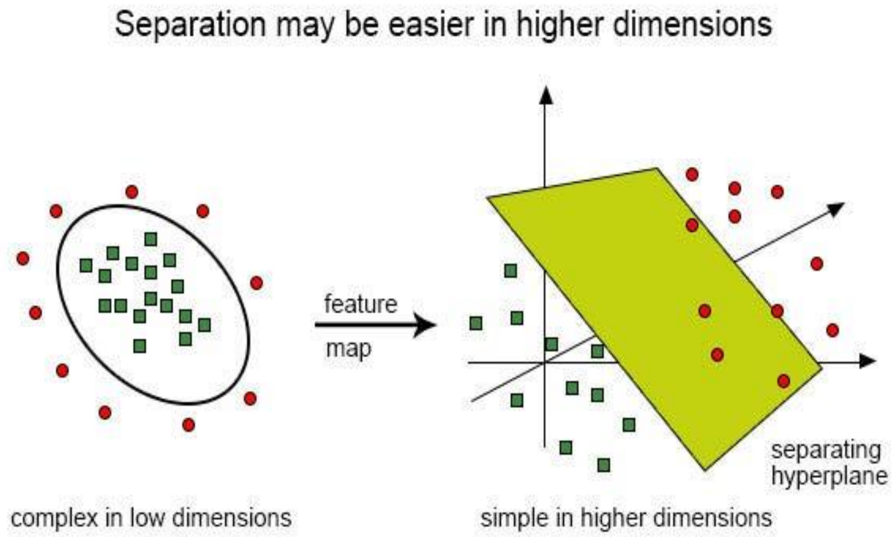
- It computes the similarity between two vectors in terms of the polynomial of the original variables.



Continue…

Radial Basis Function (RBF) Kernel:

- Suitable for data with **non-linear relationships**.
- The RBF kernel (also called the Gaussian kernel) creates circular boundaries around data points.
- It computes the similarity between two points based on the **distance** between them, which is scaled by a parameter.



Sigmoid kernel:

- The **sigmoid kernel** is a function used in Support Vector Machines (SVM) that maps input data into a higher-dimensional space where classes with complex, non-linear boundaries(s-shaped) can be more easily separated

Advantage of SVM:

- SVM is more effective in high dimensional spaces.
- SVM is relatively memory efficient
- SVM handles non-linear separations by transforming data into higher dimensions using kernels
- SVM performs effectively on small to medium-sized datasets by focusing on key support vectors rather than all data points.
- It optimally separates classes by maximizing the margin, reducing misclassifications.

SVM vs Logistic Regression

Aspects	Logistic Regression	Support Vector Machines
Multicolinearity check	Important	Not important
Outliers Handling	Cannot handle well, will skew the probability functions for labels	Can handle, outliers may not intervene with the maximum margin distance
Scaling	Important to make sure no dominance which affect coefficients	Important to ensure no dominance to affect margin distance
Optimization Function	Uses Maximum likelihood to maximize the probability of reaching to a certain label decision.	Uses Maximum Margin Distance to separate positive and negative plane by using kernels (shapes)

Thank You