SUPPORT VECTOR MACHINES(SVM) - AN EASY MATHEMATICAL EXPLANATION USING IRIS DATASET FOR BEGINNERS

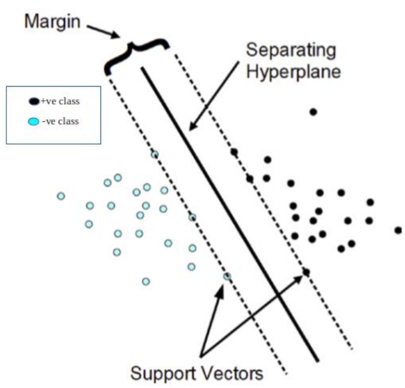
1. INTRODUCTION

Support vector machines is a form of supervised learning technique to solve classification problems, and this technique aims to find a best hyperplane which classify data set to different classes.

support vectors : These are the data points closest to the created hyperplane, these are very critical since these points lies close to the hyperplane hence the removal of these points could alter the position of hyperplane.

Margin: The concept of the margin of a separating line can be used to find the best separator line. The margin of the separator line is defined as the double of the shortest perpendicular distance between data points and separator line.

Maximum Margin Line: The line with highest value for margin is called “maximum margin line” or the “optimal separating line”. This line is otherwise called as “support vector machine”



Hyper planes: These are the subset of finite dimensional spaces, which are similar to straight lines and planes in three-dimensional space. It can be considered as a plane whose dimension one less than that of vector space. That is If you have two features,

Author: Alkha Mohan
Mohan.alkha@gmail.com

then the resulting vector space is of 2 dimention and is separated by a hyperplane of dimension 1 which is a line.

Two class dataset**:** In machine learning a two-class dataset is said to be a dataset in which the target variable takes only one of the two possible class labels. The variable whose value is being predicted is called target variable or output variable

Kernel SVM**:** Kernel SVM (Nonlinear SVM) came to the role when the data is not linearly separable. SVM uses some mathematical functions as kernels to transform the input data in to a required format to transform nonlinearity in data to linearity. There are different types of kernels used in SVM. Let us see some of the common kernels.

- Polynomial Kernel

  $k(x_i, x_j) = (x_i . x_j + 1)^d$, where d is the degree of the polynomial.

- Gaussian Kernel

$$k(x, y) = exp(\frac{-\|x - y\|^2}{2\sigma^2})$$

- Gaussian Radial Basis Kernel (RBF)

$$k(x_i, x_j) = exp(-\gamma\|x_i - x_j\|)^2, \text{ where } \gamma > 0 \text{ or } \gamma = \frac{1}{2\sigma^2}$$

2. ALGORITHM

Step 1: Calculate Euclidian distance between data points in different classes and identify the support vectors (s1,s2,..,sn)

Step 2: Use a bias to modify the support vectors

Step 3: Apply support vector equations and solve to find the α values.

$\alpha 1\ \Phi(s1) \cdot \Phi(s1) + \alpha 2\ \Phi(s2) \cdot \Phi(s1) + \ldots..+\alpha n\ \Phi(sn) \cdot \Phi(s1)$ = class label

where Φ is the kernel function

Step 4: Apply α values in weight vector w˜ equation $\tilde{w}=\sum_i \alpha_i . s_i$

Step 5: y=wx+b produce maximum margin hyperplane.

3. PROS & CONS

Advantages of SVM

- Works well if there is a clear margin of separation between classes
- Applicable to higher dimensional data
- Applicable even if the number of dimensions is greater than the number of samples.
- Works with high memory efficiency.

Disadvantages of SVM

- SVM is not appropriate for the application on large datasets
- SVM struggles when there is a presence of noise in the data
- If number of features for each data point exceeds the number of training data samples, SVM suffers performance degradation
- As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

4. APPLICATIONS of SVM

- Text Classification
- Face Detection
- Fingerprint Identification
- Image classification
- Handwriting recognition
- Geo-spatial data-based applications
- Security-based applications
- Computational biology

5. DATASET DESCRIPTION

This is the best known database to be found in the pattern recognition literature. IRIS data set is created by R.A. Fisher. Fisher's papers and datasets is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. Predicted attribute: class of iris plant.

<u>Attribute Information:</u>

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class: Iris Setosa, Iris Versicolour, Iris Virginica

6. EXPLANATION

5 samples from each classes in iris data set is selected, and given below.

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
SE1	5.1	3.5	1.4	0.2	Iris-setosa
SE2	4.9	3	1.4	0.2	Iris-setosa
SE3	4.7	3.2	1.3	0.2	Iris-setosa
SE4	4.6	3.1	1.5	0.2	Iris-setosa
SE5	5	3.6	1.4	0.2	Iris-setosa
VE1	7	3.2	4.7	1.4	Iris-versicolor
VE2	6.4	3.2	4.5	1.5	Iris-versicolor
VE3	6.9	3.1	4.9	1.5	Iris-versicolor
VE4	5.5	2.3	4	1.3	Iris-versicolor
VE5	6.5	2.8	4.6	1.5	Iris-versicolor
VI1	6.3	3.3	6	2.5	Iris-virginica
VI2	5.8	2.7	5.1	1.9	Iris-virginica
VI3	7.1	3	5.9	2.1	Iris-virginica
VI4	6.3	2.9	5.6	1.8	Iris-virginica
VI5	6.5	3	5.8	2.2	Iris-virginica

Find the Euclidian distance between the samples of first class and other two classes.

$$\boldsymbol{dist(SE1, VE1)} = \sqrt{(\mathbf{5.1} - \mathbf{7})^2 + (\mathbf{3.5} - \mathbf{3.2})^2 + (\mathbf{1.4} - \mathbf{4.7})^2 + (\mathbf{0.2} - \mathbf{1.4})^2} = \mathbf{4}$$

$$\boldsymbol{dist(SE1, VI1)} = \sqrt{(\mathbf{5.1} - \mathbf{6.3})^2 + (\mathbf{3.5} - \mathbf{3.3})^2 + (\mathbf{1.4} - \mathbf{6})^2 + (\mathbf{0.2} - \mathbf{2.5})^2} = \mathbf{5.28}$$

Similarly you can find the distance measures for all sample and can be found as below.

	VE1	VE2	VE3	VE4	VE5
SE1	4	3.62	4.16	3.09	3.79
SE2	4.1	3.69	4.24	2.98	3.81
SE3	4.28	3.85	4.42	3.15	4
SE4	4.18	3.73	4.31	2.98	3.87
SE5	4.06	3.66	4.22	3.15	3.85

	VI1	VI2	VI3	VI4	VI5
SE1	5.28	4.21	5.3	4.69	5.06
SE2	5.34	4.18	5.36	4.71	5.09
SE3	5.47	4.33	5.53	4.87	5.25
SE4	5.34	4.18	5.41	4.72	5.11
SE5	5.31	4.25	5.35	4.73	5.1

As we calculate the determinant of these distance matrices, we get a non-zero value and hence the data is linearly separable.

In order to find the support vectors find the points having minimal Euclidian distance. And those points are highlighted in below table

	VE1	VE2	VE3	VE4	VE5
SE1	4	3.62	4.16	3.09	3.79
SE2	4.1	3.69	4.24	2.98	3.81
SE3	4.28	3.85	4.42	3.15	4
SE4	4.18	3.73	4.31	2.98	3.87
SE5	4.06	3.66	4.22	3.15	3.85

The support vectors for setosa and versicolor are (SE2, SE4, VE4)

S1=SE2=[4.9, 3, 1.4, 0.2]

S2=SE4=[4.6, 3.1, 1.5, 0.2]

S3=VE4=[5.5, 2.3, 4, 1.3]

Now assume bias as 1 and the support vectors are changed to

S1=SE2=[4.9, 3, 1.4, 0.2, 1]

S2=SE4=[4.6, 3.1, 1.5, 0.2, 1]

S3=VE4=[5.5, 2.3, 4, 1.3, 1]

As you have changed support vectors you can now start writing support vector equations which maximises the margin. The equation is given below. (assume class label for setosa = -1 and versicolor = +1)

$$\alpha_1\, \Phi(s_1) \cdot \Phi(s_1) + \alpha_2\, \Phi(s_2) \cdot \Phi(s_1) + \alpha_3\, \Phi(s_3) \cdot \Phi(s_1) = -1$$

$$\alpha_1\, \Phi(s_1) \cdot \Phi(s_2) + \alpha_2\, \Phi(s_2) \cdot \Phi(s_2) + \alpha_3\, \Phi(s_3) \cdot \Phi(s_2) = -1$$

$$\alpha_1\, \Phi(s_1) \cdot \Phi(s_3) + \alpha_2\, \Phi(s_2) \cdot \Phi(s_3) + \alpha_3\, \Phi(s_3) \cdot \Phi(s_3) = +1$$

Since the problem is linearly separable Φ()= Identity , so the above equation reduces to

$$\alpha_1\, \tilde{s}_1 \cdot \tilde{s}_1 + \alpha_2\, \tilde{s}_2 \cdot \tilde{s}_1 + \alpha_3\, \tilde{s}_3 \cdot \tilde{s}_1 = -1$$

$$\alpha_1\, \tilde{s}_1 \cdot \tilde{s}_2 + \alpha_2\, \tilde{s}_2 \cdot \tilde{s}_2 + \alpha_3\, \tilde{s}_3 \cdot \tilde{s}_2 = -1$$

$$\alpha_1\, \tilde{s}_1 \cdot \tilde{s}_3 + \alpha_2\, \tilde{s}_2 \cdot \tilde{s}_3 + \alpha_3\, \tilde{s}_3 \cdot \tilde{s}_3 = +1$$

$$\begin{aligned} \tilde{s}_1 \cdot \tilde{s}_1 &= [4.9, 3, 1.4, 0.2, 1] \,.\, [4.9, 3, 1.4, 0.2, 1] \\ &= 4.9*4.9 + 3*3 + 1.4*1.4 + 0.2*0.2 + 1*1 \\ &=36 \end{aligned}$$

Similarly find the dot products and substitute in the above equation to obtain.

$$36\alpha_1 + 35\alpha_2 + 41\alpha_3 = -1$$

$$35\alpha_1 + 34\alpha_2 + 40\alpha_3 = -1$$

$$41\alpha_1 + 40\alpha_2 + 54\alpha_3 = +1$$

Solving the above equation results in α_1= -2.5, α_2= 2.25 and α_3= 0.25

The weight vector $\tilde{w}$ is given as

$$\tilde{w} = \sum_i \alpha_i . s_i$$

where i is the number of support vectors

Author: Alkha Mohan
Mohan.alkha@gmail.com

As per the equation

$$W = -2.5(4.9, 3, 1.4, 0.2, 1) + 2.25\,(4.6, 3.1, 1.5, 0.2, 1) + 0.25(5.5, 2.3, 4, 1.3, 1)$$

$$= \begin{pmatrix} -2.5 * 4.9 + 2.25 * 4.6 + 0.25 * 5.5 \\ -2.5 * 3 + 2.25 * 3.1 + 0.25 * 2.3 \\ -2.5 * 1.4 + 2.25 * 1.5 + 0.25 * 4 \\ -2.5 * 0.2 + 2.25 * 0.2 + 0.25 * 1.3 \\ -2.25 * 1 + 2.25 * 1 + 0.25 * 1 \end{pmatrix}$$

$$= (-0.525, 0.05, 0.875, 0.275, 0.25)$$

hence the marginal hyper plane is given by the equation $y = \mathrm{wx} + \mathrm{b},$ where $w = (-0.525, 0.05, 0.875, 0.275)$ and $b = 0.25$.

Now let us use the same method to find a marginal hyperplane for the setosa and verginica classes

	VI1	VI2	VI3	VI4	VI5
SE1	5.28	4.18	5.3	4.69	5.06
SE2	5.34	4.18	5.36	4.71	5.09
SE3	5.47	4.33	5.53	4.87	5.25
SE4	5.34	4.18	5.41	4.72	5.11
SE5	5.31	4.25	5.35	4.73	5.1

The support vectors for setosa and virginica are (SE1, SE2, SE4, VI2)

S1=SE1=[5.1, 3.5, 1.4, 0.2]

S2=SE2=[4.9, 3, 1.4, 0.2]

S3=SE4=[4.6, 3.1, 1.5, 0.2]

S4=VI2=[5.8, 2.7, 5.1, 1.9]

Add bias to support vectors and assume class label for setosa = -1 and virginica = +1

41α1 + 38α2 + 37α3+ 48α4 = −1

38α1 + 36α2 + 35α3+ 45α4 = -1

37α1 + 35α2 + 34α3+ 44α4 = -1

48α1 + 45α2 + 44α3+ 72α4 = +1

Solving the above equation results in α1= -0.11, α2= -1.82 , α3= 1.82 and α3= 0.11

The weight vector w˜ is given as

$$w\tilde{}=\sum_i \alpha_i . s_i$$

as per the equation

$$W = -0.11(5.1, 3.5, 1.4, 0.2,1) + -1.82\ (4.9, 3, 1.4, 0.2,1) + 1.82(4.6, 3.1, 1.5, 0.2,1) + 0.11(5.8, 2.7, 5.1, 1.9,1)$$

$$= \begin{pmatrix} -0.11 * 5.1 - 1.82 * 4.9 + 1.82 * 4.6 + 0.11 * 5.8 \\ -0.11 * 3.5 - 1.82 * 3 + 1.82 * 3.1 + 0.11 * 2.7 \\ -0.11 * 1.4 - 1.82 * 1.4 + 1.82 * 1.5 + 0.11 * 5.1 \\ -0.11 * 0.2 - 1.82 * 0.2 + 1.82 * 0.2 + 0.11 * 1.9 \\ -0.11 * 1 - 1.82 * 1 + 1.82 * 1 + 0.11 * 1 \end{pmatrix}$$

$$= (-0.469,0.094,0.589,0.187,0)$$

hence the marginal hyper plane is given by the equation $\mathrm{y} = \mathrm{wx} + \mathrm{b}$, where $w = (-0.469,0.094,0.589,0.187)$ and $b = 0$

Non Linear Separable

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
VE1	7	3.2	4.7	1.4	Iris-versicolor
VE2	6.4	3.2	4.5	1.5	Iris-versicolor
VE3	6.9	3.1	4.9	1.5	Iris-versicolor
VE4	5.5	2.3	4	1.3	Iris-versicolor
VE5	6.5	2.8	4.6	1.5	Iris-versicolor
VI1	6.3	3.3	6	2.5	Iris-virginica
VI2	5.8	2.7	5.1	1.9	Iris-virginica
VI3	7.1	3	5.9	2.1	Iris-virginica
VI4	6.3	2.9	5.6	1.8	Iris-virginica
VI5	6.5	3	5.8	2.2	Iris-virginica

Step 1: Calculate the Euclidian distance between each sample in versicolor with samples in virginica

Step 2: Check if the data is linearly separable by finding the determinant of distance matrix found in step 1

The above given data is not linearly separable, hence we need proceed with nonlinear SVM.

Step 3: Convert the samples to linearly separable data by applying appropriate kernel function. (kernel functions are mentioned in introduction section)

Step4: Now the data is linearly seperable and continue the procedure for linearly separable data as discussed above,

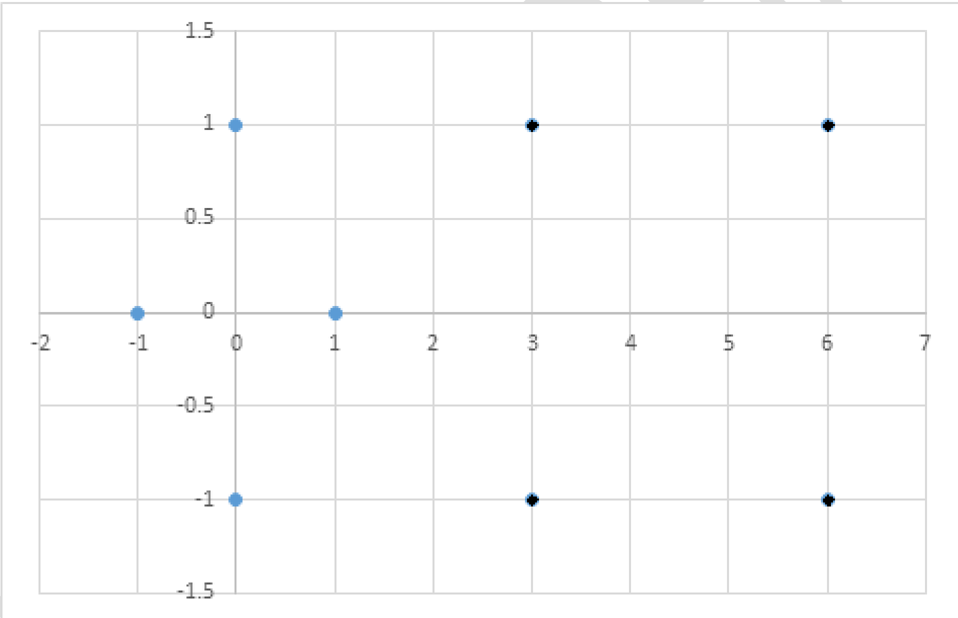
MORE SIMPLIFIED EXAMPLE

Example Problem- Linear SVM

Consider the dataset having values (3,1), (3,-1),(6,1),(6,-1) mapped to the positive class and (1,0),(0,1),(0,-1),(-1,0) mapped to negative class. Find the best line that separate these points.

Solution

Plot the points to the graph and find the support vectors as given below, black denotes positive class points and blue denotes negative class points.



From the graph, it is clear that the support vectors are s1=(1,0) s2=(3,1) and s3=(3,-1)

Now assume bias as 1 and the support vectors are changed to s1=(1,0,1) s2=(3,1,1) and s3=(3,-1,1). As you have changed support vectors you can now start writing support vector equations which maximises the margin. The equation is given below

$$\alpha 1\ \Phi(s1) \cdot \Phi(s1) + \alpha 2\ \Phi(s2) \cdot \Phi(s1) + \alpha 3\ \Phi(s3) \cdot \Phi(s1) = -1$$

$$\alpha_1 \Phi(s_1) \cdot \Phi(s_2) + \alpha_2 \Phi(s_2) \cdot \Phi(s_2) + \alpha_3 \Phi(s_3) \cdot \Phi(s_2) = +1$$

$$\alpha_1 \Phi(s_1) \cdot \Phi(s_3) + \alpha_2 \Phi(s_2) \cdot \Phi(s_3) + \alpha_3 \Phi(s_3) \cdot \Phi(s_3) = +1$$

Since the problem is linearly separable $\Phi()$= Identity, so the above equation reduces to

$$\alpha_1 \tilde{s}_1 \cdot \tilde{s}_1 + \alpha_2 \tilde{s}_2 \cdot \tilde{s}_1 + \alpha_3 \tilde{s}_3 \cdot \tilde{s}_1 = -1$$

$$\alpha_1 \tilde{s}_1 \cdot \tilde{s}_2 + \alpha_2 \tilde{s}_2 \cdot \tilde{s}_2 + \alpha_3 \tilde{s}_3 \cdot \tilde{s}_2 = +1$$

$$\alpha_1 \tilde{s}_1 \cdot \tilde{s}_3 + \alpha_2 \tilde{s}_2 \cdot \tilde{s}_3 + \alpha_3 \tilde{s}_3 \cdot \tilde{s}_3 = +1$$

Substitute the values of support vectors and calculating the dot product will result in

$$2\alpha_1 + 4\alpha_2 + 4\alpha_3 = -1$$

$$4\alpha_1 + 11\alpha_2 + 9\alpha_3 = +1$$

$$4\alpha_1 + 9\alpha_2 + 11\alpha_3 = +1$$

Solving the above equation results in α_1= -3.5, α_2=0.75 and α_3= 0.75

The weight vector $\tilde{w}$ is given as

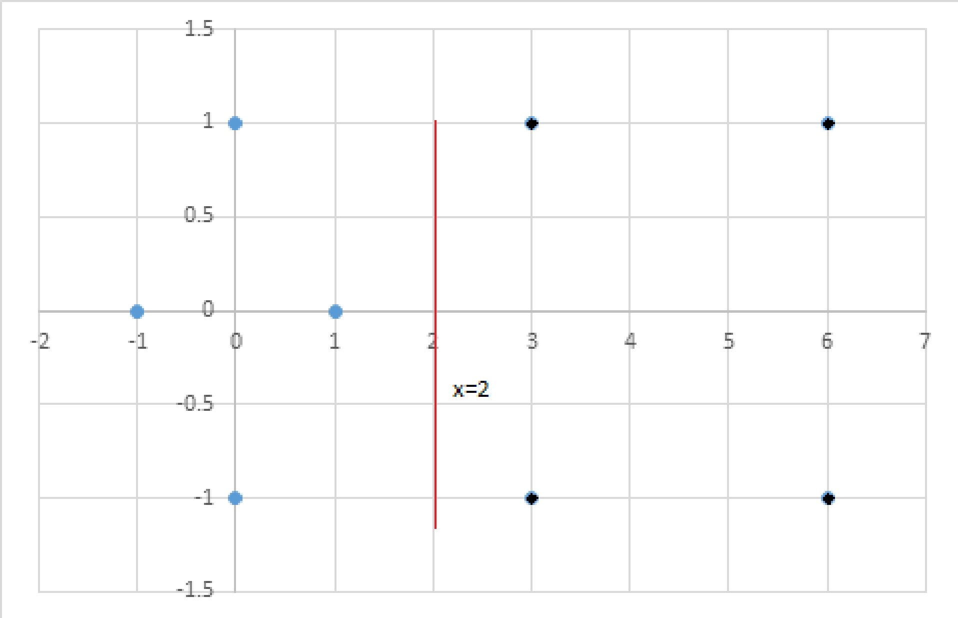
$$\tilde{w}=\sum_i \alpha_i . s_i$$

where i is the number of support vectors

as per the equation $\tilde{w}$= -3.5(1,0,1)+0.75(3,1,1)+0.75(3,-1,1)=(1,0,-2)

hence the hyper plane equation is y=wx+b, where w=(1,0) and b= -2. The marginal hyper plane as per above data is shown in below graph

Author: Alkha Mohan
Mohan.alkha@gmail.com

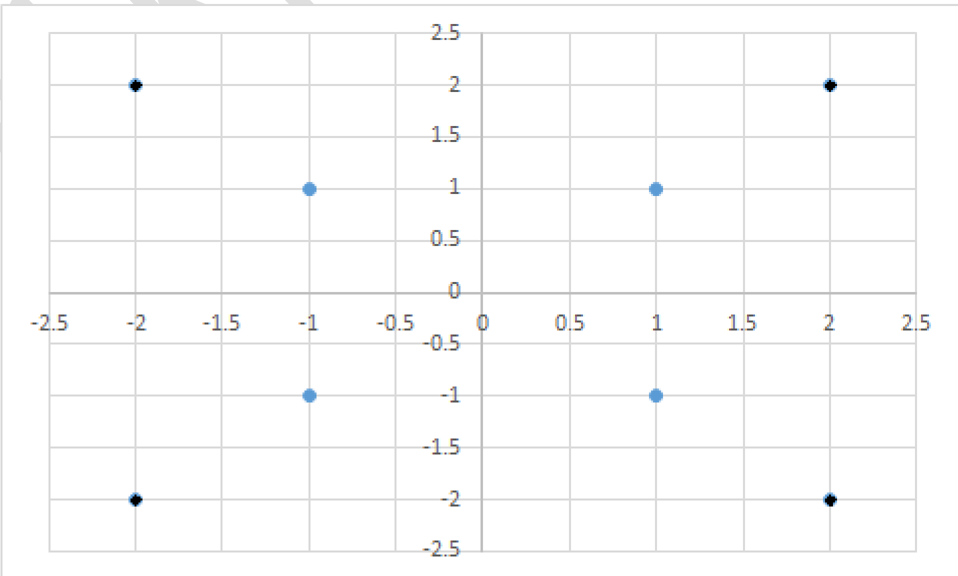


Example Problem- Non Linear SVM

Consider the dataset having values (2,2), (2,-2),(-2,-2),(-2,2) mapped to the positive class and (1,1),(1,-1),(-1,-1),(-1,1) mapped to negative class. Find the best line that separate these points.

Solution

Plot the points to the graph and find the support vectors as given below, black denotes positive class points and blue denotes negative class points.

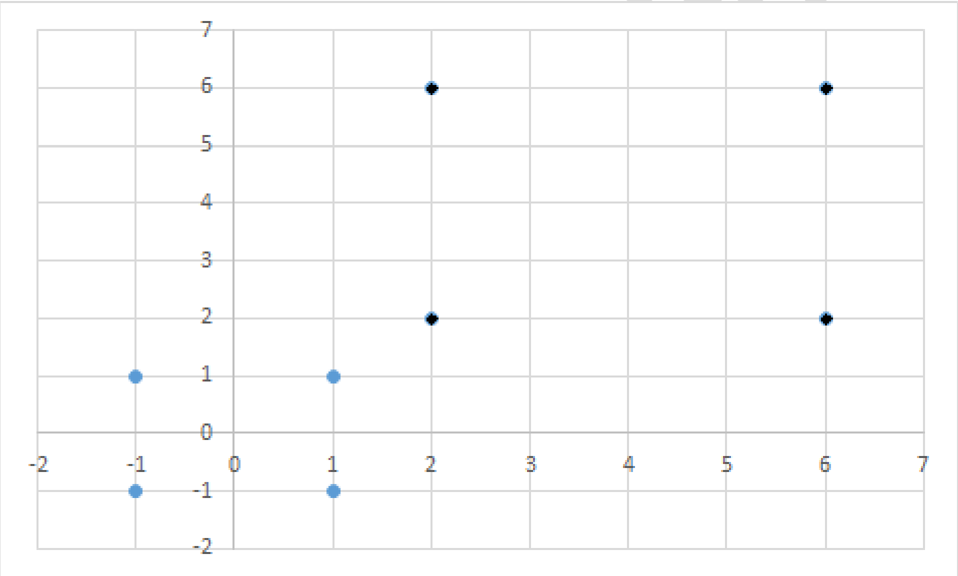


Author: Alkha Mohan
Mohan.alkha@gmail.com

From the figure it is clear that there is no hyperplane that discriminates data points into two classes. We have to use nonlinear svm with a mapping function Φ. Φ is given as below.

$$\Phi_1\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases}\begin{pmatrix}4-x_2+|x_1-x_2|\\4-x_1+|x_1-x_2|\end{pmatrix} & if\ \sqrt{{x_1}^2+{x_2}^2} > 2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & otherwaise\end{cases}$$

The data points are changes to (2,2),(6,2),(6,6),(2,6) and (1,1),(1,-1),(-1,-1),(-1,1). The graph for the below points is given below and found that these are separable.



Support vectors are s1=(1,1) and s2=(2,2)

Now as per the svm equation

α1 Φ(s1) · Φ(s1) + α2 Φ(s2) · Φ(s1) = −1

α1 Φ(s1) · Φ(s2) + α2 Φ(s2) · Φ(s2) = +1

Since the problem is linearly separable Φ()= Identity, so the above equation reduces to

$$\alpha 1\ \tilde{s}1 \cdot \tilde{s}1 + \alpha 2\ \tilde{s}2 \cdot \tilde{s}1 = -1$$

$$\alpha 1\ \tilde{s}1 \cdot \tilde{s}2 + \alpha 2\ \tilde{s}2 \cdot \tilde{s}2 = +1$$

Substitute the values of support vectors and calculating the dot product will result in

$$3\alpha 1 + 5\alpha 2 = -1$$

$$5\alpha 1 + 9\alpha 2 = +1$$

Solving the above equation results in α1= -7, α2=4

The weight vector $\tilde{w}$ is given as

$$\tilde{w} = \sum_i \alpha_i . s_i$$

where i is the number of support vectors

as per the equation $\tilde{w}$= -7(1,1,1)+4(2,2,1) =(1,1,-3)

hence the hyper plane equation is y=wx+b, where w=(1,1) and b= -3. The marginal hyper plane as per above data is shown in below graph