MODULE 13
Support Vector Machines

LESSON 28
Introduction to Support Vector Machines

Keywords: Maximum Margin, Support Vector, Classification, Decision Boundary

Introduction to Support Vector Machines (SVMs)

- **SVM as a Linear Discriminant**
  In a simplistic sense, SVM is based on linear discriminant functions. So, the classification is based on a function of the form $w^tX + b$; w and b are learnt from the training data. In a two-class problem with a positive class and a negative class, for any pattern X from the positive class $w^tX + b > 0$ and $w^tX + b < 0$ for any pattern X from the negative class when the patterns from the two classes are linearly separable.

- **SVM as a Maximum Margin Classifier**
  Consider a two-class problem where the classes are linearly separable. Let the positive class be characterized by a hyperplane $w^tX + b = 1$ and the negative class by a parallel plane $w^tX + b = -1$ so that training patterns, near the margin, from the respective classes fall on these planes. These planes are called the support planes. The decision boundary or the separating plane is characterized by $w^tX + b = 0$. It may be illustrated using the two-dimensional data shown in Figure 1.

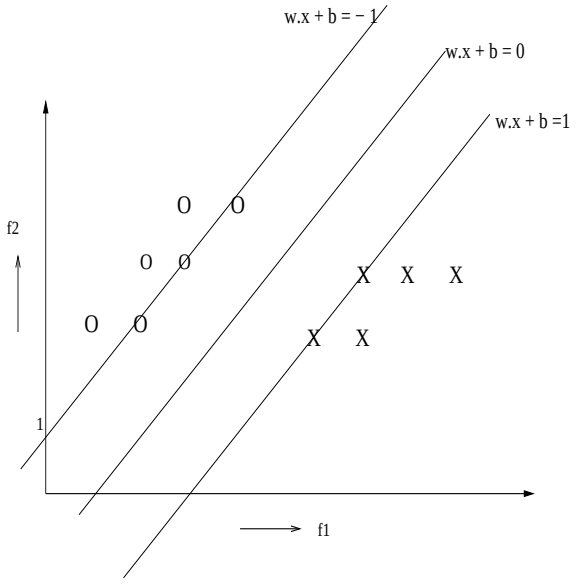

Figure 1: Support Lines Characterizing the Margin

- **Margin Between the two Lines**
  In the two-dimensional case, we have support lines, instead of planes, and the decision boundary also is a line as shown in the figure. Patterns labelled 'O' are from the negative class and those labelled 'X' are from the positive class. As discussed in lesson 26 (refer to equation (7)), the distance from any point X on the line $w^tX + b = 1$ to the decision boundary is given by

$$\frac{f(X)}{|| X ||} = \frac{1}{|| w ||}$$

  In the above, note that $f(X) = 1$ for any X on the line $w^tX + b = 1$. Similarly, the distance from a point on the line $w^tX + b = -1$ to the decision boundary is given by $\frac{-1}{||w||}$. So, the distance between the two supporting lines, or the margin, is $\frac{2}{||w||}$. Observe that this expression for margin holds good even in the d- dimensional ($d > 2$) case.

- **Maximizing the Margin**
  Maximizing the margin is achieved by maximizing $\frac{2}{||w||}$ or equivalently by finding a w that minimizes $\frac{||w||}{2}$ or for the sake of simplicity in calculus, $\frac{||w||^2}{2}$. Optimization is carried out by using constraints of the form $w^tX + b \leq -1$ for all X in the negative class and $w^tX + b \geq 1$ for all X in the positive class.

- **SVM and Data Reduction**
  SVM selects the training patterns falling on the support planes. Such vectors are called *support vectors*. These vectors are adequate to learn both w and b. This amounts to data reduction or compression.

- **Non-linear decision boundaries**
  It is possible that the two classes are not linearly separable. In such a case, the decision boundary can be non-linear. In a generalized sense, a non-linear function can be represented by a linear function. We discuss this issue next.

- **Generality of Linear Discriminants**
  The notion of linear discriminant functions is very general. This idea may be extended to handle even non-linear discriminants using the

homogeneous representation. For example, consider the non-linear decision boundary

$$f_1 \ = \ - \ f_2^2 \ + \ 5f_2 \ - \ 3 \ = 0$$

as shown in Figure 2.

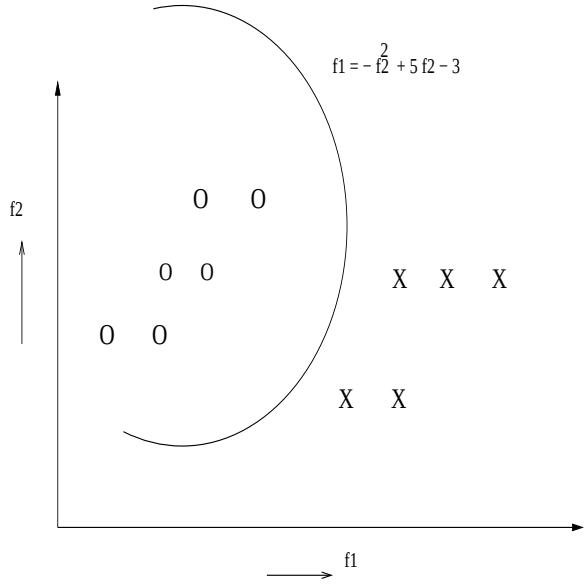

Figure 2: Classification using a Nonlinear Discriminant Function

- **Representing a Non-Linear Function Using a Linear Form**
  Such a non-linear function may be represented using the homogeneous form

$$f(X') \ = \ w^t X' \ = \ 0 \tag{1}$$

where

$$w = \begin{pmatrix} -1 \\ 5 \\ -3 \end{pmatrix}$$

$$X' = \begin{pmatrix} f_2^2 \\ f_2 \\ 1 \end{pmatrix}$$

It may be illustrated using the following example.

- **Learning the generalized linear discriminant function**
  Consider a binary classifier which assigns x to class 'O' (negative class) if $f(x) < 0$ and to class 'X' (positive class) if $f(X) > 0$, where

$$f(x) = a + bx + cx^2 \tag{2}$$

  Observe, based on the discussion above, that equivalently we can assign X to 'O' if $w^t X' < 0$ and to class 'X' if $w^t X' > 0$, where

$$w = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$$X' = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}$$

  Let us consider a set of labelled patterns that are not linearly separable. Specifically, let us consider the one-dimensional data set shown in Table 1. Observe that the data is not linearly separable. Further, the

Pattern No.	x	class
1	-1	'O'
2	-2	'O'
3	0	'O'
4	1	'X'
5	3	'X'
6	5	'O'

Table 1: Non-Linearly Separable Classes

  decision boundary is characterized by $f(x) = 0$. Now by appropriate transformation and normalization, where the components in X' are 1, value of x, and value of x^2, we get the data shown in Table 2. We can use the perceptron learning algorithm to obtain the weight vector w.

Pattern No.	1	x	x^2
1	-1	1	-1
2	-1	2	-4
3	-1	0	0
4	1	1	1
5	1	3	9
6	-1	-5	-25

Table 2: Normalized non-linearly separable data

1. Here, we start with
$$w_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$
and
$$x_1' = \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}$$
w_1 misclassifies X_1'. So, gets updated to get $w_2 = w_1 + X_1'$.
2. Continuing with the algorithm, we end up with w_{70} which classifies all the 6 patterns in Table 2 correctly. It is given by
$$w_{70} = \begin{pmatrix} -1 \\ 35 \\ -11 \end{pmatrix}$$

So, the decision boundary is given by

$$f(x) \ = -1 + 35x - 11x^2 = 0 \tag{3}$$

This example illustrates the generality of the linear discriminants. It is possible to deal with non-linearly separable classes using a linear discriminant. Further, it is possible to extend this idea to vector-valued X also.

- **SVM for Classification**
Support vector machine (SVM) generates an abstraction in the form

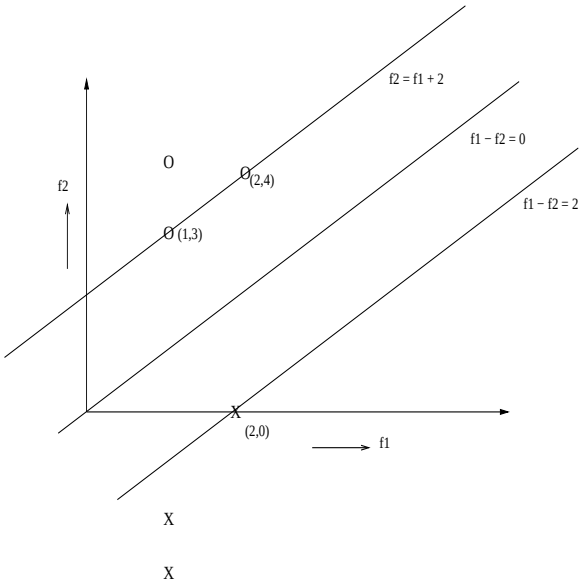

Figure 3: Illustration of Support Vector Machines

of a linear discriminant. It also selects a set of vectors called support vectors which are margin patterns and are members of the training set. We illustrate this using Figure 3.

- **Support Vectors**
  Consider three of the points shown in Figure 3. These are from two classes ('X' and 'O'). Here, $(1,3)^t$ and $(2,4)^t$ are from class 'O' and $(2,0)^t$ is from class 'X'. The lines $f_2 - f_1 = 2$ and $f_1 - f_2 = 2$ characterize the boundaries of the classes 'O' and 'X' respectively. These lines are the **support lines** and the points are the **support vectors**. These three support vectors are adequate to characterize the classifier.

- **Adequacy of support vectors**
  Now consider adding points $(1,4)^t$, $(1,5)^t$ and $(2,5)^t$ from class 'O' and $(2,-1)^t$, $(1,-2)^t$, and $(1,-3)^t$ from class 'X'. They are properly classified using the support lines (in turn using the support vectors). The region between the two lines is the margin and because the support lines correspond to maximum value of the margin, they are as far away

from each other as possible.

- **Decision Boundary**
  The line $f_1 - f_2 = 0$ is equidistant from the two decision lines and it forms the right choice for the decision boundary between the two classes. Here, points satisfying the property that $f_1 - f_2 < 0$ are classified as members of class 'O' and those satisfying the condition $f_1 - f_2 > 0$ are of class 'X'.

- **Properties of the SVM Classifier**
  From the above discussion, we make the following observations.

  1. SVM may be viewed as a binary (two class) classifier. It abstracts a linear decision boundary from the data and uses it to classify patterns belonging to the two classes.
  2. Support vectors are some of the vectors falling on the support planes in a d-dimensional space.
  3. SVMs learn, from the data, linear discriminants of the form $w^t + b$ that corresponds to the maximum margin.
  4. When the two classes are linearly separable, it is easy to compute the classifier characterizing the maximum margin.

- **Linearly Separable Case**
  Consider again the points $(1, 3)^t$ and $(2, 4)^t$ from class 'O' and $(2, 0)^t$ from class 'X' shown in Figure 3. They are linearly separable. In fact, we can draw several (possibly infinite) lines separating correctly the two classes represented by these three points. The corresponding normalized data is shown in Table 3.

Pattern No.	$feature_1$	$feature_2$	bias
1	-1	-3	-1
2	-2	-4	-1
3	2	0	1

Table 3: Normalized set of patterns in three dimensions

- **Which Linear Discriminant?**
  Using Perceptron learning algorithm, we get the decision boundary characterized by $f_1 - 3f_2 = 0$. The initial weight vector, $w_1 = 0$ misclassifies $(-1, -3, -1)^t$. So, $w_2 = (-1, -3, -1)^t$. It misclassifies $(2, 0, 1)^t$. So, $w_3 = (1, -3, 0)^t$. Note that w_3 classifies all the three patterns properly and it corresponds to the decision boundary $f_1 - 3f_2 = 0$.

- **Decision Boundary of the SVM**
  In the case of the SVM, we choose that line as decision boundary which provides maximum margin. When the patterns are linearly separable, we can get the linear decision boundary corresponding to the maximum margin solution. For example, consider the three patterns in Figure 3. In this two-dimensional example, two points of a class are adequate to characterize the support line. So, for the two points $(1, 3)^t$ and $(2, 4)^t$ from class 'O', the support line is $f_2 - f_1 = 2$. Now consider a line parallel to this support line and passing through the point $(2, 0)^t$ from class 'X'; this is $f_1 - f_2 = 2$ and it is the support line for class 'X'. These two lines, namely $f_2 - f_1 = 2$ and $f_1 - f_2 = 2$ characterize the margin. So, the decision boundary which is equidistant from these two lines is characterized by $f_1 - f_2 = 0$.

- **Learning the SVM**
  Consider two points $(2, 2)^t$ and $(1, 1)^t$ on the line $f_1 - f_2 = 0$. Because these points are on the decision boundary, we need $w = (w_1, w_2)^t$ to satisfy
  $2w_1 + 2w_2 + b = 0$ (corresponding to $(2, 2)^t$) and
  $w_1 + w_2 + b = 0$ (for the point $(1, 1)^t$). From these two equations, we get $w_1 + w_2 = 0$; so, w is of the form $(\alpha, -\alpha)^t$ and correspondingly, $b = 0$. In general, α is any constant. However, it is good to choose α in a normalized manner. For example, it is convenient to choose α such that
  $w^t X + b = 0$, or equivalently, $w^t X = 0$ for all the points on the decision boundary;
  $w^t X = 1$ for all the points on the support line $f_1 - f_2 = 2$; and
  $w^t X = -1$ for all the points on the support line $f_2 - f_1 = 2$
  This can be achieved by choosing a value of $\frac{1}{2}$ for α. So, correspondingly $w = (\frac{1}{2}, -\frac{1}{2})^t$ and $b = 0$.