

Chapter 2

Linear Discriminant Function

Abstract Linear discriminant functions (LDFs) have been successfully used in pattern classification. Perceptrons and Support Vector Machines (SVMs) are two well-known members of the category of *linear discriminant functions* that have been popularly used in classification. In this chapter, we introduce the notion of linear discriminant function and some of the important properties associated with it.

Keywords Linear classifier · Decision boundary · Linear separability · Nonlinear discriminant function · Linear discriminant function · Support vector machine · Perceptron

2.1 Introduction

We have seen in *Introduction* that a linear discriminant function $g(X)$ can be used as a classifier. The specific steps involved are as follows:

1. Consider a functional form for $g(X)$.
2. Using the two-class training data, learn $g(X)$. By learning $g(X)$ we mean obtaining the values of the coefficients of terms in $g(X)$.
3. Given a *test pattern* X_{test} , compute $g(X_{test})$. Assign X_{test} to C_- if $g(X_{test}) < 0$ else (if $g(X_{test}) > 0$) assign it to C_+ .

2.1.1 Associated Terms [1–3]

We explain the associated concepts next

- **Training Dataset:**

The training dataset or training set, \mathcal{X}_{train} , is a finite set given by

$$\mathcal{X}_{train} = \{(X_1, C_1), (X_2, C_2), \dots, (X_n, C_n)\}$$

where X_i is the i th pattern (representation) given by $X_i = (x_{i1}, x_{i2}, \dots, x_{il})$ for some finite l .

Even though it is possible to have more than two classes, we consider only two-class (*binary*) classification problems in this chapter. We will examine how to build a multiclass classifier based on a combination of binary classifiers later. So, Associated with pattern X_i is its class label C_i where $C_i \in \{C_-, C_+\}$.

- **Test Pattern:**

A test pattern, X_{test} or simply X is an l -dimensional pattern which is not yet labeled.

- **Classifier:**

A classifier *assigns a class label to a test/unlabeled pattern*.

We illustrate these notions with the help of a two-dimensional dataset shown in Fig. 2.1. We depict in the figure, a set of children and a set of adults. Each *child* is depicted using C and each *adult* using A . In addition there are four test patterns X_1, X_2, X_3 , and X_4 . Each pattern is represented by its *Height* and *Weight*.

In Fig. 2.1 three classifiers are shown, a *decision tree classifier*, an *LDF based classifier*, and a *nonlinear discriminant based classifier*.

Each of the three classifiers in the figure belongs to a different category. Here,

- The *Linear discriminant/classifier* depicted by the thin broken line is a *linear classifier*. Any point X falling on the left side of the line (or $g(X) < 0$) is a *child* and

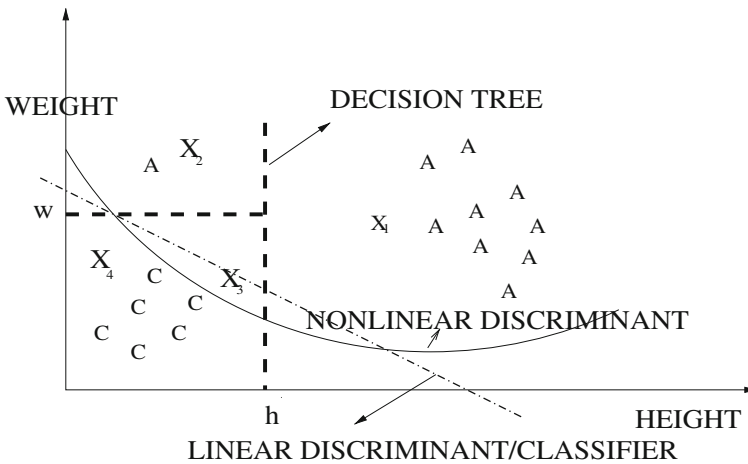


Fig. 2.1 An example dataset

a point X to the right (or $g(X) > 0$) is classified as *adult*.

- The *Nonlinear discriminant* shown by the curved line in the figure corresponds to a *nonlinear classifier*. An X such that $g(X) < 0$ is assigned the label *child*. If $g(X) > 0$, then X is assigned *adult*.
- The *decision tree* classifier depicted by the piecewise linear region in the figure is not linear and it could be called a *piecewise linear classifier*. It may be described by

Adult : $(HEIGHT > h) \vee [(HEIGHT < h) \wedge (WEIGHT > w)]$.

Child : $(HEIGHT < h) \wedge (WEIGHT < w)$.

In this simple case, test patterns X_1 and X_2 are assigned to class *Adult* or equivalently X_1 and X_2 are assigned the class label *Adult* by all the three classifiers.

Similarly, test pattern X_4 is assigned the label *child* by all the three classifiers. However, X_3 is assigned the label *adult* by the nonlinear discriminant-based classifier and the other two classifiers assign X_3 to class *child*.

It is possible to extend these ideas to more than two-dimensional spaces. In high-dimensional spaces,

- the linear discriminant is characterized by a *hyperplane* instead of a line as in the two-dimensional case.
- the nonlinear discriminant is characterized by a *manifold* instead of a curve.
- the piecewise linear discriminant characterizing the decision tree classifier continues to be piecewise linear discriminant, perhaps involving a larger size conjunction. So, learning a decision tree classifier in high-dimensional spaces could be computationally prohibitive.

However, it is possible to classify X based on the value of $g(X)$ irrespective of the dimensionality of X (or the value of l). This needs obtaining an appropriate $g(X)$. In this chapter, we will concentrate on *linear classifiers*.

2.2 Linear Classifier [2–4]

A *linear classifier* is characterized by a *linear discriminant function* $g(X) = W^t X + b$, where $W = (w_1, w_2, \dots, w_l)^t$ and $X = (x_1, x_2, \dots, x_l)^t$. We assume without loss of generality that W and $X \in \mathbb{R}^l$ and $b \in \mathbb{R}$.

Note that both the components of W and X are in linear form in $g(X)$. It is also possible to express $g(X)$ as

$$g(X) = b + \sum_{i=1}^l w_i x_i$$

If we augment X and W appropriately and convert them into $l + 1$ dimensional vectors, we can have a more acceptable and simpler form for $g(X)$. The *augmented* form is given by $X_a = (1, x_1, \dots, x_l)^t$ and $W_a = (b, w_1, \dots, w_l)^t$, where X_a and W_a are augmented versions of X and W , respectively. Note that both X_a and W_a are $l + 1$ dimensional vectors.

Now

$$g(X) = W_a^t X_a = b + \sum_{i=1}^l w_i x_i$$

If we use the augmented vectors, then $g(X)$ satisfies the two properties of *linear systems* as shown below.

- **Homogeneity:** For $c \in \mathbb{R}$, $g(cX) = cg(X)$

$$g(cX) = W_a^t (cX_a) = cW_a^t X_a = cg(X)$$

- **Additivity:** For X_1 and $X_2 \in \mathbb{R}^l$, $g(X_1 + X_2) = g(X_1) + g(X_2)$

$$g(X_1 + X_2) = W_a^t (X_{1a} + X_{2a}) = W_a^t X_{1a} + W_a^t X_{2a} = g(X_1) + g(X_2).$$

Note that if W and X are used in their l -dimensional form, then *homogeneity* and *additivity* are not satisfied. However, *convexity* is satisfied as shown below.

- **Convexity:** For some $\alpha \in [0, 1]$, $g(\alpha X_1 + (1 - \alpha)X_2) \leq \alpha g(X_1) + (1 - \alpha)g(X_2)$

$$g(\alpha X_1 + (1 - \alpha)X_2) = b + W^t (\alpha X_1 + (1 - \alpha)X_2)$$

$$= \alpha b + (1 - \alpha)b + \alpha W^t X_1 + (1 - \alpha)W^t X_2$$

$$= \alpha(b + W^t X_1) + (1 - \alpha)(b + W^t X_2) = \alpha g(X_1) + (1 - \alpha)g(X_2)$$

- **Classification of augmented Vectors using W_a :**

We will illustrate classification of patterns using the augmented representations of the six patterns shown in Fig. 1.3. We show the augmented patterns in Table 2.1 along with these value of $W_a^t X_a$ for $W_a = (-14, 1, 5)^t$.

Table 2.1 Classification of augmented patterns using $W_a = (-14, 1, 5)^t$

Pattern number	Class label	1	x_1	x_2	$W_a^t X_a$
1	–	1	1	1	–8
2	–	1	2	2	–2
3	+	1	2	3	3
4	+	1	6	2	2
5	+	1	7	2	3
6	+	1	7	3	8

2.3 Linear Discriminant Function [2]

We have seen earlier in this chapter that a linear discriminant function is of the form $g(X) = W^t X + b$ where W is a column vector of size l and b is a scalar. $g(X)$ divides the space of vectors into three parts. They are

2.3.1 Decision Boundary

In the case of linear discriminant functions, $g(x) = W^t X + b = 0$ characterizes the *hyperplane* (line in a two-dimensional case) or the *decision boundary*. The decision boundary corresponding to $g(X)$ (DB_g) could also be viewed as

$$DB_g = \{X | g(X) = 0\}$$

2.3.2 Negative Half Space

This may be viewed as the set of all patterns that belong to C_- . Equivalently, the negative half space corresponding to $g(X)$ (NHS_g) is the set

$$NHS_g = \{X | g(X) < 0\} = C_-$$

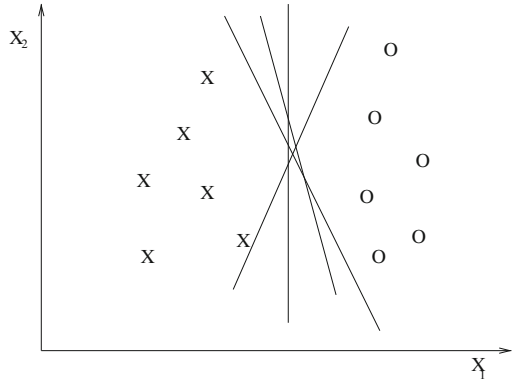
2.3.3 Positive Half Space

This is the set of all patterns belonging to C_+ . Equivalently, the positive half space corresponding to $g(X)$ (PHS_g) is given by

$$PHS_g = \{X | g(X) > 0\} = C_+$$

Note that each of these parts is a potentially infinite set. However, the training dataset and the collection of test patterns that one encounters are finite.

Fig. 2.2 Linearly separable dataset



2.3.4 Linear Separability

Let \mathcal{X} be a set of labeled patterns given by

$$\mathcal{X} = \{(X_1, C_1), (X_2, C_2), \dots, (X_n, C_n)\}.$$

We say the set \mathcal{X} is *linearly separable* if there is a W and b such that $W^T X_i + b > 0$ if $C_i = C_+$ and $W^T X_i + b < 0$ if $C_i = C_-$ for $i = 1, 2, \dots, n$.

We can think of employing linear classifiers when the samples/set of patterns is linearly separable. Consider the two-dimensional patterns shown in Fig. 2.2. They are linearly separable. If they are linearly separable, then we can have infinite number of LDFs associated as shown in the figure.

2.3.5 Linear Classification Based on a Linear Discriminant Function

A linear classifier is *abstracted* by the corresponding *ldf*, $g(X) = W^T X + b$. The three regions associated with $g(X)$ are important in appreciating the classifier as shown in Fig. 2.3.

1. **The decision boundary** or the hyperplane associated with $g(X)$ is the *separator* between the two classes, the *negative* and *positive* classes. Any point X on the decision boundary satisfies $g(X) = 0$.

If X_1 and X_2 are two different points on the decision boundary, then

$$W^T X_1 + b = W^T X_2 + b = 0 \Rightarrow W^T (X_1 - X_2) = 0.$$

This means W is **orthogonal** to $(X_1 - X_2)$ or the line joining the two points X_1 and X_2 or the decision boundary. So, W is *orthogonal to the Decision boundary*.

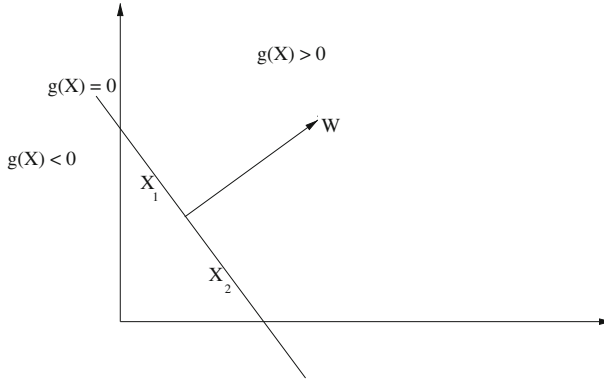


Fig. 2.3 Three regions associated with $g(X) = W^T X + b$

This means that there is a natural association between W and the decision boundary; in a sense if we specify one, the other gets fixed.

2. **The Positive Half Space:** Any pattern X in this region satisfies the property that $g(X) = W^T X + b > 0$. We can interpret it further as follows:

- a. **Role of b :** We can appreciate the role of b by considering the value of $g(X)$ at the *origin*. Let $b > 0$ and X is the *origin*. Then $g(0) = W^T 0 + b = 0 + b = b > 0$. So, at the origin 0, $g(0) > 0$; hence the origin 0 is in the positive half space or PHS_g .

If $b > 0$, then the origin is in the positive half space of $g(X)$.

Now consider the situation where $b = 0$. So, $g(X) = W^T X + b = W^T X$. If X is at the origin, then $g(X) = g(0) = W^T 0 = 0$. So, the origin satisfies the property that $g(X) = 0$ and hence it is on the decision boundary.

So, if $b = 0$, then the origin is on the decision boundary.

- b. **Direction of W :** Consider an LDF $g(X)$ where $b = 0$. Then $g(X) = W^T X$. If X is in the positive half space, then $g(X) = W^T X > 0$. We have already seen that W is orthogonal to the decision boundary $g(X) = 0$. Now we will examine whether W is oriented toward the positive half space or the negative half space.

If $b = 0$ and X is in the positive half space, then $g(X) = W^T X > 0$. Now relate $W^T X$ with the cosine of the angle between W and X . We have

$$\text{cosine}(W, X) = \frac{W^T X}{\|W\| \|X\|} \Rightarrow W^T X = \text{cosine}(W, X) \|W\| \|X\|.$$

So, given that $W^T X > 0$, we have $\text{cosine}(W, X) \|W\| \|X\| > 0$

We know that $\|W\| > 0$ and $\|X\| > 0$. So,

$$\text{cosine}(W, X) > 0.$$

This can happen when the angle, θ , between W and X is such that $-90 < \theta < 90$ which can happen when W is pointing toward the positive half space as X is in the positive half space.

3. **The Negative Half Space:** Any point X in the negative half space is such that $g(X) < 0$. Again if we let $b = 0$ and consider a pattern, X , in the negative class, then $W^T X < 0$. This means the angle, θ , between X and W is such that $90 < \theta < 270$. This also ratifies that W points toward the positive half space.

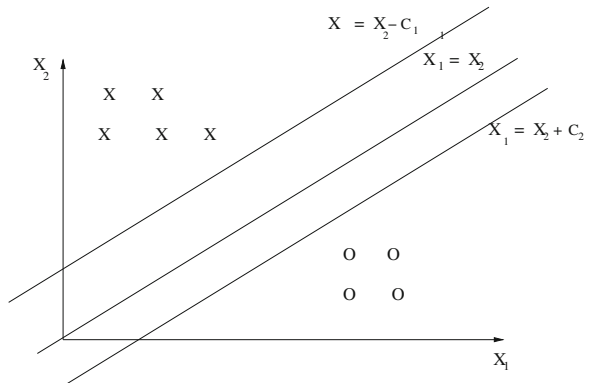
Further, note that for $b < 0$ and X in the negative half space, $g(X) = W^T X + b < 0$ and evaluated at the origin, $g(0) = W^T 0 + b = b < 0$. So, if $b < 0$, then the origin is in the negative half space.

So, the roles of W and b in the LDF $g(X) = W^T X + b$ are given by

- *The value of b decides the location of the origin.* The origin is in the PHS_g if $b > 0$; it is in the NHS_g if $b < 0$ and the origin is on the decision boundary if $b = 0$. It is illustrated in Fig. 2.4

Note that there are patterns from two classes and the samples are linearly separable. There are three linear discriminant functions with different b values and correspondingly the origin is in the negative space in one case ($x_1 = x_2 - C_1$), on the decision boundary in the second case ($x_1 = x_2$) and it is in the positive space

Fig. 2.4 Three decision boundaries with same W



in the third ($x_1 = x_2 + C_2$). However, W is the same for all the three functions as the decision boundaries are all parallel to each other.

- W is orthogonal to the decision boundary and it points toward the positive half space of g as shown in Fig. 2.3.

2.4 Example Linear Classifiers [2]

It is possible to show that the *MDC*, *Naïve Bayes* classifier and others are linear classifiers. Consider

2.4.1 Minimum-Distance Classifier (MDC)

In the case of MDC we assign X to C_- if

$$\|X - m_-\|^2 < \|X - m_+\|^2 \Rightarrow X^T X + m_-^T m_- - 2m_-^T X < X^T X + m_+^T m_+ - 2m_+^T X.$$

We can simplify by canceling $X^T X$ that is common to both sides and bringing all the terms to the left-hand side, we get

$$\text{assign } X \text{ to } C_- \text{ if } (m_+ - m_-)^T X + \frac{1}{2}(m_-^T m_- - m_+^T m_+) < 0.$$

This is the same as assigning X to C_- if $W^T X + b < 0$ where

$$W = (m_+ - m_-) \text{ and } b = \frac{1}{2}(m_-^T m_- - m_+^T m_+).$$

So, MDC is a linear classifier characterized by an LDF of the form $W^T X + b$.

2.4.2 Naïve Bayes Classifier (NBC)

In the case of NBC, we have

$$P(C_-|X) = \prod_{i=1}^l P(x_i|C_-)P(C_-)$$

and

$$P(C_+|X) = \prod_{i=1}^l P(x_i|C_+)P(C_+)$$

We assign X to C_- if $P(C_-|X) > P(C_+|X)$ or equivalently when

$$\prod_{i=1}^l P(x_i|C_-)P(C_-) > \prod_{i=1}^l P(x_i|C_+)P(C_+).$$

By applying logarithm both sides and rearranging terms, we have

$$\sum_{i=1}^l n_i \log \frac{P(x_i|C_-)}{P(x_i|C_+)} + \log \frac{P(C_-)}{P(C_+)} > 0$$

where n_i is the number of times the feature x_i occurred in X . If X is a binary pattern, then n_i is either 1 or 0. If X is a document, then n_i is the number of times term x_i occurred in X .

So, we assign X to C_- if

$$\sum_{i=1}^l w_i n_i + b > 0$$

where

$$w_i = \log \frac{P(x_i|C_-)}{P(x_i|C_+)}, \quad b = \log \frac{P(C_-)}{P(C_+)}.$$

So, *Naïve Bayes Classifier* is a linear classifier.

2.4.3 Nonlinear Discriminant Function

It is possible to view a nonlinear discriminant function as a linear discriminant function in a higher dimensional space. For example, consider the two-dimensional dataset of six patterns shown in Fig. 1.6.

We have seen that a nonlinear discriminant function given by $x_1^2 + 32x_2 - 76$ can be used to classify the six patterns.

Here, X is a two-dimensional column vector given by $X = (x_1, x_2)^t$. However, if we map it to a six-dimensional representation given by $\phi(X) = (1, x_1, x_2, x_1^2, x_2^2, x_1x_2)^t$ where

$$\phi_1(X) = 1, \quad \phi_2(X) = x_1, \quad \phi_3(X) = x_2, \quad \phi_4(X) = x_1^2, \quad \phi_5(X) = x_2^2, \quad \phi_6(X) = x_1x_2.$$

So, ϕ is a mapping from \mathbb{R}^2 to \mathbb{R}^6 such that

$$\phi : (x_1, x_2)^t \rightarrow (1, x_1, x_2, x_1^2, x_2^2, x_1x_2)^t.$$

Then the nonlinear discriminant function $x_1^2 + 32x_2 - 76$ in \mathbb{R}^2 is linear in \mathbb{R}^6 corresponding to the $\phi(X)$ space.

If we choose $W = (-76, 0, 32, 1, 0, 0)$ then, $g(X) = W^t\phi(X)$ which is a linear discriminant function in $\phi(X)$.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
2. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley (1970)
3. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (2013)
4. Zhao, W., Chellappa, R., Nandhakumar, N.: Empirical performance analysis of linear discriminant classifiers, In: Proceedings of Computer Vision and Pattern Recognition, 25–28 June 1998, pp. 164–169. Santa Barbara, CA, USA (1998)



<http://www.springer.com/978-3-319-41062-3>

Support Vector Machines and Perceptrons
Learning, Optimization, Classification, and Application to
Social Networks

Murty, M.N.; Raghava, R.

2016, XIII, 95 p. 25 illus., Softcover

ISBN: 978-3-319-41062-3