GENERALIZED DECISION FUNCTIONS

Classes which do not share a single pattern may always be separated. However, decision boundaries which separate between classes, may not always be linear (see Fig. 2.1.2). The complexity of these boundaries may sometimes request the use of highly nonlinear surfaces. A popular approach to generalize the concept of linear decision functions is to consider a *generalized decision function* defined as

$$d(\boldsymbol{x}) = w_1 f_1(\boldsymbol{x}) + \ldots + w_N f_N(\boldsymbol{x}) + w_{N+1} \tag{2.3.1}$$

where $f_i(\boldsymbol{x}),\ 1 \le i \le N$ are scalar functions of the pattern $\boldsymbol{x},\ \boldsymbol{x} \in R^n$. Introducing $f_{N+1}(\boldsymbol{x}) = 1$ we get

$$d(\boldsymbol{x}) = \sum_{i=1}^{N+1} w_i f_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}^* \tag{2.3.2}$$

where

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \\ w_{N+1} \end{pmatrix}, \quad \boldsymbol{x}^* = \begin{pmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_N(\boldsymbol{x}) \\ f_{N+1}(\boldsymbol{x}) \end{pmatrix} \tag{2.3.3}$$

The representation of $d(\boldsymbol{x})$ by Eqs. (2.3.2) and (2.3.3) implies that *any* decision function defined by Eq. (2.3.1) can be treated as linear, provided that we first transform *all* the original patterns $\boldsymbol{x}$ into $\boldsymbol{x}^*$ by calculating $f_i(\boldsymbol{x})$, $1 \le i \le N$ for every individual $\boldsymbol{x}$. Although $d(\boldsymbol{x})$ is linear in the $(N+1)-$ dimensional space whose dimension $N+1$ is usually considerably greater than n, it certainly maintains its nonlinearity characteristics in R^n.

As expected, the most commonly used generalized decision function is $d(\boldsymbol{x})$ for which $f_i(\boldsymbol{x}),\ 1 \le i \le N$ are polynomials. If these functions are all linear in R^n, then $d(\boldsymbol{x})$ can be rewritten as

$$d(\boldsymbol{x}) = (\boldsymbol{w}^*)^T \boldsymbol{x} \tag{2.3.4}$$

where $\boldsymbol{w}^*$ is a new weight vector, which can be calculated from the original $\boldsymbol{w}$ and the original linear $f_i(\boldsymbol{x}),\ 1 \le i \le N$ in Eq. (2.3.1). The expression in Eq. (2.3.4) is identical to that in Eq. (2.1.2) from the previous section.

Let us now consider quadratic decision functions. For 2-dimensional patterns (i.e. $n = 2$), the most general decision function is

$$d(\boldsymbol{x}) = w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_2^2 + w_4 x_1 + w_5 x_2 + w_6 \tag{2.3.5}$$

i.e. $\boldsymbol{w} = (w_1, w_2, \ldots, w_6)^T$ and $\boldsymbol{x}^* = (x_1^2, x_1 x_2, x_2^2, x_1, x_2, 1)^T$. For patterns $\boldsymbol{x} \in R^n$ the most general quadratic decision function is given by

$$d(\boldsymbol{x}) = \sum_{i=1}^{n} w_{ii} x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} x_i x_j + \sum_{i=1}^{n} w_i x_i + w_{n+1} \tag{2.3.6}$$

The number of terms at the right-hand side of Eq. (2.3.6) is

$$l = N + 1 = n + \frac{n(n-1)}{2} + n + 1 = \frac{(n+1)(n+2)}{2} \tag{2.3.7}$$

This is the total number of weights which are the free parameters of the problem. If for example $n = 3$, the vector $\boldsymbol{x}^*$ is 10-dimensional. For $n = 10$ we already have a considerably large $N = 65$.

In the case of polynomial decision functions of order m, a typical $f_i(\boldsymbol{x})$ in Eq. (2.3.1) is given by

$$f_i(\boldsymbol{x}) = x_{i_1}^{e_1} \; x_{i_2}^{e_2} \; \ldots \; x_{i_m}^{e_m} \tag{2.3.8}$$

where $1 \leq i_1, i_2, \ldots, i_m \leq n$ and e_i, $1 \leq i \leq m$ is 0 or 1. It is clearly a polynomial with a degree between 0 and m. To avoid repetitions we request $i_1 \leq i_2 \leq \ldots \leq i_m$.

■ **Theorem 2.3.1.** Let $d^m(\boldsymbol{x})$ denote the most general polynomial decision function of order m. Then

$$d^m(\boldsymbol{x}) = \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \cdots \sum_{i_m=i_{m-1}}^{n} w_{i_1 i_2 \ldots i_m} x_{i_1} x_{i_2} \ldots x_{i_m} + d^{m-1}(\boldsymbol{x}) \tag{2.3.9}$$

where $d^0(\boldsymbol{x}) = w_{n+1}$.

The proof using mathematical induction is straightforward.

■ **Example 2.3.1** Let $n = 3$ and $m = 2$. Then

$$d^2(\boldsymbol{x}) = \sum_{i_1=1}^{3} \sum_{i_2=i_1}^{3} w_{i_1 i_2} x_{i_1} x_{i_2} + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4$$

$$= w_{11} x_1^2 + w_{12} x_1 x_2 + w_{13} x_1 x_3 + w_{22} x_2^2 + w_{23} x_2 x_3 + w_{33} x_3^2$$

$$+ w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4$$

■ **Example 2.3.2** Let $n = 2$ and $m = 3$. Then

$$d^3(\boldsymbol{x}) = \sum_{i_1=1}^{2} \sum_{i_2=i_1}^{2} \sum_{i_3=i_2}^{2} w_{i_1 i_2 i_3} x_{i_1} x_{i_2} x_{i_3} + d^2(\boldsymbol{x})$$

$$= w_{111} x_1^3 + w_{112} x_1^2 x_2 + w_{122} x_1 x_2^2 + w_{222} x_2^3 + d^2(\boldsymbol{x})$$

$$d^2(\boldsymbol{x}) = \sum_{i_1=1}^{2} \sum_{i_2=i_1}^{2} w_{i_1 i_2} x_{i_1} x_{i_2} + d^1(\boldsymbol{x})$$

$$= w_{11}x_1^2 + w_{12}x_1x_2 + w_{22}x_2^2 + w_1x_1 + w_2x_2 + w_3$$

♠

The number of terms needed to represent a general quadratic decision function is $\frac{(n+1)\,(n+2)}{2}$ where n is the original patterns space's dimension. It can be shown, that in the case of order m, this number is

$$M(n,m) = \binom{n+m}{m} = \frac{(n+m)!}{n!m!} \quad (2.3.10)$$

2.4 GEOMETRICAL DISCUSSION

Since linear decision functions play a significant role in pattern recognition, it is essential to provide a complete geometrical interpretation of their properties. Such an interpretation which includes the concepts of hyperplanes and dichotomies is given below.

2.4.1 Hyperplanes

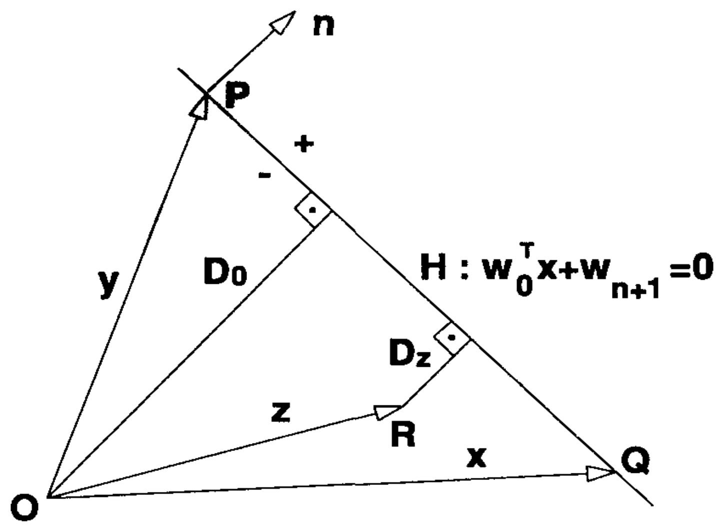
Let R^n be the original patterns' space and consider a two-class or a multiclass problem. A linear decision function which separates one class from another, is determined by an equation such as

$$d(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + w_{n+1} = 0 \qquad (2.4.1)$$

which defines a *linear decision boundary*. The linear decision function itself, is the left-hand side of Eq. (2.4.1). For $n = 2$, the linear decision boundary is a straight line. It is a plane for $n = 3$ and a *hyperplane* for $n > 3$. The vector form of Eq. (2.4.1) is

$$d(\boldsymbol{x}) = \boldsymbol{w}_0^T \boldsymbol{x} + w_{n+1} = 0 \tag{2.4.2}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ and $\boldsymbol{w}_0 = (w_1, w_2, \ldots, w_n)^T$.



■ **Figure 2.4.1** Basic properties of hyperplane.

Consider now the hyperplane H of Eq. (2.4.2) as shown in Fig. 2.4.1. Let n be a unit normal vector at some point P of H, pointing to its positive side. Let $y = \overrightarrow{OP}$ and let $x = \overrightarrow{OQ}$ denote any arbitrary point on the hyperplane. Then, the equation of the hyperplane can be rewritten as

$$n^T \cdot \overrightarrow{QP} = n^T \cdot (x - y) = 0 \tag{2.4.3}$$

or as

$$n^T x = -n^T y \tag{2.4.4}$$

To compare with Eq. (2.4.2) we normalize the previous equation and divide it by

$$\|\boldsymbol{w}_0\| = (w_1^2 + w_2^2 + \ldots + w_n^2)^{1/2}$$

to get

$$\frac{\boldsymbol{w}_0^T \boldsymbol{x}}{\|\boldsymbol{w}_0\|} = -\frac{w_{n+1}}{\|\boldsymbol{w}_0\|} \tag{2.4.5}$$

Since Eqs. (2.4.4) and (2.4.5) represent the same hyperplane and since $\boldsymbol{n}$ and $\boldsymbol{w}_0/\|\boldsymbol{w}_0\|$ are unit vectors, we must have either $\boldsymbol{n} = \boldsymbol{w}/\|\boldsymbol{w}_0\|$ or $\boldsymbol{n} = -\boldsymbol{w}_0/\|\boldsymbol{w}_0\|$. But $\boldsymbol{n}$ was chosen to point to the positive side of the hyperplane, implying

$$\boldsymbol{w}_0^T(\boldsymbol{y}+\boldsymbol{n})+w_{n+1}>0 \tag{2.4.6}$$

and since $\boldsymbol{w}_0^T\boldsymbol{y}+w_{n+1}=0$ we get $\boldsymbol{w}_0^T\boldsymbol{n}>0$. Therefore

$$\boldsymbol{n}=\frac{\boldsymbol{w}_0}{\|\boldsymbol{w}_0\|} \tag{2.4.7}$$

and consequently, by virtue of Eqs. (2.4.4) and (2.4.5)

$$\boldsymbol{n}^T\boldsymbol{y}=\frac{-w_{n+1}}{\|\boldsymbol{w}_0\|} \tag{2.4.8}$$

The quantity $|\boldsymbol{n}^T\boldsymbol{y}|$ measures the normal distance D_0 between the origin and the hyperplane H. Thus

$$D_0=\frac{|w_{n+1}|}{\|\boldsymbol{w}_0\|} \tag{2.4.9}$$

The distance between an arbitrary point R, associated to a vector z, from the hyperplane, is

$$D_z = \left| n^T (y - z) \right| = \left| n^T (z - y) \right| \qquad (2.4.10)$$

and by applying Eqs. (2.4.7) and (2.4.8) we get

$$D_z = \left| \frac{w_0^T}{\|w_0\|}(z - y) \right| = \left| \frac{w_0^T z + w_{n+1}}{\|w_0\|} \right| \qquad (2.4.11)$$

In the particular case $w_{n+1} = 0$, the hyperplane H passes through the origin, since $D_0 = 0$.

■ **Example 2.4.1** Consider the decision boundary

$$3x_1 + 4x_2 - 5 = 0$$

in R^2. Here $\|\boldsymbol{w}_0\| = (3^2 + 4^2)^{1/2} = 5$ and the normal unit vector pointing at the positive side of the straight line is $\boldsymbol{n} = \boldsymbol{w}_0 / \|\boldsymbol{w}_0\| = \left(\frac{3}{5}, \frac{4}{5}\right)^T$. The distance of a pattern located at $(1,2)^T$ from the decision boundary is

$$D_{(1,2)} = \left| \frac{(3,4)\,(1,2)^T - 5}{5} \right| = \left| \frac{3+8-5}{5} \right| = 1.2$$

■ **Example 2.4.2** Consider a two-class pattern classification of a given 3-D pattern set, using the plane

$$2x_1 - x_2 + 2x_3 - 7 = 0$$

as a linear decision boundary. If patterns whose normal distance from the plane is less than 0.01 are excluded, one should eliminate all the patterns (y_1, y_2, y_3) for which

$$\left| \frac{2y_1 - y_2 + 2y_3 - 7}{\|w_0\|} \right| = \left| \frac{2y_1 - y_2 + 2y_3 - 7}{3} \right| < 0.01$$

If a pattern is located at (0.51,0,3), it is excluded since

$$\left| \frac{2 \cdot 0.51 - 0 + 2 \cdot 3 - 7}{3} \right| = \frac{0.02}{3} < 0.01$$

Dichotomies

■ **Definition 2.4.1** An $m-$pattern set in R^n is said to be *regularly distributed*, if none of its $(n+1)-$pattern subsets is located on a hyperplane in R^n.

■ **Theorem 2.4.1** Given a regularly distributed $m-$pattern set in R^n, the number of its linear dichotomies is

$$D(m,n)=\begin{cases} 2\sum_{i=0}^{n}\binom{m-1}{i}, & m>n \\ 2^m, & m\leq n \end{cases} \tag{2.4.13}$$

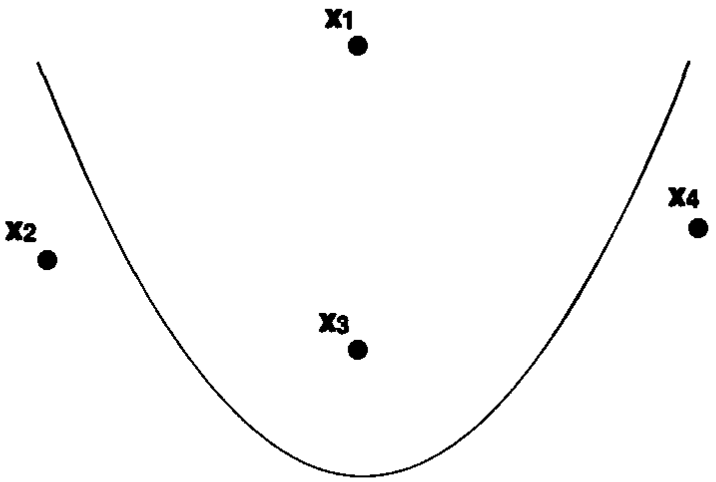
Let us consider a regularly distributed $m-$pattern set and generalized decision functions which transform the original $n-$dimensional patterns into $N-$dimensional ones. The number of linear dichotomies that can be obtained is $D(m,N)$, compared with the total number of two-class groupings which is 2^m. Thus, the probability for a random dichotomy (i.e. a random two-class grouping of the pattern set) to be *linearly implementable* is

$$p(m,N)=\frac{D(m,N)}{2^m}=\begin{cases}2^{-(m-1)}\sum_{i=0}^{N}\binom{m-1}{i}, & m>N\\ 1, & m\le N\end{cases} \tag{2.4.14}$$

Consequently, if the number of patterns does not exceed the new dimensionality of the pattern space, each two disjoint pattern classes whose union is the whole pattern set, are linearly separable in the N- dimensional space.

■ **Example 2.4.5** Consider the four 2-D patterns in Fig. 2.4.4. There is no way that the classes $\{x_1, x_3\}$, $\{x_2, x_4\}$ will be linearly separated. However, by using quadratic decision functions and boundaries, we get

$N = 5$ and since $m = 4 < 5$ linear separation in R^5 is possible. In the original space, we simply use a quadratic parabola.



■ **Figure 2.4.4** A 2-D problem linearly separable only in R^5.