Notes on Linear Models for Classification

Your Name

October 29, 2025

Contents

Chapter 1

Discriminant Functions

1.1 Discriminant Functions for Two Classes

This section covers the simplest case of a linear discriminant for a two-class classification problem.

Definition 1.1.1 (Linear Discriminant Function (2 Classes))**.** A linear discriminant function is defined by taking a linear function of the input vector $\mathbf{x}$:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{1.1}$$

where $\mathbf{w}$ is the **weight vector** and w_0 is the **bias**. The negative of the bias, $-w_0$, is sometimes referred to as a **threshold**.

1.1.1 Decision Boundary and Classification Rule

Definition 1.1.2 (Classification Rule)**.** An input vector $\mathbf{x}$ is assigned to class $\mathcal{C}_1$ if $y(\mathbf{x}) \geq 0$ and to class $\mathcal{C}_2$ otherwise (i.e., if $y(\mathbf{x}) < 0$).

Definition 1.1.3 (Decision Surface)**.** The **decision boundary** (or decision surface) is the set of points where the discriminant function is zero. It is defined by the relation:

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0 \tag{1.2}$$

For a D-dimensional input space $\mathbf{x}$, this equation defines a $(D-1)$-dimensional hyperplane.

1.1.2 Geometric Properties of the Decision Surface

We can derive several key geometric properties from the definition of the linear discriminant.

Proposition 1.1.4 (Orientation of the Decision Surface)**.** *The weight vector* $\mathbf{w}$ *is orthogonal (perpendicular) to every vector lying within the decision surface. Therefore,* $\mathbf{w}$ *determines the orientation of the decision surface.*

Proof. Let $\mathbf{x}_A$ and $\mathbf{x}_B$ be any two distinct points that lie on the decision surface. By definition, $y(\mathbf{x}_A) = 0$ and $y(\mathbf{x}_B) = 0$.

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_A + w_0 &= 0 \\ \mathbf{w}^T\mathbf{x}_B + w_0 &= 0 \end{aligned}$$

Subtracting the second equation from the first gives:

$$\begin{aligned} (\mathbf{w}^T\mathbf{x}_A + w_0) - (\mathbf{w}^T\mathbf{x}_B + w_0) &= 0 - 0 \\ \mathbf{w}^T\mathbf{x}_A - \mathbf{w}^T\mathbf{x}_B &= 0 \\ \mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) &= 0 \end{aligned}$$

The vector $(\mathbf{x}_A - \mathbf{x}_B)$ is a vector that lies in the decision surface (it connects two points on the surface). Since its dot product with $\mathbf{w}$ is zero, $\mathbf{w}$ must be orthogonal to this vector. This holds for any two points $\mathbf{x}_A, \mathbf{x}_B$ on the surface, proving the proposition. □

Proposition 1.1.5 (Location of the Decision Surface)**.** *The bias parameter w_0 determines the location of the decision surface relative to the origin. Specifically, the normal distance from the origin to the hyperplane is $\frac{-w_0}{\|\mathbf{w}\|}$.*

Proof. Let $\mathbf{x}_{\text{ds}}$ be any point on the decision surface. The perpendicular distance from the origin to the hyperplane is the projection of the vector $\mathbf{x}_{\text{ds}}$ onto the normal vector $\mathbf{w}$. The unit normal vector is $\frac{\mathbf{w}}{\|\mathbf{w}\|}$. The distance (as a scalar) is the dot product of $\mathbf{x}_{\text{ds}}$ with this unit normal:

$$\text{Distance} = \mathbf{x}_{\text{ds}}^T \left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) = \frac{\mathbf{w}^T \mathbf{x}_{\text{ds}}}{\|\mathbf{w}\|}$$

From the definition of the decision surface, we know that $\mathbf{w}^T\mathbf{x}_{\text{ds}} + w_0 = 0$, which implies $\mathbf{w}^T\mathbf{x}_{\text{ds}} = -w_0$. Substituting this into our distance equation, we get:

$$\text{Distance from origin} = \frac{-w_0}{\|\mathbf{w}\|} \tag{1.3}$$

Thus, the location of the plane is controlled by w_0 (relative to the magnitude of $\mathbf{w}$). □

Proposition 1.1.6 (Perpendicular Distance from a Point $\mathbf{x}$)**.** *The value of $y(\mathbf{x})$ provides a signed measure of the perpendicular distance r from the point $\mathbf{x}$ to the decision surface. The distance is given by:*

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|} \tag{1.4}$$

Proof. Let $\mathbf{x}$ be an arbitrary point in the input space. Let $\mathbf{x}_\perp$ be its orthogonal projection onto the decision surface, so $y(\mathbf{x}_\perp) = 0$. Let r be the signed perpendicular distance from $\mathbf{x}_\perp$ to $\mathbf{x}$. The vector from $\mathbf{x}_\perp$ to $\mathbf{x}$ is parallel to the normal vector $\mathbf{w}$. We can therefore write this vector as $r\frac{\mathbf{w}}{\|\mathbf{w}\|}$. We can decompose the vector $\mathbf{x}$ as the sum of its projection on the plane and this normal component:

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

Now, let's evaluate the discriminant function $y(\mathbf{x})$ by multiplying by $\mathbf{w}^T$ and adding w_0:

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{w}^T\mathbf{x} + w_0 \\ &= \mathbf{w}^T\left(\mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\ &= (\mathbf{w}^T\mathbf{x}_\perp + w_0) + \mathbf{w}^T\left(r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) \end{aligned}$$

We know that $y(\mathbf{x}_\perp) = \mathbf{w}^T\mathbf{x}_\perp + w_0 = 0$, because $\mathbf{x}_\perp$ is on the decision surface.

$$\begin{aligned} y(\mathbf{x}) &= 0 + r\left(\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|}\right) \\ &= r\left(\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|}\right) \\ &= r\|\mathbf{w}\| \end{aligned}$$

Rearranging this result to solve for the distance r, we find:

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|} \tag{1.5}$$

This confirms that $y(\mathbf{x})$ is proportional to the signed perpendicular distance from $\mathbf{x}$ to the boundary. □

1.1.3 Augmented Input Space

It is often convenient to use a more compact notation by augmenting the input vector $\mathbf{x}$.

Definition 1.1.7 (Augmented Vectors)**.** We introduce a "dummy" input $x_0 = 1$ and define the augmented input vector $\tilde{\mathbf{x}}$ and augmented weight vector $\tilde{\mathbf{w}}$ as:

$$\tilde{\mathbf{x}} = (x_0, x_1, \ldots, x_D)^T = (1, \mathbf{x})^T \tag{1.6}$$

$$\tilde{\mathbf{w}} = (w_0, w_1, \ldots, w_D)^T = (w_0, \mathbf{w})^T \tag{1.7}$$

Proposition 1.1.8. *The linear discriminant function $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ can be written in the augmented space as:*

$$y(\mathbf{x}) = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}} \tag{1.8}$$

Proof.

$$\begin{aligned}
\tilde{\mathbf{w}}^T\tilde{\mathbf{x}} &= \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}^T \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_D \end{pmatrix} = \begin{pmatrix} w_0 \\ \mathbf{w} \end{pmatrix}^T \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \\
&= w_0 \cdot 1 + w_1x_1 + \cdots + w_Dx_D \\
&= w_0 + \mathbf{w}^T\mathbf{x} = y(\mathbf{x})
\end{aligned}$$

In this $(D+1)$-dimensional augmented space, the decision surface $y(\mathbf{x}) = 0$ is defined by $\tilde{\mathbf{w}}^T\tilde{\mathbf{x}} = 0$, which is a D-dimensional hyperplane that passes directly through the origin. □

Proposition 1.1.9 (Perpendicular Distance from a Point $\mathbf{x}$)**.** *The value of $y(\mathbf{x})$ provides a signed measure of the perpendicular distance r from the point $\mathbf{x}$ to the decision surface. The distance is given by:*

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|} \tag{1.9}$$

Proof 1 (Geometric Projection). Let $\mathbf{x}$ be an arbitrary point in the input space. Let $\mathbf{x}_\perp$ be its orthogonal projection onto the decision surface, which means $y(\mathbf{x}_\perp) = 0$.

Let r be the signed perpendicular distance from $\mathbf{x}_\perp$ to $\mathbf{x}$. The vector from $\mathbf{x}_\perp$ to $\mathbf{x}$ is parallel to the normal vector $\mathbf{w}$. We can therefore write this vector as $r\frac{\mathbf{w}}{\|\mathbf{w}\|}$.

We can decompose the vector $\mathbf{x}$ as the sum of its projection on the plane and this normal component:

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

Now, let's evaluate the discriminant function $y(\mathbf{x})$ by multiplying by $\mathbf{w}^T$ and adding w_0:

$$\begin{aligned}
y(\mathbf{x}) &= \mathbf{w}^T\mathbf{x} + w_0 \\
&= \mathbf{w}^T\left(\mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\
&= (\mathbf{w}^T\mathbf{x}_\perp + w_0) + \mathbf{w}^T\left(r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right)
\end{aligned}$$

We know that $y(\mathbf{x}_\perp) = \mathbf{w}^T\mathbf{x}_\perp + w_0 = 0$, because $\mathbf{x}_\perp$ is on the decision surface.

$$\begin{aligned}
y(\mathbf{x}) &= 0 + r\left(\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|}\right) \\
&= r\left(\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|}\right) \\
&= r\|\mathbf{w}\|
\end{aligned}$$

Rearranging this result to solve for the distance r, we find:

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|} \tag{1.10}$$

This confirms that $y(\mathbf{x})$ is proportional to the signed perpendicular distance from $\mathbf{x}$ to the boundary. The absolute distance is $\frac{|y(\mathbf{x})|}{\|\mathbf{w}\|}$. □

Proof 2 (by Optimization, based on your image). The perpendicular distance is the minimum distance from $\mathbf{x}$ to any point $\mathbf{v}$ on the hyperplane.

$$\text{distance} = \min_{\mathbf{v}}\{\|\mathbf{x}-\mathbf{v}\|\} \quad \text{subject to} \quad \mathbf{w}^T\mathbf{v} + w_0 = 0.$$

Let the closest point on the plane be $\mathbf{v}$. The vector from $\mathbf{v}$ to $\mathbf{x}$ must be normal to the plane, so $\mathbf{x}-\mathbf{v}$ is parallel to $\mathbf{w}$. We can write:

$$\mathbf{x}-\mathbf{v} = k\mathbf{w} \implies \mathbf{v} = \mathbf{x} - k\mathbf{w}$$

for some scalar k. We find k by enforcing the constraint $\mathbf{w}^T\mathbf{v} + w_0 = 0$:

$$\begin{aligned}
\mathbf{w}^T(\mathbf{x}-k\mathbf{w}) + w_0 &= 0\\
\mathbf{w}^T\mathbf{x} - k(\mathbf{w}^T\mathbf{w}) + w_0 &= 0\\
(\mathbf{w}^T\mathbf{x}+w_0) - k\|\mathbf{w}\|^2 &= 0\\
y(\mathbf{x}) = k\|\mathbf{w}\|^2 &\implies k = \frac{y(\mathbf{x})}{\|\mathbf{w}\|^2}
\end{aligned}$$

This confirms that the vector $\mathbf{x}-\mathbf{v} = \frac{y(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}$. The distance is the magnitude of this vector:

$$\|\mathbf{x}-\mathbf{v}\| = \left\|\frac{y(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}\right\| = \left|\frac{y(\mathbf{x})}{\|\mathbf{w}\|^2}\right|\|\mathbf{w}\| = \frac{|y(\mathbf{x})|}{\|\mathbf{w}\|}$$

To show this is the minimum, let $\mathbf{u}$ be any other point on the plane ($\mathbf{w}^T\mathbf{u} + w_0 = 0$).

$$\begin{aligned}
\|\mathbf{x}-\mathbf{u}\|^2 &= \|\mathbf{x}-\mathbf{v}+\mathbf{v}-\mathbf{u}\|^2\\
&= \|\mathbf{x}-\mathbf{v}\|^2 + \|\mathbf{v}-\mathbf{u}\|^2 + 2(\mathbf{x}-\mathbf{v})^T(\mathbf{v}-\mathbf{u})
\end{aligned}$$

The cross-term is $2(\mathbf{x}-\mathbf{v})^T(\mathbf{v}-\mathbf{u}) = 2\left(\frac{y(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}\right)^T(\mathbf{v}-\mathbf{u}) = 2\frac{y(\mathbf{x})}{\|\mathbf{w}\|^2}\mathbf{w}^T(\mathbf{v}-\mathbf{u})$. Since $\mathbf{w}^T\mathbf{v} = -w_0$ and $\mathbf{w}^T\mathbf{u} = -w_0$, the term $\mathbf{w}^T(\mathbf{v}-\mathbf{u}) = 0$. Thus, $\|\mathbf{x}-\mathbf{u}\|^2 = \|\mathbf{x}-\mathbf{v}\|^2 + \|\mathbf{v}-\mathbf{u}\|^2$. Because $\|\mathbf{v}-\mathbf{u}\|^2 \geq 0$, we have $\|\mathbf{x}-\mathbf{u}\|^2 \geq \|\mathbf{x}-\mathbf{v}\|^2$. This proves the minimum distance is $\|\mathbf{x}-\mathbf{v}\| = \frac{|y(\mathbf{x})|}{\|\mathbf{w}\|}$. □

1.2 Multiple Classes (K > 2)

Simple approaches to creating a K-class discriminant from multiple two-class discriminants, such as the **one-versus-the-rest** or **one-versus-one** schemes, run into difficulties. Both methods can create ambiguous regions in the input space where the classification is not clearly defined.

1.2.1 K-Class Discriminant Function

We can avoid these problems by defining a single K-class discriminant composed of K separate linear functions, one for each class $\mathcal{C}_k$:

Definition 1.2.1 (K-Class Discriminant)**.** The discriminant is defined by a set of K linear functions of the form:

$$y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0} \tag{1.11}$$

for $k = 1, \ldots, K$. Each class $\mathcal{C}_k$ has its own weight vector $\mathbf{w}_k$ and bias w_{k0}.

1.2.2 Decision Boundaries

The decision boundary between any two classes, $\mathcal{C}_k$ and $\mathcal{C}_j$, is the set of points $\mathbf{x}$ where their discriminant functions are equal (i.e., they "tie").

$$y_k(\mathbf{x}) = y_j(\mathbf{x}) \tag{1.12}$$

We can derive the explicit form of this boundary:

$$\begin{aligned} \mathbf{w}_k^T\mathbf{x} + w_{k0} &= \mathbf{w}_j^T\mathbf{x} + w_{j0} \\ \mathbf{w}_k^T\mathbf{x} - \mathbf{w}_j^T\mathbf{x} + w_{k0} - w_{j0} &= 0 \\ (\mathbf{w}_k - \mathbf{w}_j)^T\mathbf{x} + (w_{k0} - w_{j0}) &= 0 \end{aligned}$$

Remark 1.2.2 (Analogy to Two-Class Case). This resulting boundary equation has the exact same form as the linear discriminant for the two-class case, $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$. In this multi-class context, we can think of the boundary between $\mathcal{C}_k$ and $\mathcal{C}_j$ as being defined by an equivalent weight vector $\mathbf{w} = (\mathbf{w}_k - \mathbf{w}_j)$ and an equivalent bias $w_0 = (w_{k0} - w_{j0})$. This shows that the boundary between any two classes is a single $(D-1)$-dimensional hyperplane.

1.2.3 Convexity of Decision Regions

The decision regions formed by this discriminant are always singly connected and convex.

Proposition 1.2.3. *The decision region $\mathcal{R}_k$ for class $\mathcal{C}_k$ (the set of all points $\mathbf{x}$ assigned to $\mathcal{C}_k$) is convex.*

Proof. Consider two points, $\mathbf{x}_A$ and $\mathbf{x}_B$, both of which lie inside the decision region $\mathcal{R}_k$. By definition, this means that for $\mathbf{x}_A$ and $\mathbf{x}_B$, the discriminant y_k is larger than all other discriminants:

$$\begin{aligned} \forall j \neq k, \quad y_k(\mathbf{x}_A) > y_j(\mathbf{x}_A) \\ \forall j \neq k, \quad y_k(\mathbf{x}_B) > y_j(\mathbf{x}_B) \end{aligned}$$

Now, consider any point $\hat{\mathbf{x}}$ that lies on the line segment connecting $\mathbf{x}_A$ and $\mathbf{x}_B$. Such a point can be written as:

$$\hat{\mathbf{x}} = \lambda\mathbf{x}_A + (1-\lambda)\mathbf{x}_B \tag{1.13}$$

where $0 \leq \lambda \leq 1$. Let's evaluate the discriminant function y_k at this point $\hat{\mathbf{x}}$. Due to the linearity of the function y_k:

$$\begin{aligned} y_k(\hat{\mathbf{x}}) &= \mathbf{w}_k^T\hat{\mathbf{x}} + w_{k0} \\ &= \mathbf{w}_k^T(\lambda\mathbf{x}_A + (1-\lambda)\mathbf{x}_B) + (\lambda + 1 - \lambda)w_{k0} \\ &= \lambda(\mathbf{w}_k^T\mathbf{x}_A + w_{k0}) + (1-\lambda)(\mathbf{w}_k^T\mathbf{x}_B + w_{k0}) \\ &= \lambda y_k(\mathbf{x}_A) + (1-\lambda)y_k(\mathbf{x}_B) \end{aligned}$$

The same linearity holds for any other discriminant $y_j(\mathbf{x})$.

$$y_j(\hat{\mathbf{x}}) = \lambda y_j(\mathbf{x}_A) + (1-\lambda)y_j(\mathbf{x}_B)$$

Now we use our initial assumptions. Since $y_k(\mathbf{x}_A) > y_j(\mathbf{x}_A)$ and $y_k(\mathbf{x}_B) > y_j(\mathbf{x}_B)$, and given that $\lambda \geq 0$ and $(1-\lambda) \geq 0$:

$$\begin{aligned} \lambda y_k(\mathbf{x}_A) \geq \lambda y_j(\mathbf{x}_A) \\ (1-\lambda)y_k(\mathbf{What}_B) \geq (1-\lambda)y_j(\mathbf{x}_B) \end{aligned}$$

Adding these two inequalities, we get:

$$\lambda y_k(\mathbf{x}_A) + (1-\lambda)y_k(\mathbf{x}_B) > \lambda y_j(\mathbf{x}_A) + (1-\lambda)y_j(\mathbf{x}_B)$$

Substituting the linear combinations:

$$y_k(\hat{\mathbf{x}}) > y_j(\hat{\mathbf{x}}), \quad \text{for all } j \neq k$$

This shows that the point $\hat{\mathbf{x}}$ also lies inside the decision region $\mathcal{R}_k$. Since this is true for any point on the line segment between $\mathbf{x}_A$ and $\mathbf{x}_B$, the region $\mathcal{R}_k$ is, by definition, convex. $\square$

1.2.4 Classification Rule

The discriminant functions are used to classify new points with a "winner-takes-all" rule.

Definition 1.2.4 (Classification Rule (K-Classes))**.** A new input vector $\mathbf{x}$ is assigned to the class $\mathcal{C}_k$ whose discriminant function $y_k(\mathbf{x})$ has the largest value:

$$\text{Assign } \mathbf{x} \text{ to } \mathcal{C}_k \quad \text{if} \quad y_k(\mathbf{x}) > y_j(\mathbf{x}) \text{ for all } j \neq k$$