LINEAR CLASSIFICATION

Scalar Products

Definition

For two vectors **x** and **y** in $\mathbb{R}^d$, the **scalar product** of **x** and **y** is

$$\langle \mathbf{x}, \mathbf{y} \rangle \quad := \quad x_1 y_1 + \ldots + x_d y_d \quad = \quad \sum_{i=1}^{d} x_i y_i$$

Note: $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2$, so the Euclidean norm (= the length) of **x** is $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Linearity

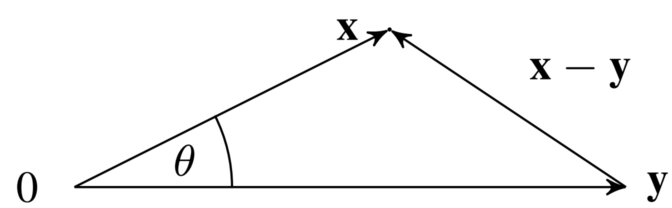
The scalar product is additive in both arguments,

$$\langle \mathbf{x} + \mathbf{z}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{z}, \mathbf{y} \rangle \qquad \text{and} \qquad \langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$$

and scales as

$$\langle c \cdot \mathbf{x}, \mathbf{y} \rangle = c \cdot \langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, c \cdot \mathbf{y} \rangle \qquad \text{for any } c \in \mathbb{R} \ .$$

Functions that are additive and scale-equivariant are called **linear**, so the scalar product is linear in both arguments.



Recall: The cosine rule

If two vectors **x** and **y** enclose an angle θ, then

$$\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\cos\theta\|\mathbf{x}\|\|\mathbf{y}\|$$

(If θ is a right angle, then $\cos\theta = 0$, and this becomes Pythogoras' $\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.)

Cosine rule for scalar products

It is easy to check that

$$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 = 2\langle\mathbf{x},\mathbf{y}\rangle$$

Substituting gives

$$2\cos\theta\|\mathbf{x}\|\|\mathbf{y}\| = 2\langle\mathbf{x},\mathbf{y}\rangle$$

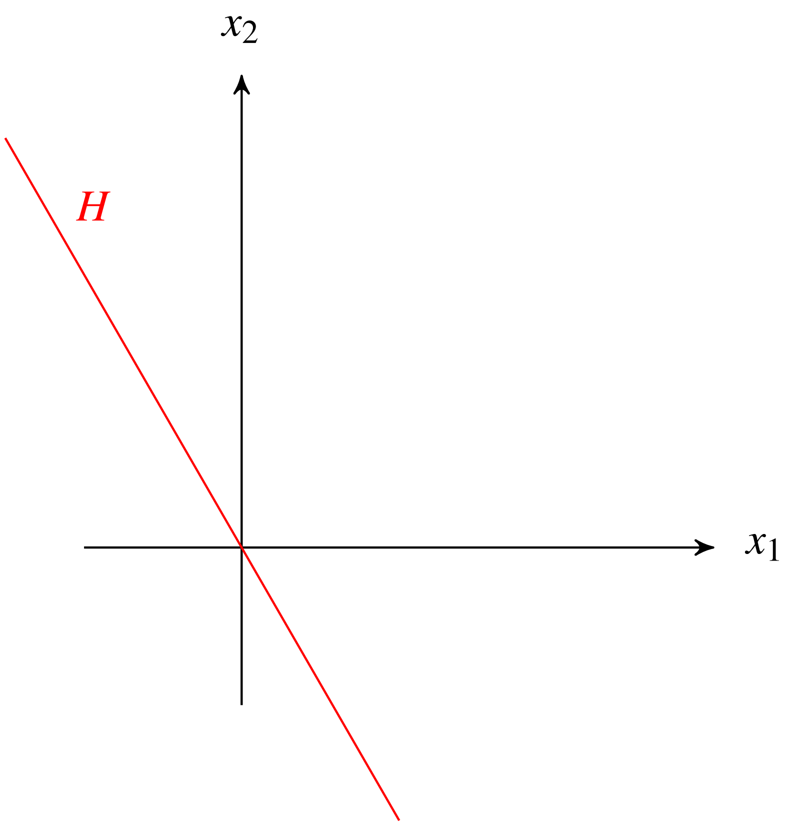
and hence

$$\cos\theta = \frac{\langle\mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

Consequences of the cosine rule

The scalar product satisfies $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \|\mathbf{y}\|$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are parallel, and

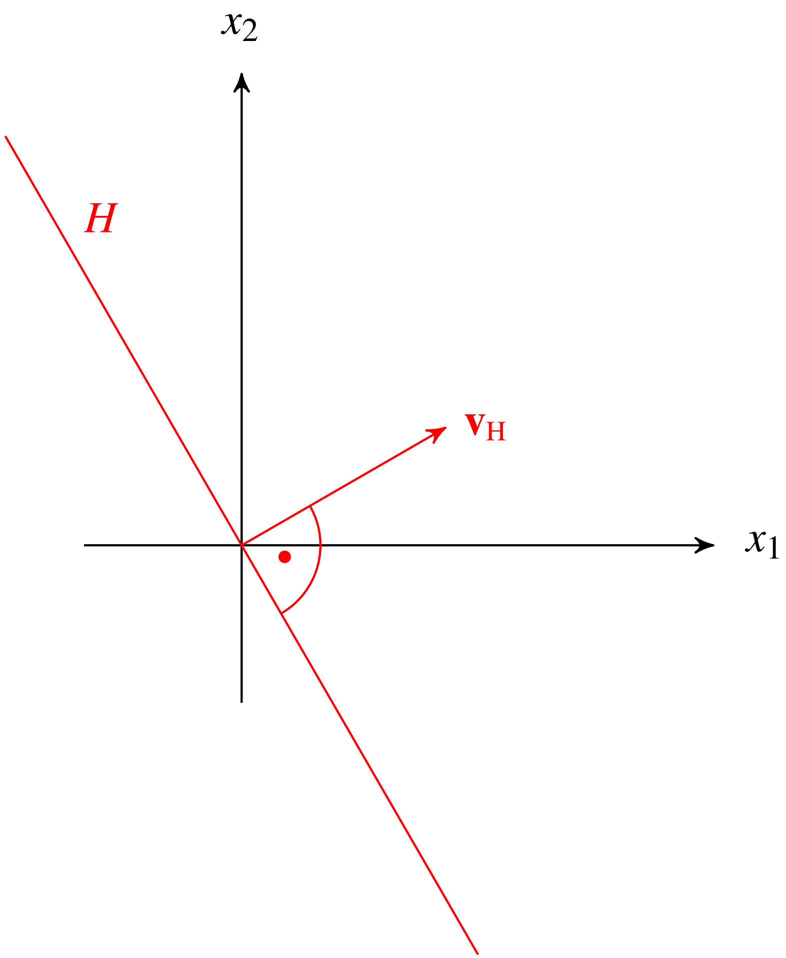
$$\langle \mathbf{x}, \mathbf{y} \rangle = 0 \qquad \text{if and only if } \mathbf{x} \text{ and } \mathbf{y} \text{ are orthogonal.}$$



Hyperplanes

A **hyperplane** in $\mathbb{R}^d$ is a linear subspace of dimension $(d-1)$.

- A hyperplane in $\mathbb{R}^2$ is a line.
- A hyperplane in $\mathbb{R}^3$ is a plane.
- A hyperplane always contains the origin, since it is a linear subspace.



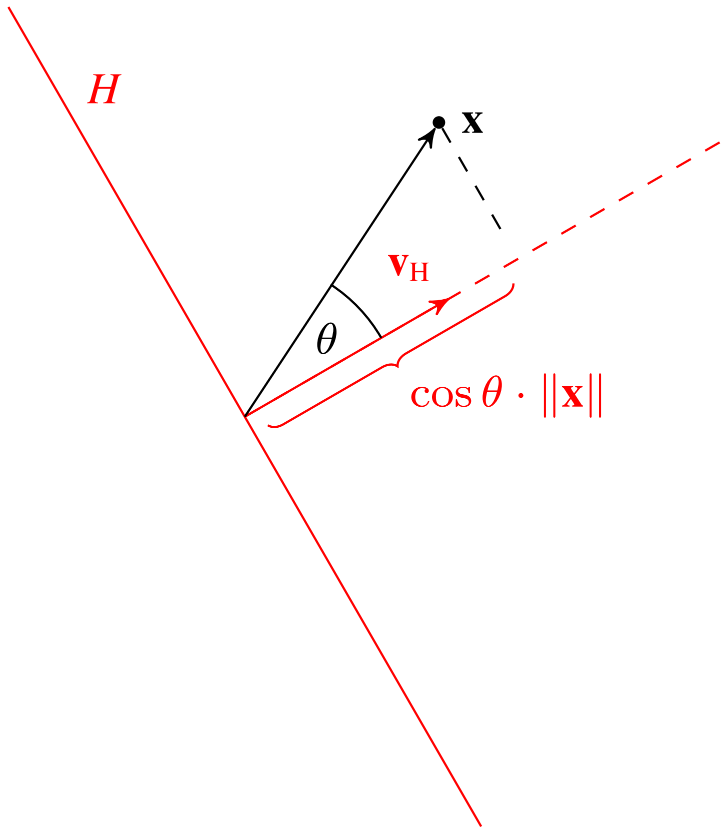
Hyperplanes

- Consider a hyperplane H in $\mathbb{R}^d$. Think of H as a set of points.
- Each point $\mathbf{x}$ in H is a vector $\mathbf{x} \in \mathbb{R}^d$.
- Now draw a vector $\mathbf{v}_{\mathrm{H}}$ that is orthogonal to H.
- Then any vector $\mathbf{x} \in \mathbb{R}^d$ is a point in H if and only if $\mathbf{x}$ is orthogonal to $\mathbf{v}_{\mathrm{H}}$.
- Hence:

$$\mathbf{x} \in H \qquad \Leftrightarrow \qquad \langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle = 0 \; .$$

- If we choose $\mathbf{v}_{\mathrm{H}}$ to have length $\|\mathbf{v}_{\mathrm{H}}\| = 1$, then $\mathbf{v}_{\mathrm{H}}$ is called a **normal vector** of H.

$$H = \{\mathbf{x} \in \mathbb{R}^d \mid \langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle = 0\} \; .$$



Distance from the plane

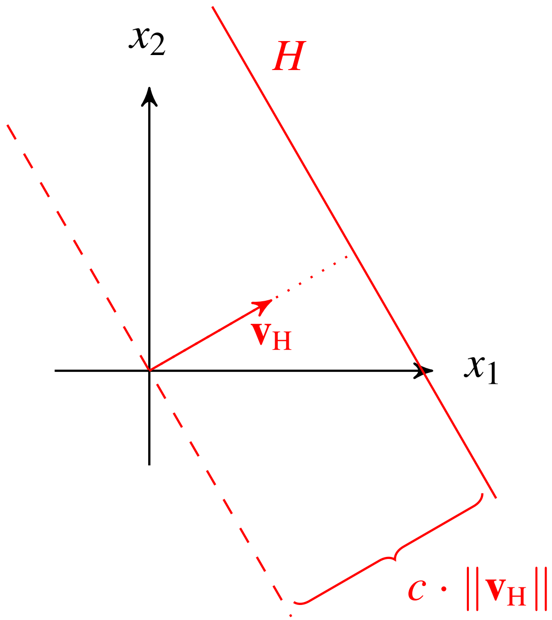
- The projection of $\mathbf{x}$ onto the direction of $\mathbf{v}_{\mathrm{H}}$ has length $\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle$ *measured in units of* $\mathbf{v}_H$, i.e. length $\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle / \|\mathbf{v}_{\mathrm{H}}\|$ in the units of the coordinates.
- By cosine rule: The distance of $\mathbf{x}$ from the plane is

$$d(\mathbf{x}, H) = \frac{\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle}{\|\mathbf{v}_{\mathrm{H}}\|} = \cos\theta \cdot \|\mathbf{x}\| \ .$$

Which side of the plane?

- The cosine satisfies $\cos\theta > 0$ iff $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$.
- We can decide which side of the plane $\mathbf{x}$ is on using

$$\operatorname{sgn}(\cos\theta) = \operatorname{sgn}\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle \ .$$



Affine Hyperplanes

- An **affine hyperplane** $H_{\mathbf{w}}$ is a hyperplane shifted by a vector $\mathbf{w}$,

$$H_{\mathbf{w}} = H + \mathbf{w} \ .$$

(That means $\mathbf{w}$ is added to each point $\mathbf{x}$ in H.)

- We choose $\mathbf{w}$ in the direction of $\mathbf{v}_{\mathrm{H}}$, so

$$\mathbf{w} = c \cdot \mathbf{v}_{\mathrm{H}} \qquad \text{for some } c > 0 \ .$$

Which side of the plane are we on?

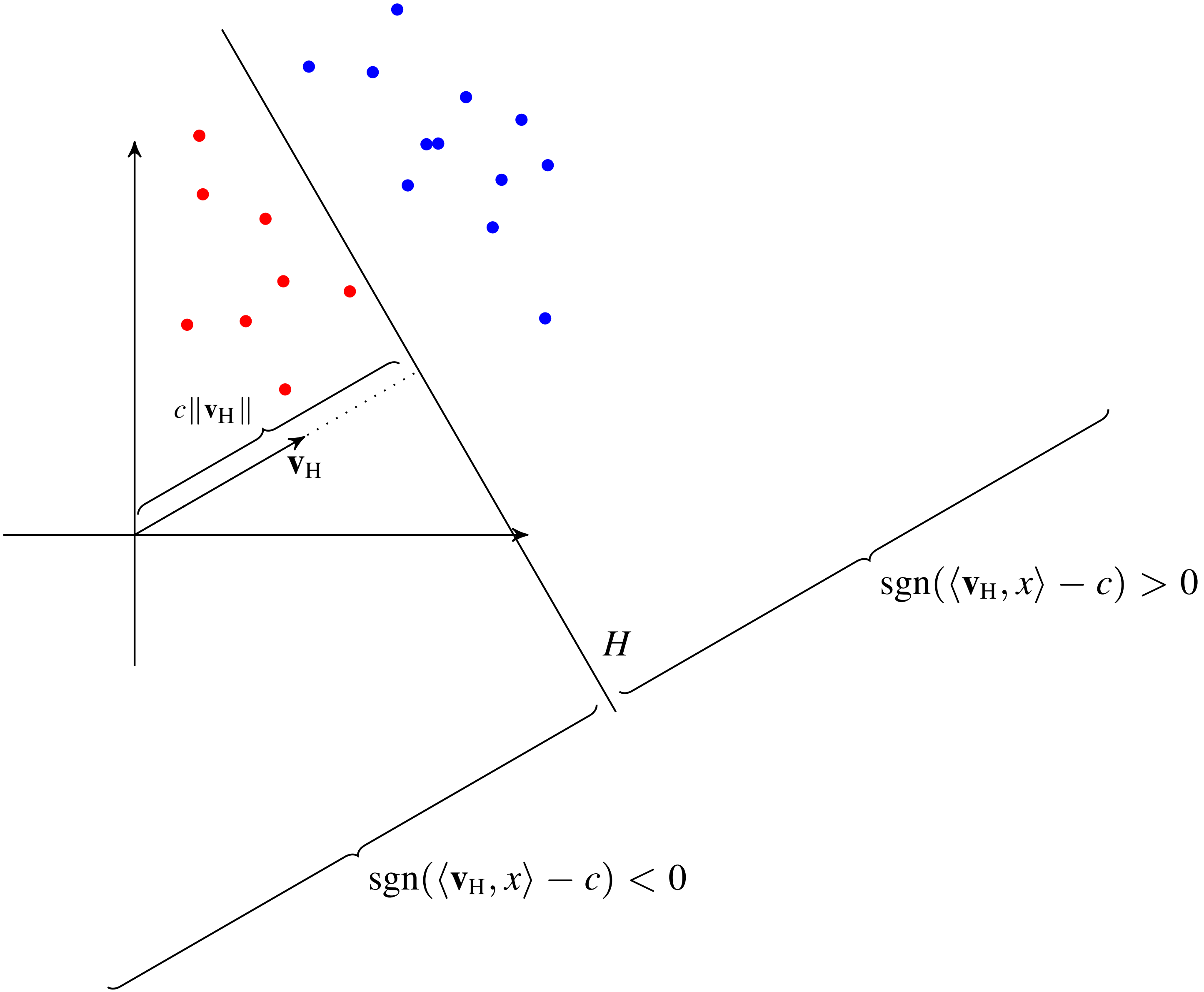
- Which side of $H_{\mathbf{w}}$ a point $\mathbf{x}$ is on is determined by

$$\operatorname{sgn}(\langle \mathbf{x} - \mathbf{w}, \mathbf{v}_{\mathrm{H}} \rangle) = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle - c \, \langle \mathbf{v}_{\mathrm{H}}, \mathbf{v}_{\mathrm{H}} \rangle) = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle - c\|\mathbf{v}_{\mathrm{H}}\|^2) \ .$$

- If $\mathbf{v}_{\mathrm{H}}$ is a unit vector, we can use

$$\operatorname{sgn}(\langle \mathbf{x} - \mathbf{w}, \mathbf{v}_{\mathrm{H}} \rangle) = \operatorname{sgn}(\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle - c) \ .$$

Classification with Affine Hyperplanes



Definition

A **linear classifier** is a function of the form

$$f_{\mathrm{H}}(\mathbf{x}) := \operatorname{sgn}(\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle - c) \ ,$$

where $\mathbf{v}_{\mathrm{H}} \in \mathbb{R}^d$ is a vector and $c \in \mathbb{R}_+$.

Note:

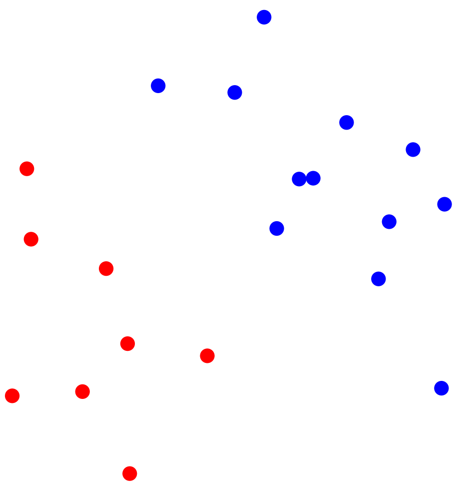
- We usually assume $\mathbf{v}_{\mathrm{H}}$ to be a unit vector. If it is not, f_{H} still defines a linear classifier, but c describes a shift of a different length.
- Specifying a linear classifier in $\mathbb{R}^d$ requires $d + 1$ scalar parameters.

Definition

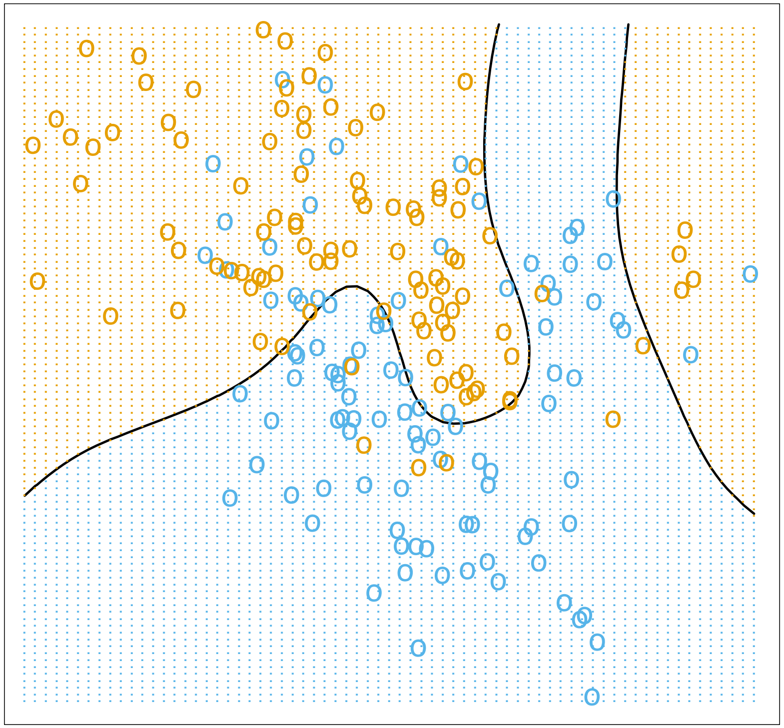
Two sets $A, B \in \mathbb{R}^d$ are called **linearly separable** if there is an affine hyperplane H which separates them, i.e. which satisfies

$$\langle \mathbf{x}, \mathbf{v}_{\mathrm{H}} \rangle - c = \begin{cases} < 0 & \text{if } \mathbf{x} \in A \\ > 0 & \text{if } \mathbf{x} \in B \end{cases}$$

Linear Separability



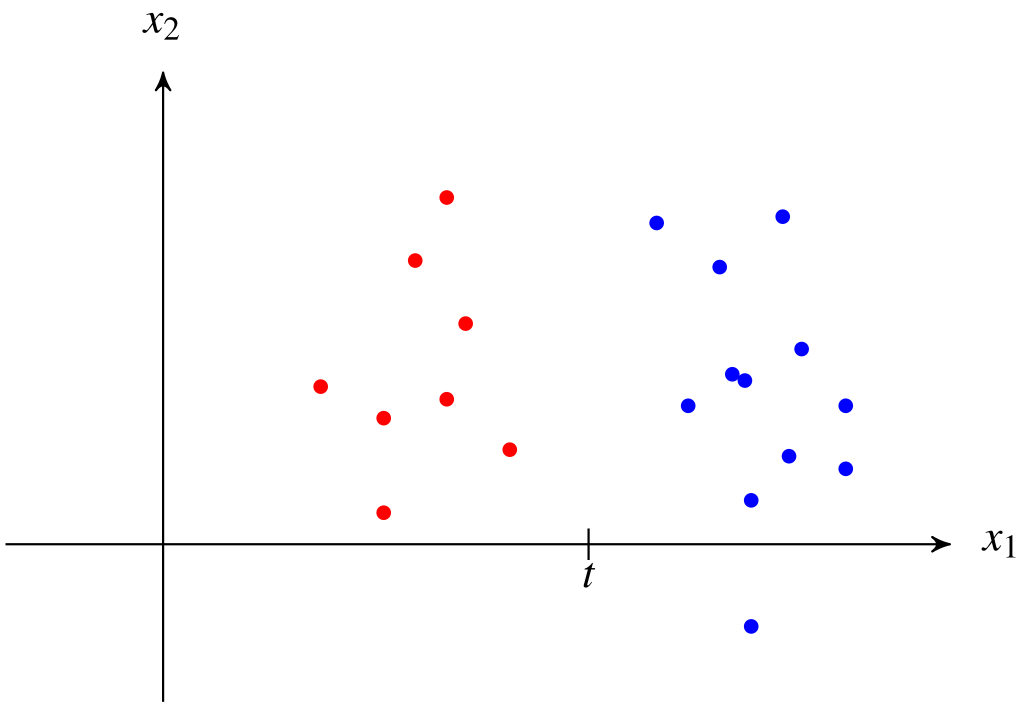
linearly separable



not linearly separable

LINEAR SEPARABILITY

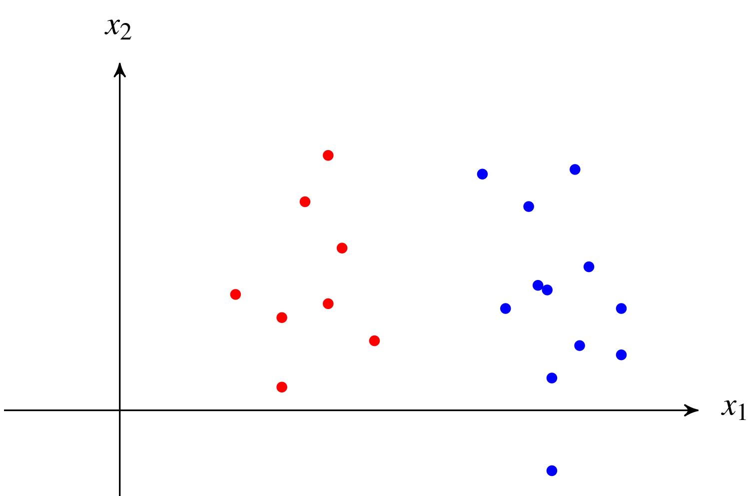
- Recall that when data is represented by points in $\mathbb{R}^d$, each axis represents a quantity that is measured (a “variable”).
- If there exists a single variable that distinguishes two classes, these classes can be distinguished along a single axis.



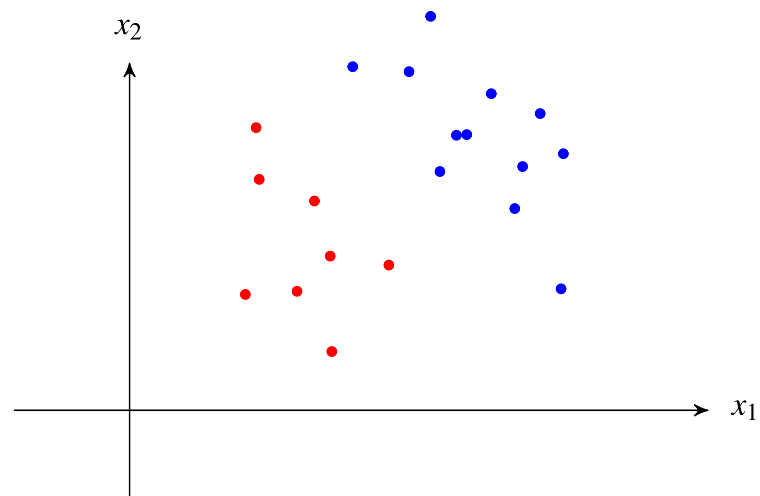
- In this illustration, we could classify by a “threshold point” t on the line.

LINEAR SEPARABILITY

- Even if classes cannot be distinguished by a single variable, they may be distinguishable by a combination of several variables.
- That is the case for linearly separable data. The threshold point along x_1 is now a function of the threshold point along x_2, and vice versa. Linearly separable also implies this function is linear.

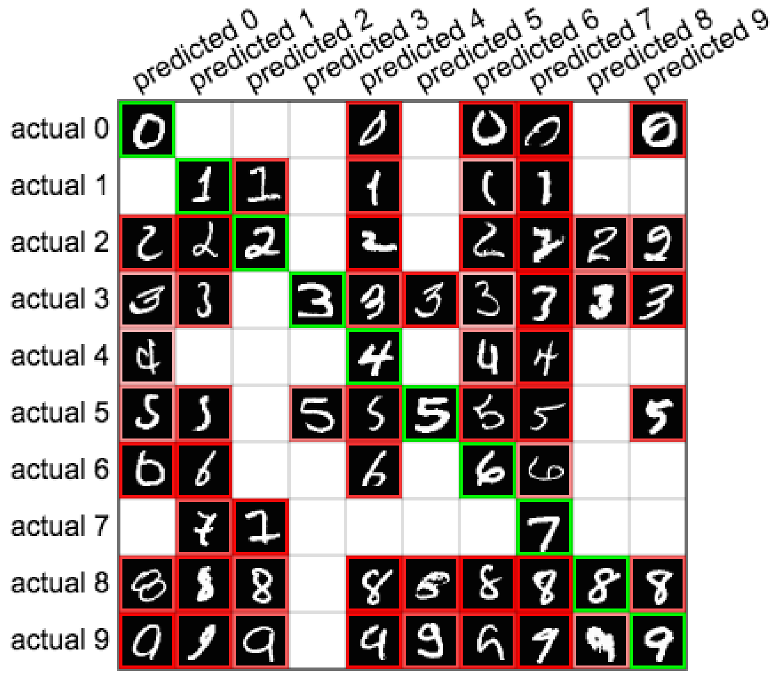


separable by a single measurement



not separable by a single measurement

but linearly separable



predicted 0
predicted 1
predicted 2
predicted 3
predicted 4
predicted 5
predicted 6
predicted 7
predicted 8
predicted 9
actual 0
actual 1
actual 2
actual 3
actual 4
actual 5
actual 6
actual 7
actual 8
actual 9