Linear Classification

- **Regression** → output y is continuous ($y \in \mathbb{R}$)
- **Classification** → output y belongs to one of C predetermined classes ($y \in \{1, \dots, C\}$)

Classification problem

Given:

- observations
  $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathbb{R}^D$
- set of possible classes
  $C = \{1, \dots, C\}$
- labels
  $y = \{y_1, y_2, \dots, y_N\}$, $y_i \in C$

Find:

- function $f: \mathbb{R}^D \to C$ that maps observations x_i to class labels y_i:
  $y_i = f(x_i)$ for $i \in \{1, \dots, N\}$

Error: Zero-one loss

Denotes the number of misclassified samples $\hat{y}$ → prediction

$$l_{01}(y, \hat{y}) = \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i \neq y_i)$$

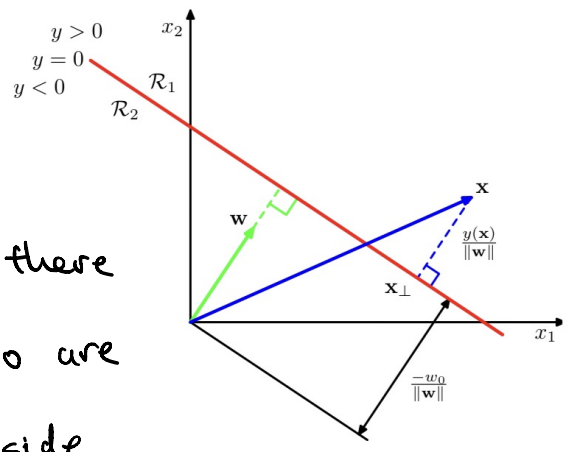
$$\mathbb{I}(a) \begin{cases} 1 \to \text{if } a \text{ is True} \\ 0 \to \text{else} \end{cases}$$

Hyperplane as a decision boundary

We can try to separate points from the classes by a hyperplane

Hyperplane is defined by a normal vector w and an offset w_0

$$w^T x + w_0 \begin{cases} = 0 & \text{if } x \text{ on the plane} \\ > 0 & \text{if } x \text{ on normal's side} \\ < 0 & \text{else} \end{cases}$$



A data set $D = \{(x_i, y_i)\}$ is **linearly separable** if there exists a hyperplane for which all x_i with $y_i = 0$ are on one side and all x_i with $y_i = 1$ on the other side

Perceptron

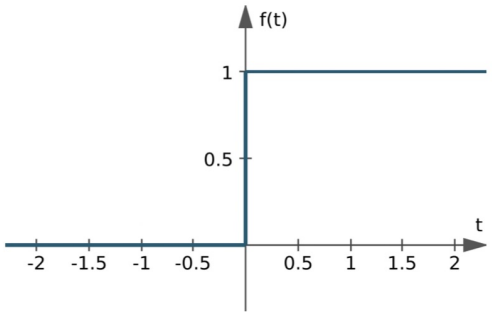
One of the oldest methods for binary classification

Decision rule:

$$\hat{y} = f(w^T x + w_0)$$

f is the step function defined as:

$$f(t) = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$



Learning rule:

1. Initialize parameters to any value, e.g., a zero vector: $w, w_0 \leftarrow 0$
2. For each misclassified sample x_i, in the training set update

$$w \leftarrow \begin{cases} w + x_i & \text{if } y_i = 1 \\ w - x_i & \text{if } y_i = 0 \end{cases}$$

$$w_0 \leftarrow \begin{cases} w_0 + 1 & \text{if } y_i = 1 \\ w_0 - 1 & \text{if } y_i = 0 \end{cases}$$

until all samples are classified correctly

This method takes a finite number of steps to converge to a (w, w_0) discriminating between two classes if exists.

If data is not linearly separable → Basis functions

Apply a nonlinear transformation $\phi: \mathbb{R}^D \to \mathbb{R}^M$

But there are limitations for hard-decision based classifiers such as:
no measure of uncertainty / can't handle noisy data / poor generalization / difficult to optimize

→ Solution: Probabilistic models for classification

label y, given data x:

$$p(y=c \mid x) = \frac{p(x \mid y=c) \cdot p(y=c)}{p(x)}$$

Probabilistic generative models for linear classification

- Generative model

$$p(y=c \mid x) \propto \underbrace{p(x \mid y=c)}_{\text{class conditional}} \cdot \overbrace{p(y=c)}^{\text{prior (a priori probability of a point belonging to class C)}}$$

class conditional
(probability of generating a point x, given that it belongs to class c)

$p(x \mid y=c, \psi)$
$p(y=c \mid \theta)$
} estimate the parameters of our model $\{\psi, \theta\}$ from the data D

1. Learning (using maximum likelihood - obtain estimates $\{\hat{\psi}, \hat{\theta}\}$)

$$p(y=c \mid x, \hat{\psi}, \hat{\theta}) \propto p(x \mid y=c, \hat{\psi})\, p(y=c \mid \hat{\theta})$$

2. Inference

Additionally, we can generate new data

$$y_{new} \sim p(y \mid \hat{\theta})$$
$$x_{new} \sim p(x \mid y = y_{new}, \hat{\psi})$$

$$\theta_c^{MLE} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = c) \qquad \| \quad p(x \mid y=c) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu_c)^T \Sigma^{-1} (x-\mu_c) \right\}$$

$$= \mathcal{N}(x \mid \mu_c, \Sigma)$$

multivariate normal distribution

$$p(y=1 \mid x) = \frac{1}{1+\exp(-a)} =: \sigma(a)$$

$$a = \log \frac{p(x \mid y=1)\, p(y=1)}{p(x \mid y=0)\, p(y=0)} = w^T x + w_0$$

$$\begin{cases} w = \Sigma^{-1} (\mu_1 - \mu_0) \\ w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 + \log \frac{p(y=1)}{p(y=0)} \end{cases}$$

Posterior distribution

$$p(y=1|x) = \frac{1}{1+\exp[-(w^T x + w_0)]} = \sigma(w^T x + w_0)$$

$$y|x \sim \text{Bernoulli}(\sigma(w^T x + w_0))$$