# Classification binaire

- $(X, Y)$ couple de v.a. de distribution $P$ **inconnue**
- $Y \in \{-1, +1\}$ est une **étiquette** binaire
- $X \in \mathbb{R}^d$ modélise une observation permettant de prédire $Y$

## Objectif

Construire $g : \mathbb{R}^d \to \mathbb{R}$ de façon à minimiser l'**erreur de classification**

$$L(g) = \mathbb{P}\{Y \neq g(X)\}$$

## Exemples

- Diagnostic médical
- Ciblage marketing
- Risque de crédit

# Règle optimale



- **Probabilité a posteriori**

$$\forall x \in \mathbb{R}^d, \quad \eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$$

- **Règle naïve de Bayes**

$$\forall x \in \mathbb{R}^d, \quad g^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1$$

## Risque minimum

$$L(g^*) = \min_g L(g)$$

- L'erreur $L(g)$ est **inconnue**... comme $P$ !

# Apprentissage statistique

## Minimisation du Risque Empirique

- Données observées

$$(X_1, Y_1) \ \ldots, \ (X_n, Y_n)$$

- On cherche à concevoir un algorithme qui "minimise" le **risque empirique**

$$\widehat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \neq g(X_i)\}$$

sur une collection "pas trop complexe" de règles $\mathcal{G}$

- Minimiseur du risque empirique sur $\mathcal{G}$

$$\widehat{g}_n = arg \min_{g \in \mathcal{G}} \widehat{L}_n(g)$$

# Minimisation du Risque Empirique - Propriétés

$$L(\hat{g}_n) - L(g^*) \leq 2 \sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid + \left( \inf_{g \in \mathcal{G}} L(g) - L^* \right)$$

**Contrôle de la complexité :**

$$\mathbb{E}\left[ \sup_{g \in \mathcal{G}} \mid L_n(g) - L(g) \mid \leq C \sqrt{\frac{V}{n}} \right]$$

si la **dimension de Vapnik-Chervonenkis** de $\mathcal{G}$ est finie, égale à $V$.

Un exemple: l'algorithme du Perceptron pour la classification

**Classification**

Séparateurs affines

Le Perceptron mono-couche

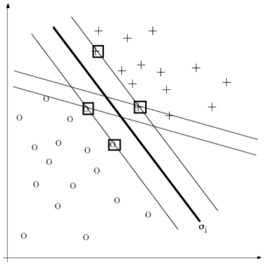
Stéphan Clémençon

# Hyperplans séparateurs

- On cherche à séparer l'espace d'entrée $\mathbb{R}^d$ en deux régions par un **hyperplan affine** : $(w, \theta) \in \mathbb{R}^d \times \mathbb{R}$

$$\mathcal{H}_{w,\theta} : \quad {}^t w \cdot X + \theta = 0$$

- Règle associée : $g_{w,\theta}(X) = sign({}^t w \cdot X + \theta)$



Un exemple: l'algorithme du Perceptron pour la classification

Classification

Séparateurs affines

Le Perceptron mono-couche

Stéphan Clémençon

# Le Perceptron de F. Rosenblatt (1957)

- On remplace

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{-Y_i({}^t w \cdot X_i + \theta) > 0\},$$

par la fonction **lisse**

$$-\frac{1}{n}\sum_{i=1}^{n} Y_i({}^t w \cdot X_i + \theta)$$

1. Initialisation : $(w, \theta)$ choisi au hasard
2. On tire au hasard $(X_i, Y_i)$ :
   - Si la règle courante classe correctement le point choisi, on ne modifie pas l'hyperplan
   - Sinon, on effectue une **descente de gradient** de pas $\rho$

$$\begin{pmatrix} w \\ \theta \end{pmatrix} \leftarrow \begin{pmatrix} w \\ \theta \end{pmatrix} + \rho\begin{pmatrix} y_i x_i \\ y_i \end{pmatrix}$$

# Propriétés

- L'algorithme converge ssi les données sont **linéairement séparables**
- Cas "linéairement séparable" : le nombre d'itérations de l'algorithme peut être contrôlé (Théorème de Novikoff, '62)
- L'algorithme s'adapte immédiatement au cas **séquentiel**
- Extensions : Réseaux de Neurones, Machines à vecteurs support (SVM)