Semaine IMT Grand-Est

# Introduction à l'apprentissage automatique, la science de l'intelligence artificielle

## Séance 4

## *Classification supervisée linéaire: le perceptron*

Frédéric Sur

https://members.loria.fr/FSur/enseignement/IMT_GE/

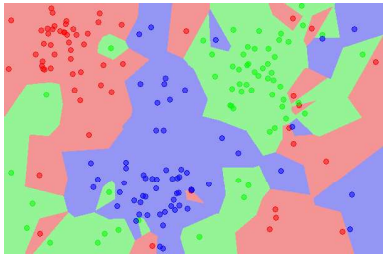


# Plan

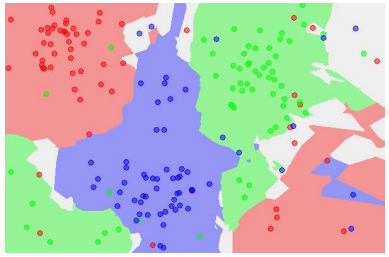




# Rappel : classification aux plus proches voisins



1-p.p.v.



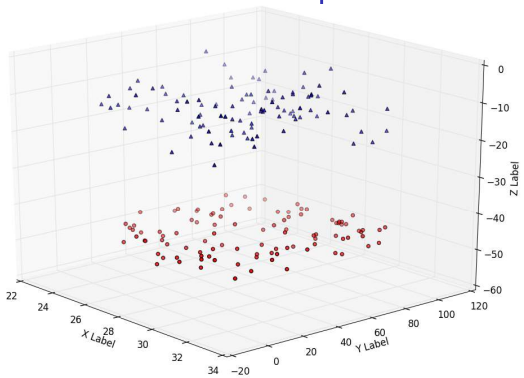
5-p.p.v.

**Inconvénients** : beaucoup trop de « paramètres », algorithmiquement lourd, frontières de séparation trop complexes

**Objectif dans la suite du cours** : classification robuste au « bruit », nécessitant peu de paramètres, algorithmiquement efficace



# Données linéairement séparables




Données : $(x_i, y_i)$

$x_i \in \mathbb{R}^d$

$y_i \in \{-1, 1\}$

Fonction discriminante : $f(x) = \beta_0 + \beta_1 \cdot x \quad (\beta_0 \in \mathbb{R},\ \beta_1 \in \mathbb{R}^d)$

autre notation : $f(x) = \beta_0 + \beta_1^T x \quad (= \beta_0 + \sum_{j=1}^{d} \beta_{1,j} x_j)$

Règle de classification :

$f(x) > 0 \Rightarrow x \in \mathcal{C}_1$, et $f(x) < 0 \Rightarrow x \in \mathcal{C}_{-1}$,

**Question du jour** : comment trouver un (hyper-)plan séparateur ?

→ c'est à dire $\beta_0$ et $\beta_1$ permettant de classer une nouvelle observation $x$ selon le signe de $\beta_0 + \beta_1 \cdot x$

## Moindres carrés ?

$(\beta_0, \beta_1)$ minimisant $\sum_{i=1}^{N} \|\beta_0 + \beta_1 \cdot x_i - y_i\|^2$

$\rightarrow$ pas une bonne idée...

**Problème** : influence des observations "trop correctes"...

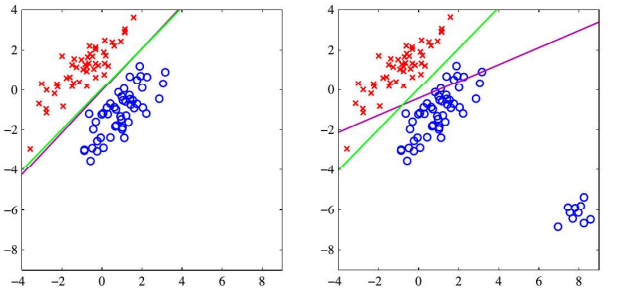


**Figure 4.4** The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

Illustration : C. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006



## Rappel : classifieur de la régression logistique

**Classification de la régression logistique** :
$x$ dans $\mathcal{C}_1 \iff p(\mathcal{C}_1|x) > 1/2 \iff f(x) = \beta_0 + \beta_1 \cdot x > 0$
$\rightarrow$ il s'agit d'un classifieur linéaire
(les deux classes sont séparées par un plan)

**Remarque** : il est possible d'interpréter la sortie $f(x)$ de la régression logistique avec : $\sigma(f(x)) = p(\mathcal{C}_1|x)$

**Interprétation** de la maximisation de la log-vraisemblance cond. :

$$\ell_{(x_i,y_i)_{1\leq i\leq N}}(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \log(\sigma(f(x_i))) + (1 - y_i) \log(\sigma(-f(x_i)))$$

- si $y_i = 1$ on a intérêt à avoir $f(x_i) = \beta_0 + \beta_1 \cdot x_i > 0$
- si $y_i = 0$ on a intérêt à avoir $f(x_i) = \beta_0 + \beta_1 \cdot x_i < 0$
- la sigmoïde est peu sensible à l'éloignement de $x_i$ au plan $[f = 0]$

(cf. distance Euclidienne signée : $d(x, P) = (\beta_0 + \beta_1 \cdot x)/||\beta_1||$)



## Plan

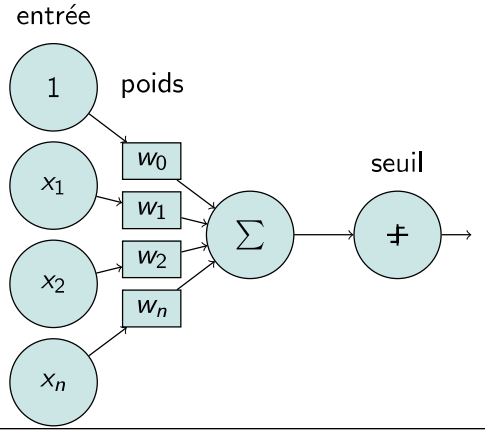




## Le perceptron (neurone artificiel)

**Rappel** : classifieur linéaire
$w_0 + w_1 x_1 + \cdots + w_n x_n > 0 : x \in \mathcal{C}_1$
$w_0 + w_1 x_1 + \cdots + w_n x_n < 0 : x \in \mathcal{C}_{-1}$

**Perceptron** : Frank Rosenblatt 1950-1960
$\rightarrow$ naissance du *machine learning* et de l'IA




**Apprentissage** :
trouver les $w_i$

**Nouvelle observation** :
$(x_1, \ldots, x_n) \in \mathcal{C}_{-1}$ ou $\mathcal{C}_1$ ?
$\rightarrow$ il suffit de calculer
la sortie du perceptron

## Algorithme d'apprentissage du perceptron

Base d'apprentissage : $N$ observations $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$

Remarque : quitte à considérer les $x_i \in \mathbb{R}^{d+1}$ avec première composante $= 1$, on peut toujours se ramener à des hyperplans d'équation $w \cdot x$.

**Algorithme d'apprentissage** (Rosenblatt 1957)

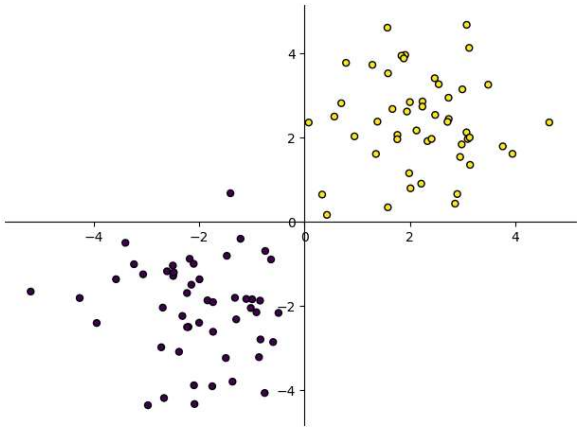
Initialisation : $w = 0$
Tant qu'il y a des mises à jours :
  Pour i $= 1$ à $N$ :
    $\widehat{y}_i = \text{signe}(w \cdot x_i)$
    Si $\widehat{y}_i \neq y_i$ alors : $w \leftarrow w + y_i x_i$
Retourner $w$

**Remarque** : $w$ déterminé à une constante multiplicative près



## Illustration de l'algorithme d'apprentissage

jaune : 1 mauve : -1

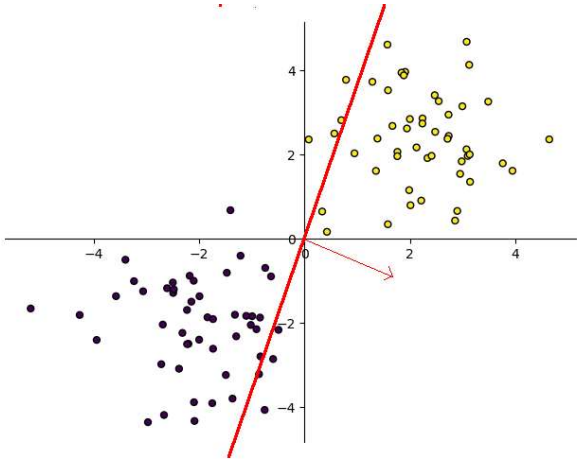


Données



## Illustration de l'algorithme d'apprentissage

jaune : 1 mauve : -1

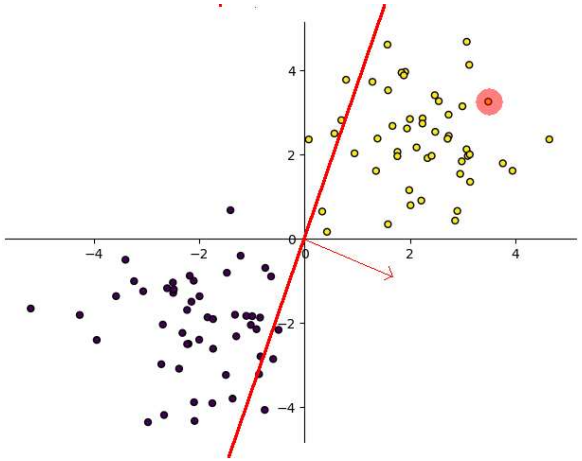


Droite séparatrice à l'étape $n$



## Illustration de l'algorithme d'apprentissage

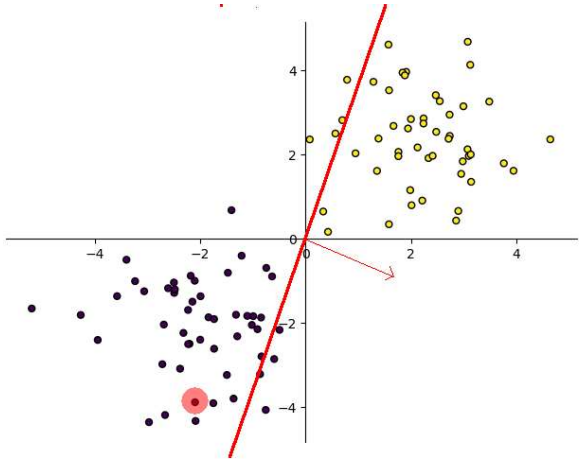
jaune : 1 mauve : -1



$y_i = 1$, $\widehat{y}_i = 1$ : pas de mise à jour

## Illustration de l'algorithme d'apprentissage

**jaune** : 1 **mauve** : -1

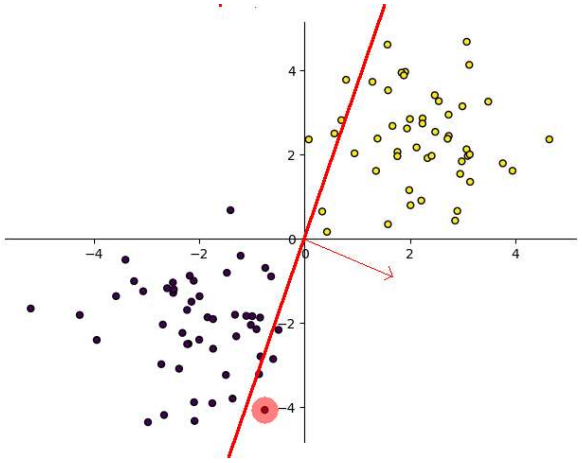


$y_i = -1$, $\widehat{y}_i = -1$ : pas de mise à jour



## Illustration de l'algorithme d'apprentissage

**jaune** : 1 **mauve** : -1

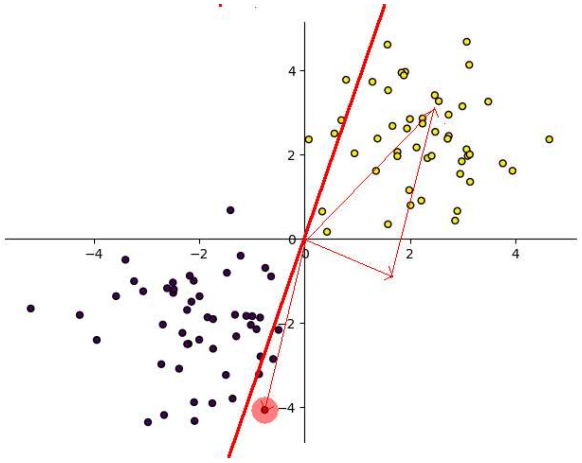


$y_i = -1$, $\widehat{y}_i = 1$ : mise à jour de $w$



## Illustration de l'algorithme d'apprentissage

**jaune** : 1 **mauve** : -1

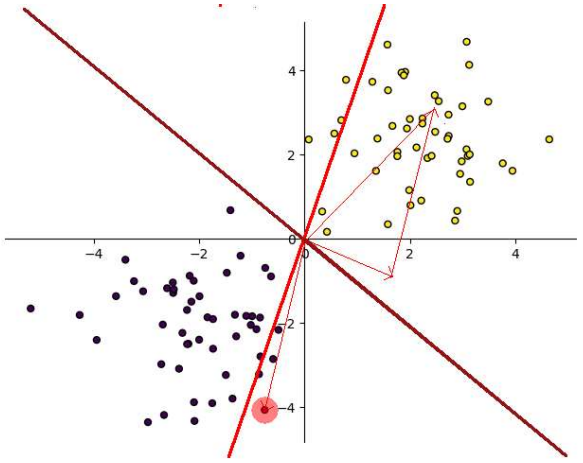


$w \leftarrow w + y_i x_i$



## Illustration de l'algorithme d'apprentissage

**jaune** : 1 **mauve** : -1



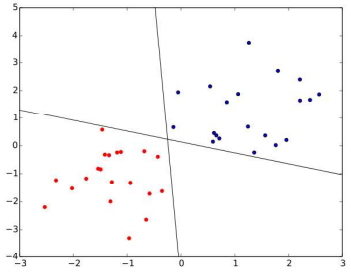
Droite séparatrice à l'étape $n + 1$

## Convergence de l'algorithme du perceptron

**Théorème** (Novikoff 1962) : si les données sont linéairement séparables, alors l'algorithme du perceptron **converge en un nombre fini d'étapes** vers un hyperplan séparateur.

**Remarque** :
vers **un** des hyperplans...

c'est un problème pour la prédiction de la classe de nouvelles observations



By Qwertyus - Own work, CC BY-SA 4.0
https://commons.wikimedia.org/w/index.php?curid=41351528

→ si données pas séparables, pas de convergence
→ il faut alors arrêter après un certain nombre de parcours des données (*epoch*)



## Plan





## Cas de la classification multiclasse

classes : $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ $(K > 2)$

- « **Nativement** » **multiclasse :**
Naive Bayes, plus proches voisins...

- **Algorithmes dédiés** :
régression logistique multiclasse, perceptron multiclasse...

- **À partir d'un classifieur biclasse** :

*One-vs-all* : entraînement de $K$ classifieurs $f_k$ discriminant entre $\mathcal{C}_k$ et l'ensemble des autres classes.
Puis : $f(x) = \text{argmax}_k f_k(x)$
**Problème** : échelle des $f_k$ ?

*One-vs-one* : $K(K-1)/2$ classifieurs discriminant entre $\mathcal{C}_i$ et $\mathcal{C}_j$
Puis : vote
**Problème** : complexité, ambiguïté ?

- Algorithme exact dans `scikit-learn` à vérifier :
`http://scikit-learn.org/stable/modules/multiclass.html`



## Le perceptron multiclasse

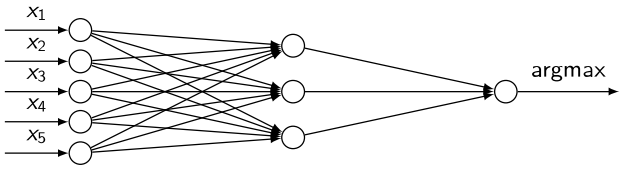



**Figure 4.8** Illustration of the Mark 1 perceptron hardware. The photograph on the left shows how the inputs were obtained using a simple camera system in which an input scene, in this case a printed character, was illuminated by powerful lights, and an image focussed onto a $20 \times 20$ array of cadmium sulphide photocells, giving a primitive 400 pixel image. The perceptron also had a patch board, shown in the middle photograph, which allowed different configurations of input features to be tried. Often these were wired up at random to demonstrate the ability of the perceptron to learn without the need for precise wiring, in contrast to a modern digital computer. The photograph on the right shows one of the racks of adaptive weights. Each weight was implemented using a rotary variable resistor, also called a potentiometer, driven by an electric motor thereby allowing the value of the weight to be adjusted automatically by the learning algorithm.

Illustration : C. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006

## Battage médiatique autour du perceptron

1957 : **Perceptron** - Frank Rosenblatt (1928-1971)

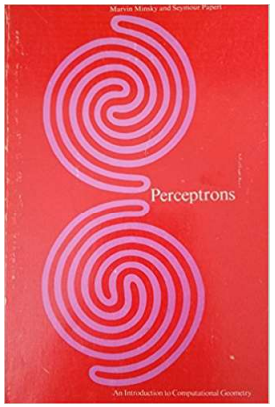
> ***NEW NAVY DEVICE LEARNS BY DOING; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser***
>
> SPECIAL TO THE NEW YORK TIMES JULY 8, 1958
>
> WASHINGTON, July 7 (UPI) -- The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The New Yorker, 6 décembre 1958 :

« *Dr. Rosenblatt defined the perceptron as the first non-biological object which will achieve an organization of its external environment in a meaningful way. [...] It can tell the difference betwen a cat and a dog [...]. Right now it is of no practical use, Dr. Rosenblatt conceded, but he said that one day it might be useful to send one into outer space to take in impressions for us.* »



## « Critique » du perceptron




*Perceptrons* - 1969

Marvin Minsky (1927-2016), MIT
Seymour Papert (1928-2016), MIT (Logo)

→ AI winter

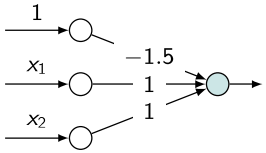


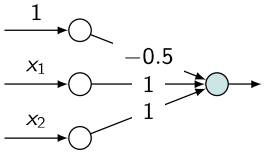
## Perceptron et relations logiques

Tables logiques :

| $x_1$ | | $x_2$ | $z$ |
|---|---|---|---|
| 0 | AND | 0 | 0 |
| 0 | AND | 1 | 0 |
| 1 | AND | 0 | 0 |
| 1 | AND | 1 | 1 |

| $x_1$ | | $x_2$ | $z$ |
|---|---|---|---|
| 0 | OR | 0 | 0 |
| 0 | OR | 1 | 1 |
| 1 | OR | 0 | 1 |
| 1 | OR | 1 | 1 |




→ représentation graphique ?



## Perceptron et « ou exclusif »

| $x_1$ | | $x_2$ | $z$ |
|---|---|---|---|
| 0 | XOR | 0 | 0 |
| 0 | XOR | 1 | 1 |
| 1 | XOR | 0 | 1 |
| 1 | XOR | 1 | 0 |

→ représentation graphique ?

→ limite rédhibitoire ?

## Plan





## Conclusion

On cherche des modèles de classification supervisée robustes, avec peu de paramètres et un algorithme d'apprentissage simple.
$\rightarrow$ intérêt des classifieurs linéaires

**Suite du cours** :

- ambiguïté de l'hyperplan séparateur ?
  quel est le "meilleur" hyperplan ?
  $\rightarrow$ régression logistique
  $\rightarrow$ **Machines à Vecteurs Supports** (SVM)
- données non linéairement séparables ?
  $\rightarrow$ hyperplan optimal robuste aux mauvaises classifications
  $\rightarrow$ **perceptron multicouche** (réseaux de neurones)
  $\rightarrow$ **machines à noyau**
  ($k(x, y) = \phi(x) \cdot \phi(y)$, avec $\phi$ plongement des observations dans un espace de $\ll$ plus grande dimension $\gg$ ; généralise la notion de produit scalaire)