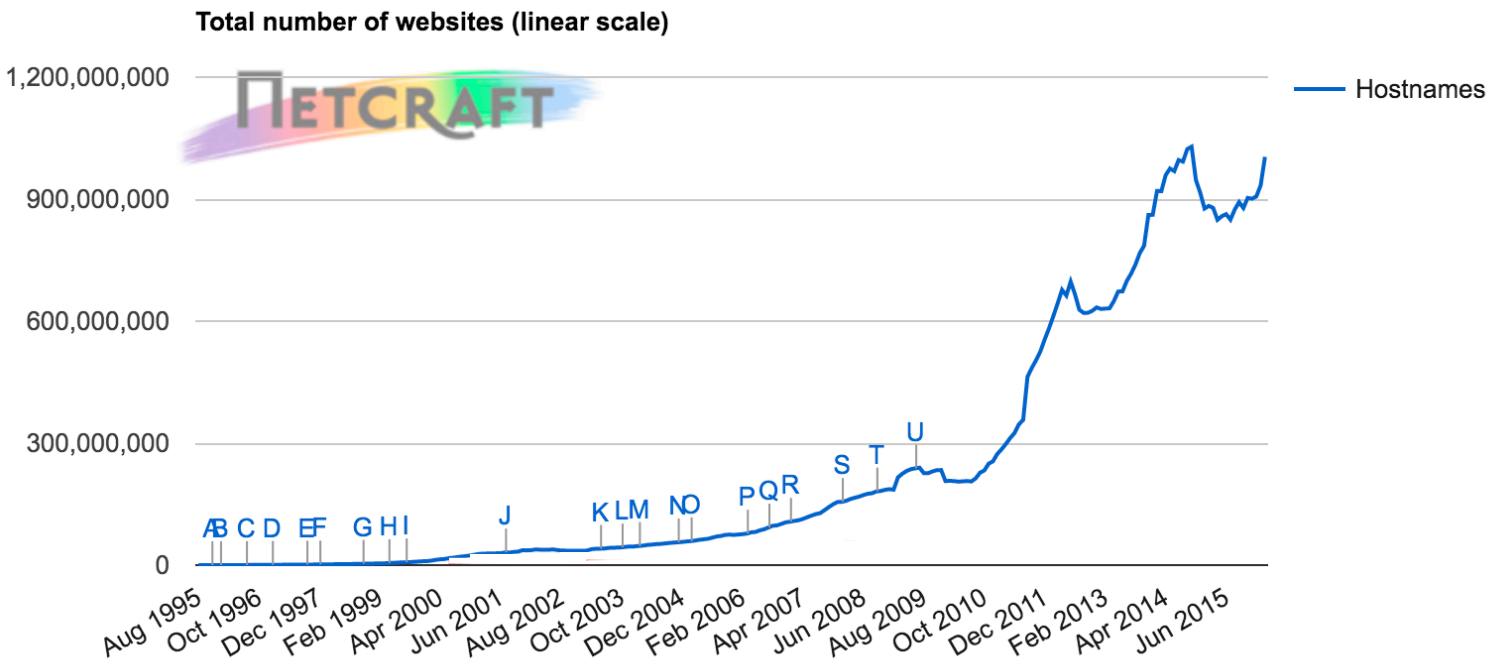
Modèles supervisés pour Data-Science

Massih-Reza Amini

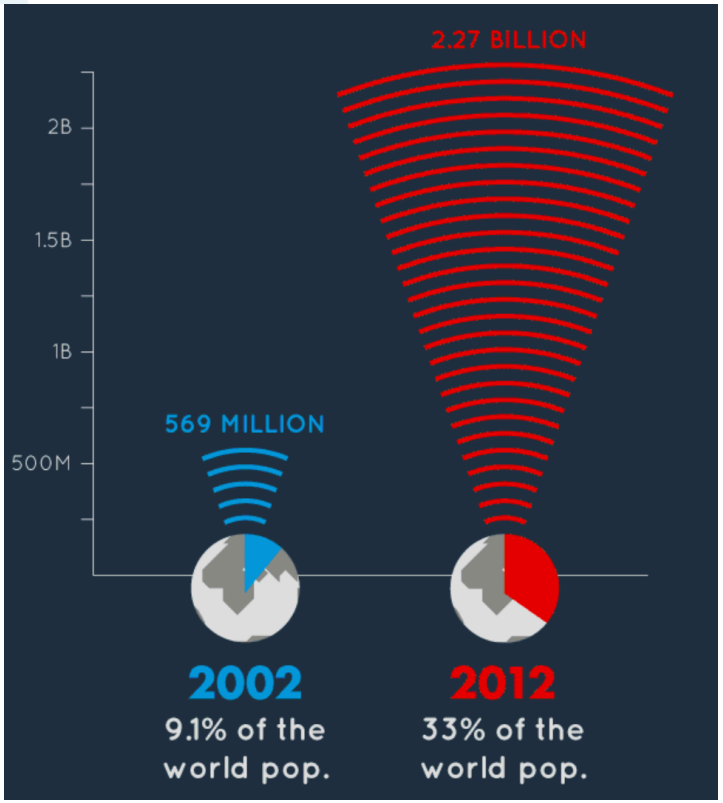
Université Grenoble Alpes
Laboratoire d'Informatique de Grenoble
Massih-Reza.Amini@imag.fr

Ère de Big Data

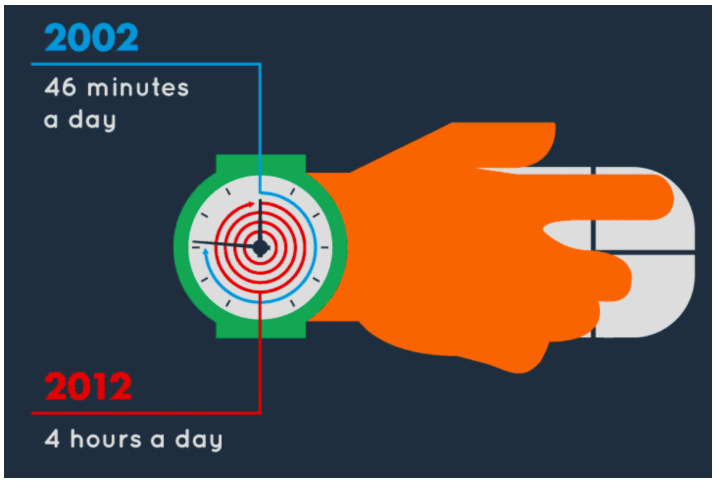
- En grande partie dû au développement rapide du Web ces 20 dernières années,



Big Data: nouvelles habitudes et pratiques

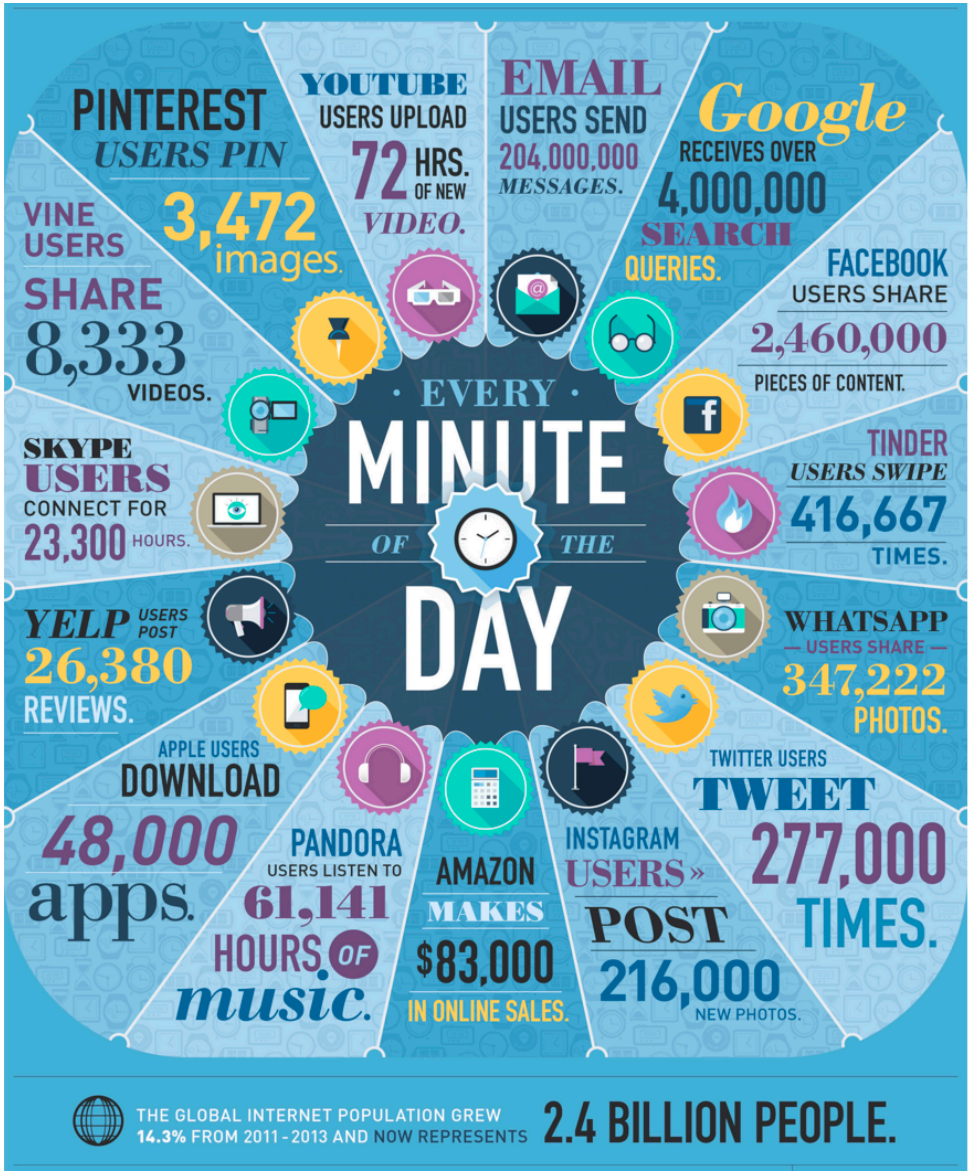


Nombre d'internautes



Temps de connexion

Big Data: génération accrue de données



Big Data: valeur des données

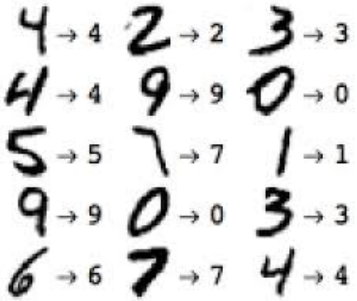
- D'après les prévisions du projet EMC[1], en 2020 il y aura 40 zetta octets (40×10^{21} octets) de données non-structurées sur la Toile.
- Ces données sont considérées comme le pétrole du XXI^e siècle.[2]
- Nécessité de développer de nouveaux outils automatiques pour la recherche et accès à l'information.

[1] http://www.emc.com/leadership/digital-universe/index.htm

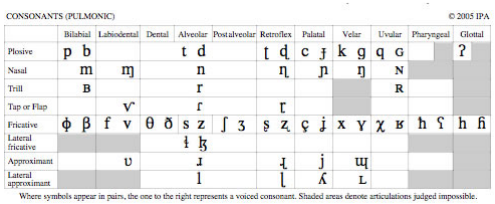
[2] http://www.lepoint.fr/technologie/les-data-petrole-du-xxie-siecle-14-03-2012-1441346_58.php

Programme

Spam detection
$K = 2$



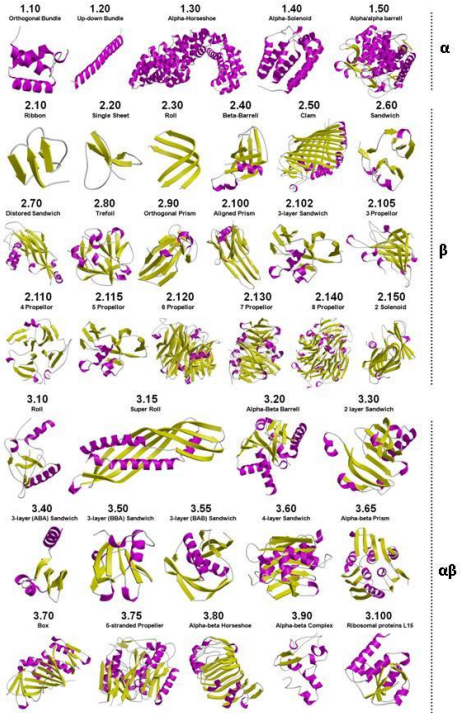
Digit recognition
$K = 10$


CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

Phoneme recognition
$K = 50$



Protein classification
$K \in [300 - 600]$

Programme

1. Modèles de classification
   - Perceptron,
   - Perceptron à marge,
   - Regression Logistique,
   - Adaboost.

Apprentissage et Inférence

Le processus de l'inférence est faite suivant trois étapes:

1. Observer un phénomène,
2. Construire un modèle associé au phénomène,
3. Faire des prédictions.

Apprentissage et Inférence

Le processus de l'inférence est faite suivant trois étapes:

1. Observer un phénomène,
2. Construire un modèle associé au phénomène,
3. Faire des prédictions.

- Ces trois étapes sont impliquées dans plus ou moins toutes les sciences naturelles!
  All that is necessary to reduce the whole nature of laws similar to those which Newton discovered with the aid of calculus, is to have a sufficient number of observations and a mathematics that is complex enough (Marquis de Condorcet, 1785)

Apprentissage et Inférence

Le processus de l'inférence est faite suivant trois étapes:

1. Observer un phénomène,
2. Construire un modèle associé au phénomène,
3. Faire des prédictions.

- Ces trois étapes sont impliquées dans plus ou moins toutes les sciences naturelles!
- Le but de l'apprentissage machine est d'automatiser ce processus.

Induction vs. déduction

- Induction est le processus de dériver des principes généraux à partir des faits particuliers ou instances.

- Déduction est, quant à elle, le processus de raisonner à partir d'un ensemble d'hypothèses pour en découler une conclusion, comme celui qui permet aux mathématiciens de prouver des théorèmes à partir des axiomes.

Exemple

En reconnaissance de formes:

- Les données sont constituées de paires d'exemples (Le vecteur de représentation d'une observation, son étiquette de classe),
- Les étiquettes de classes sont souvent $\mathcal{Y} = \{1, \ldots, K\}$ avec K (Classiquement les problèmes étudiés sont binaires $\mathcal{Y} = \{-1, +1\}$),
- L'algorithme d'apprentissage construit une fonction de prédiction qui associe un vecteur de représentation à son étiquette de classe,
- Objectif: Faire peu d'erreurs de prédiction sur les nouveaux exemples.

Reconnaissance des Formes (Exemple)

Classification des IRISes, Ronald Fisher (1936)

Iris Setosa

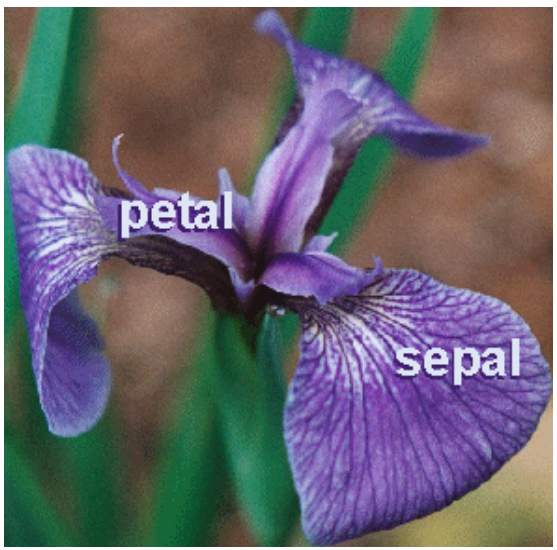
Iris Versicolor

Iris Virginica

Reconnaissance des Formes (Exemple)

- La première étape consiste à vectoriser la perception que nous avons des fleurs en utilisant des caractéristiques communes pertinentes.

- Cette étape requiert généralement la connaissance d'un expert.



Reconnaissance des Formes (Exemple)

- Si les observations proviennent d'un champs d'irises

Reconnaissance des Formes (Exemple)

- Si les observations proviennent d'un champs d'irises elles deviennent alors

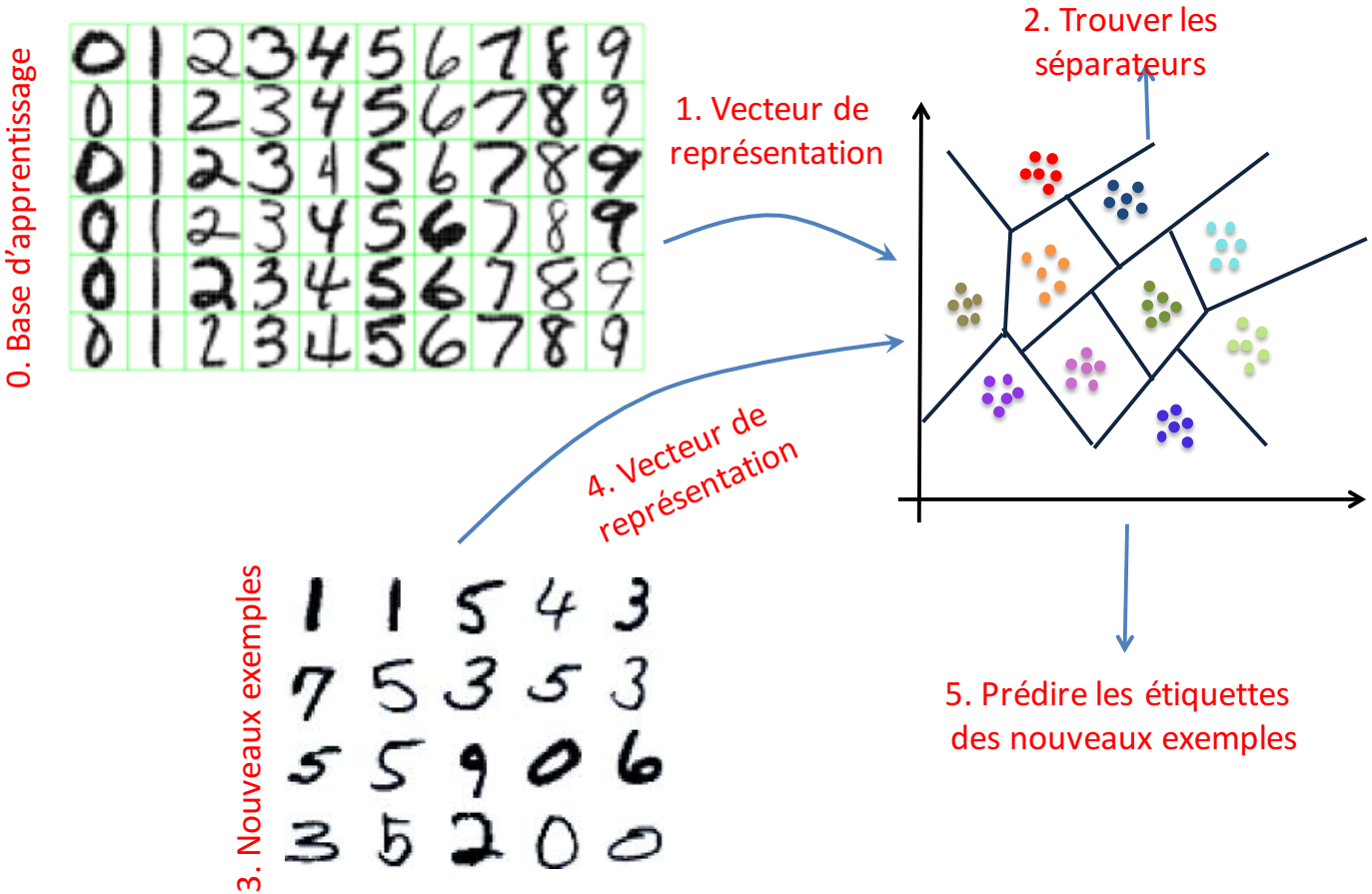
Fisher's *Iris* Data

longueur des sépales (en cm) *(Sepal length)*	largeur des sépales (en cm) *(Sepal width)*	longueur des pétales (en cm) *(Petal length)*	largeur des pétales (en cm) *(Petal width)*	Espèce *(Species)*
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
...				
7.0	3.2	4.7	1.4	*I. versicolor*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
...				
6.3	3.3	6.0	2.5	*I. virginica*
5.8	2.7	5.1	1.9	*I. virginica*
7.1	3.0	5.9	2.1	*I. virginica*
6.3	2.9	5.6	1.8	*I. virginica*
6.5	3.0	5.8	2.2	*I. virginica*

Reconnaissance des Formes (Exemple)

- La constitution des vecteurs d'observations ainsi que leurs étiquettes associées prend généralement beaucoup de temps.
- Beaucoup d'études s'intéressent maintenant à l'apprentissage de représentation en utilisant les réseaux de neurones profonds
- La deuxième étape consiste à apprendre une fonction de prédiction qui associe les entrées aux sorties

Reconnaissance des Formes

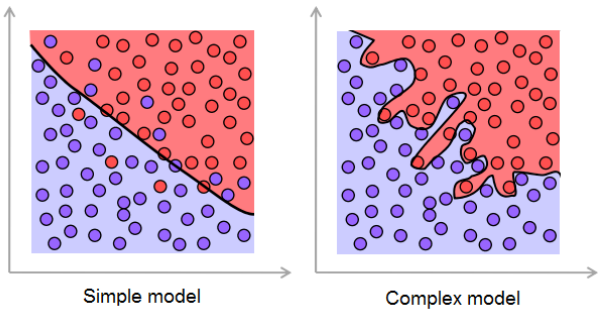


Approximation - Interpolation

Il est toujours possbile de construire une fonction qui s'ajuste exactement aux données.

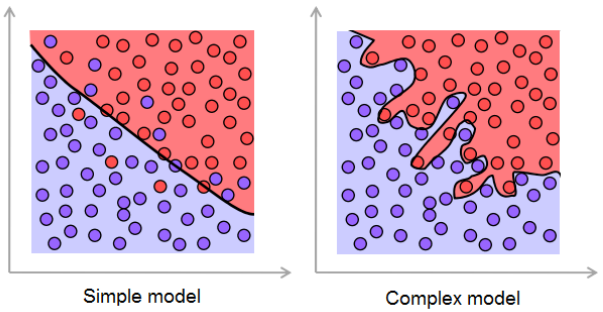
Approximation - Interpolation

Il est toujours possbile de construire une fonction qui s'ajuste exactement aux données.



Approximation - Interpolation

Il est toujours possbile de construire une fonction qui s'ajuste exactement aux données.



Est-il raisonnable?

Rasoir d'Occam

Idée: Chercher des régularités (ou des répétitions) dans le phénomène observé, la généralisation est faite depuis les observations passées aux nouvelles ⇒ Prendre le modèle le plus simple ...
Comment mesurer la simplicité ?

1. Nombre de constants,
2. Nombre de paramètres,
3. ...

Hypothèses de base

Deux types d'hypothèses:

- Les observations passées sont liées aux nouvelles
  → Le phénomène est stationnaire

- Les observations sont indépendamment générées à partir d'une source
  → Notion d'indépendance

Buts

→ Comment faire de la prédiction à partir des données passées? Quelles sont les hypothèses?

- Donner une définition formelle de l'apprentissage, de la généralisation, du sur-apprentissage,
- Caractériser la performance des algorithmes d'apprentissage,
- Construire de meilleurs algorithmes.

Modèle Probabiliste

Relations entre les observations passées et futures.

- Indépendance: Chaque nouvelle observation apporte un maximum d'information individuelle,
- identiquement distribuées: Les observations apporte de l'information sur le phénomène qui génère les observations.

Formellement

Nous considérons un espace d'entrée $\mathcal{X} \subseteq \mathbb{R}^d$ et un espace de sortie $\mathcal{Y}$.

Assumption: Les paires d'exemples $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ sont *identiquement* et *indépendamment* distribuées (`i.i.d`) d'après une distribution de probabilité fixe $\mathcal{D}$.

Samples: Nous observons une séquence de m paires d'exemples $(\mathbf{x}_i, y_i)$ générées `i.i.d` suivant $\mathcal{D}$.

Aim: Construire une fonctions de prédiction $f : \mathcal{X} \to \mathcal{Y}$ qui prédit la sortie y d'un nouvel exemple $\mathbf{x}$ avec la plus petite probabilité d'erreur.

Apprentissage supervisé

- Les modèles discriminants trouve directement la fonction de classification $f : \mathcal{X} \to \mathcal{Y}$ à partir d'une classe de fonctions $\mathcal{F}$;
- La fonction trouvée devrait être celle qui minimise la probabilité d'erreur

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}} L(x, y) = \int_{\mathcal{X}\times\mathcal{Y}} L(f(x), y)\, d\mathcal{D}(x, y)$$

Où L est la fonction de risque définie comme

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$

La fonction de risque usuellement considérée est l'erreur de classification:

$$\forall (x, y); L(f(x), y) = \mathbb{1}_{f(x)\neq y}$$

Où $\mathbb{1}_\pi$ est la fonction indicatrice.

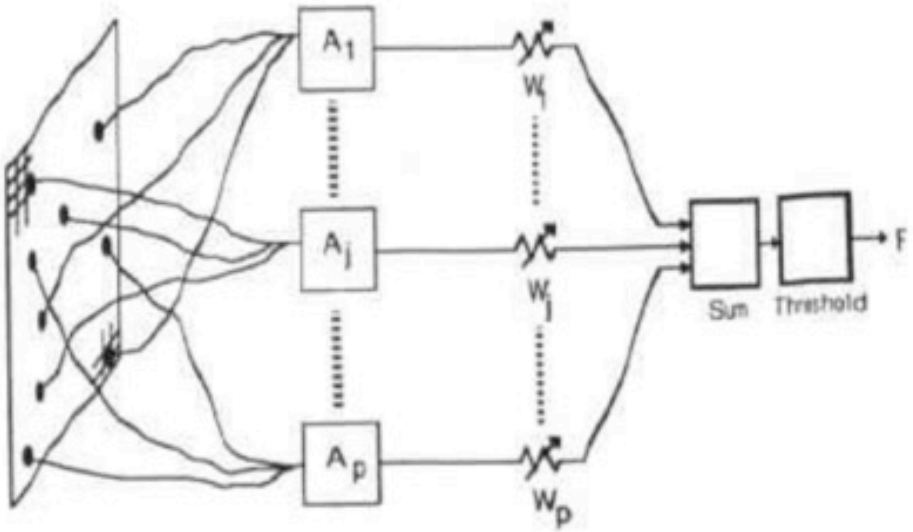
Principe de la minimisation du risque empirique

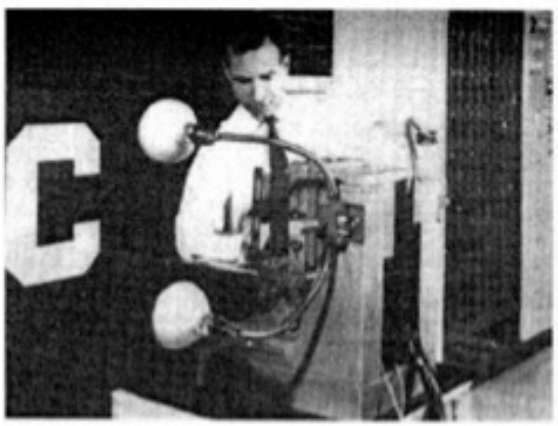
- Comme la distribution de probabilité $\mathcal{D}$ est inconnue, la forme analytique du risque ne peut pas être calculée, la fonction de prédiction ne peut ainsi être calculée directement de $R(f)$.
- Principe de la minimisation du risque empirique: Trouve f en minimisant l'estimateur non-biaisé de R sur une base d'apprentissage $S = (x_i, y_i)_{i=1}^m$:

$$\hat{R}_m(f, S) = \frac{1}{m} \sum_{i=1}^{m} L(f(x_i), y_i)$$

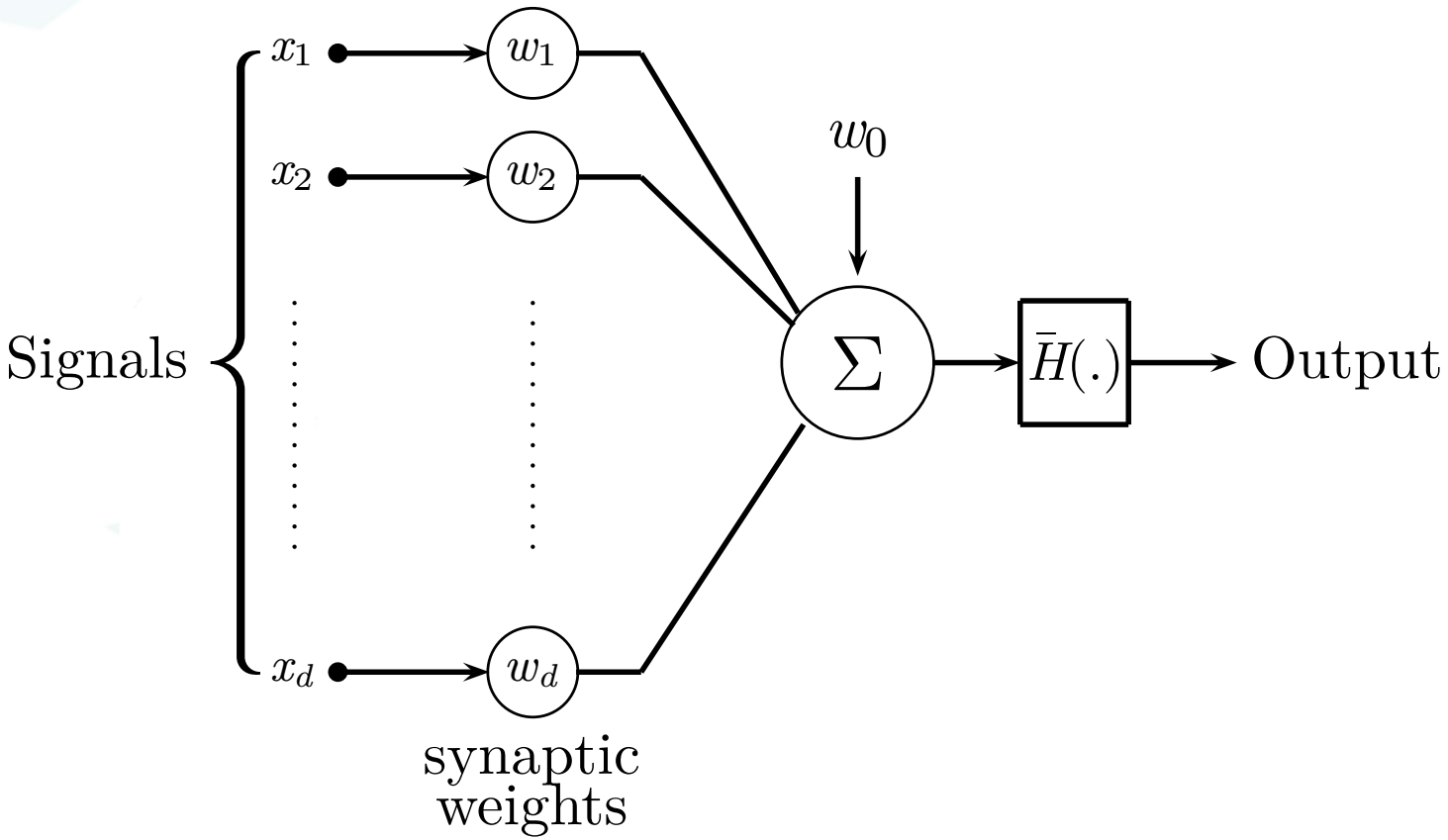
- Ceci ne peut pas se faire sans restreindre la classe de fonctions (rasoir d'Occam) ...

Perceptron [Rosenblatt, 1958]





Perceptron [Rosenblatt, 1958]



- Fonction de prédiction linéaire

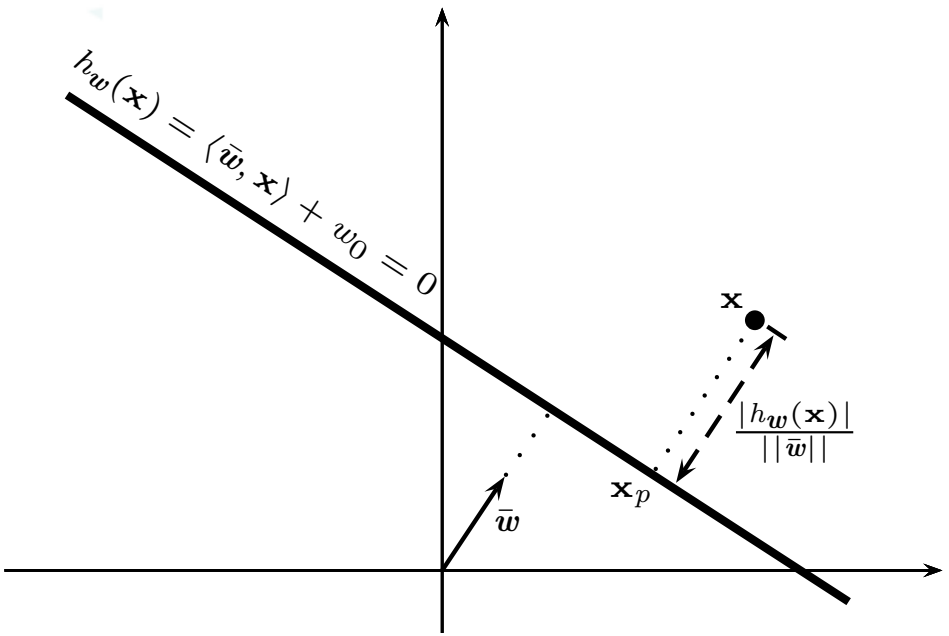
$$\begin{aligned} h_{\boldsymbol{w}} : \mathbb{R}^d &\rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto \langle \bar{\boldsymbol{w}}, \mathbf{x} \rangle + w_0 \end{aligned}$$

Perceptron [Rosenblatt, 1958]

- Fonction de prédiction linéaire

$$\begin{array}{rccc} h_{\boldsymbol{w}} : \mathbb{R}^d & \rightarrow & \mathbb{R} \\ \mathbf{x} & \mapsto & \langle \bar{\boldsymbol{w}}, \mathbf{x} \rangle + w_0 \end{array}$$

- Trouve les paramètres $\boldsymbol{w} = (\bar{\boldsymbol{w}}, w_0)$ en minimisant la distance entre les exemples mal-classés et la frontière de décision.



Apprendre les paramètres du modèle

- La fonction objectif

$$\hat{\mathcal{L}}(\boldsymbol{w}) = -\sum_{i' \in \mathcal{I}} y_{i'} \left(\langle \bar{\boldsymbol{w}}, \mathbf{x}_{i'} \rangle + w_0 \right)$$

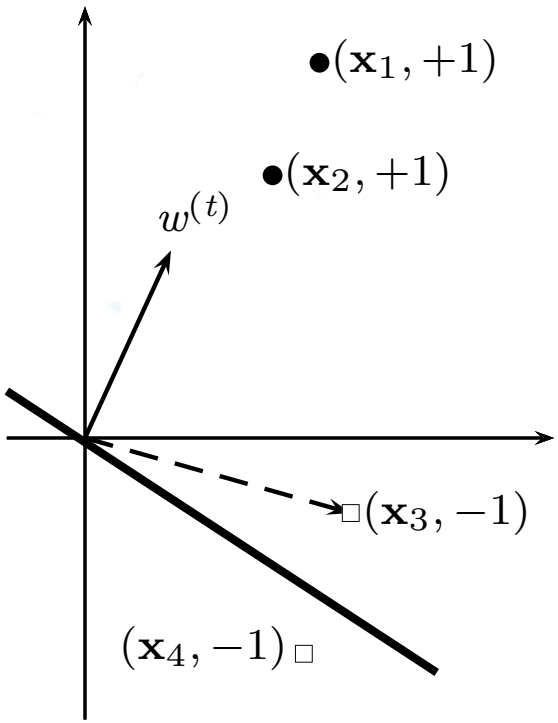
- Les dérivées de $\hat{\mathcal{L}}(\boldsymbol{w})$ par rapport aux paramètres

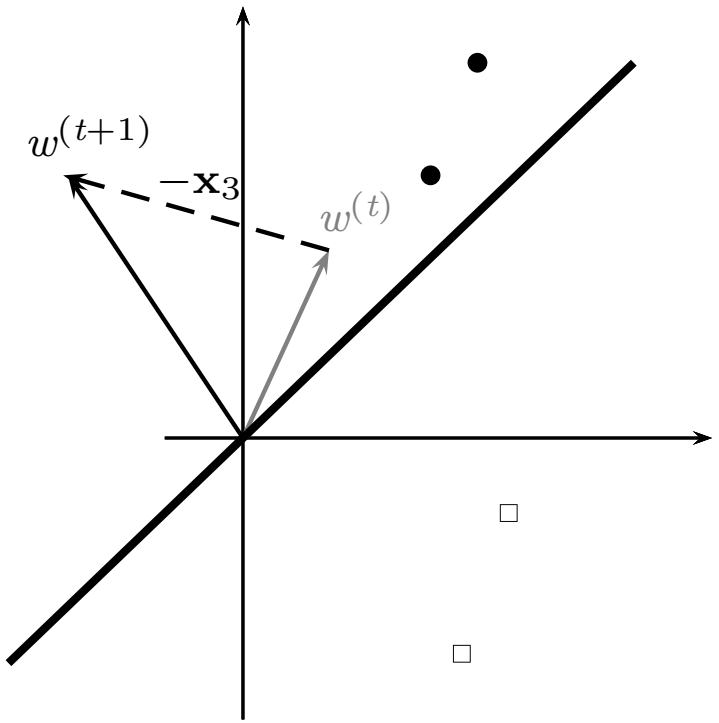
$$\begin{aligned} \frac{\partial \hat{\mathcal{L}}(\boldsymbol{w})}{\partial w_0} &= -\sum_{i' \in \mathcal{I}} y_{i'}, \\ \nabla \hat{\mathcal{L}}(\bar{\boldsymbol{w}}) &= -\sum_{i' \in \mathcal{I}} y_{i'} \mathbf{x}_{i'} \end{aligned}$$

- Perceptron: Mise à jour en-ligne des paramètres

$$\forall (\mathbf{x}, y), \text{ si } y(\langle \bar{\boldsymbol{w}}, \mathbf{x} \rangle + w_0) \leq 0 \text{ alors } \begin{pmatrix} w_0 \\ \bar{\boldsymbol{w}} \end{pmatrix} \leftarrow \begin{pmatrix} w_0 \\ \bar{\boldsymbol{w}} \end{pmatrix} + \eta \begin{pmatrix} y \\ y\mathbf{x} \end{pmatrix}$$

Mise à jour en-ligne des paramètres





Perceptron (algorithme)

Algorithm 1 L'algorithme de perceptron

1: Base d'apprentissage $S = \{(x_i, y_i) \mid i \in \{1, \ldots, m\}\}$
2: Initialiser les poids $w^{(0)} \leftarrow 0$
3: $t \leftarrow 0$
4: Pas d'apprentissage $\eta > 0$
5: **repeat**
6: Choisir aléatoire un exemple $(x^{(t)}, y^{(t)}) \in S$
7: **if** $y \left\langle w^{(t)}, x^{(t)} \right\rangle < 0$ **then**
8: $w_0^{(t+1)} \leftarrow w_0^{(t)} + \eta \times y^{(t)}$
9: $w^{(t+1)} \leftarrow w^{(t)} + \eta \times y^{(t)} \times x^{(t)}$
10: **end if**
11: $t \leftarrow t + 1$
12: **until** $t > T$

☞ Est-ce que cet algorithme converge?

References

Massih-Reza Amini
Apprentissage Machine de la théorie à la pratique
éditions Eyrolles, 2015.

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalker
Foundations of Machine Learning
2012.

F. Rosenblatt
The perceptron: A probabilistic model for information storage and organization in the brain.
Psychological Review, 65: 386–408.
1958