Un panorama sur

l' Apprentissage Artificiel

Approches, questions, enjeux

Antoine Cornuéjols

UMR AgroParisTech - INRA MIA-518

www.lri.fr/~antoine

antoine.cornuejols@agroparistech.fr

http://www.lri.fr/~antoine/Papers/Seminaires/AG-MIA-2010

The idea of a learning machine may appear paradoxical to some readers.

Alan Turing (1912 - 1954), 1950.

La prédiction est difficile, surtout lorsqu'il s'agit de l'avenir

Niels Bohr (1885 - 1962).

1. Introduction

1.1 Place de l'Apprentissage Artificiel

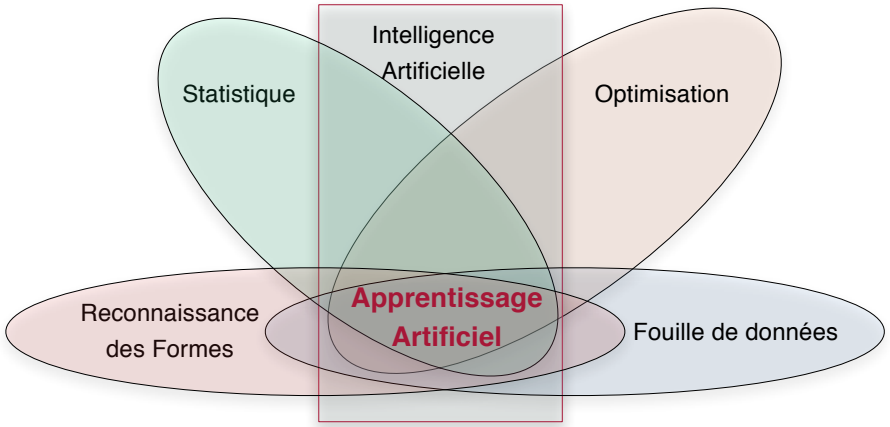
- **Qu'est-ce que c'est ?**
- **À quoi ça sert ?**
- **Comment ça marche ? Quels en sont les fondements ?**
- **En quoi est-ce singulier ?**
- **Quelles questions sont sur l'agenda de la « communauté apprentissage » ?**

1.1 Place de l'Apprentissage Artificiel

L'Apprentissage Artificiel ...

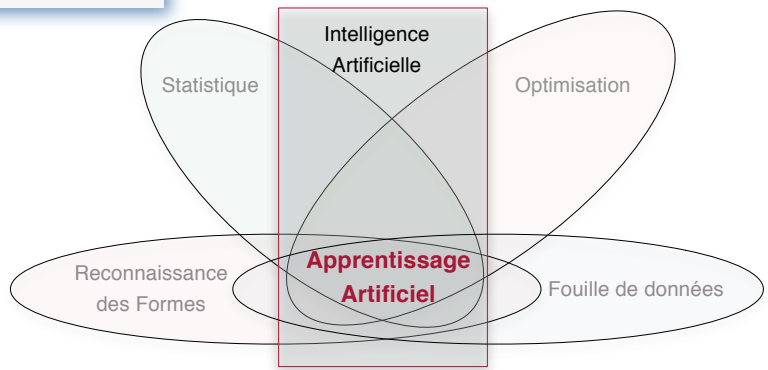
- **Science de modélisation**
  - Recherche des **régularités sous-jacentes** aux données d'observation
  - Cherche un modèle du monde permettant la **décision** et la **prédiction**
- Science de l'**adaptation**
  - Apprentissage par renforcement
  - Évolution simulée

1.1 Place de l'Apprentissage Artificiel



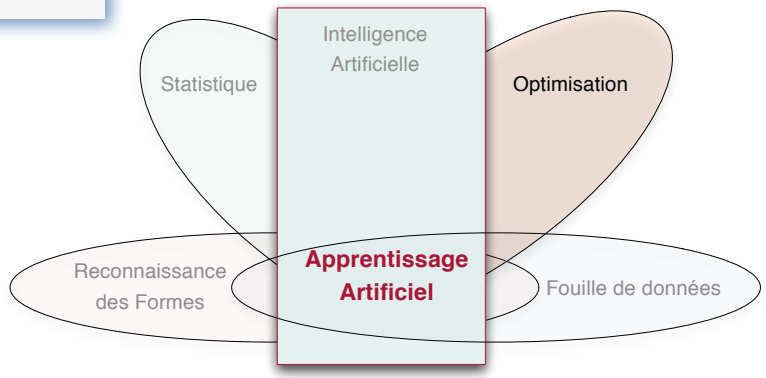
1.1 Place de l'Apprentissage Artificiel

- Représentation des connaissances
- Règles d'inférence
- Exploration de graphe



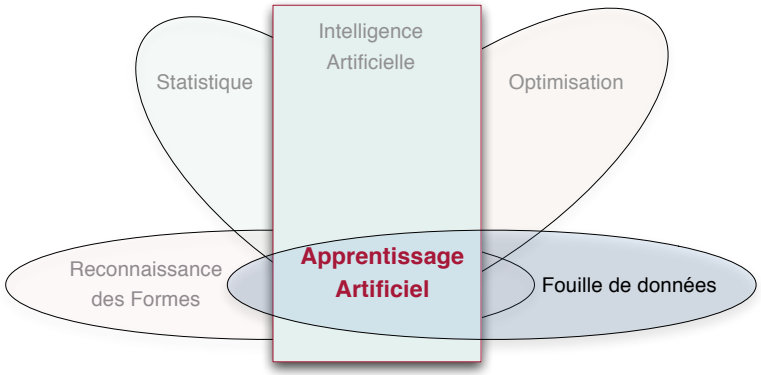
1.1 Place de l'Apprentissage Artificiel

- Résolution de problème inverse (régularisation)
- Méthodes d'optimisation
- Heuristiques



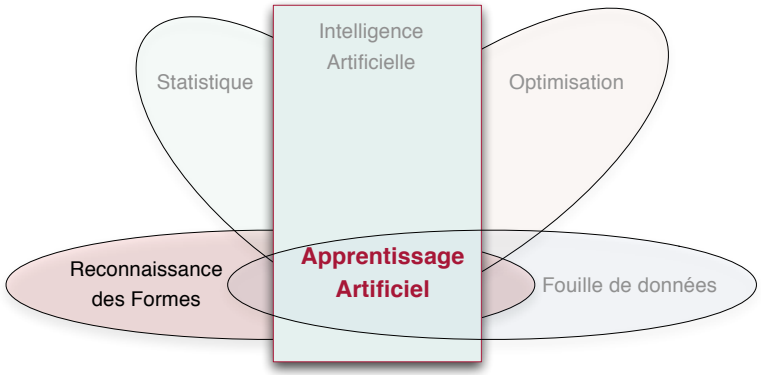
1.1 Place de l'Apprentissage Artificiel

- Mesures d'intérêt
- Algorithmes pour masses de données (vectorielles)



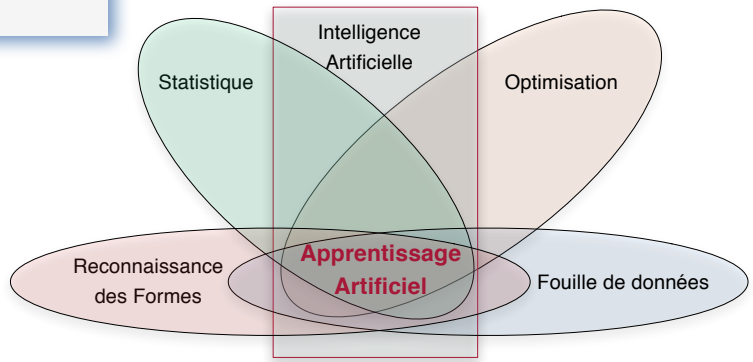
1.1 Place de l'Apprentissage Artificiel

- Apprentissage supervisé
- Algorithmes variés



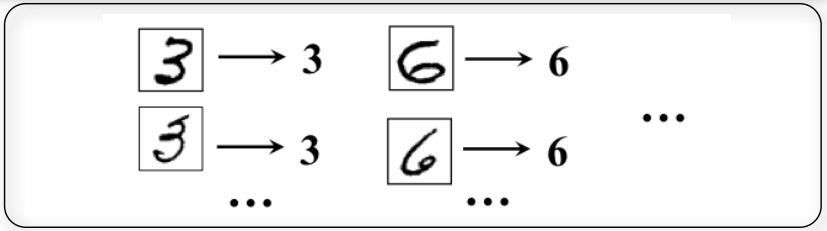
1.1 Place de l'Apprentissage Artificiel

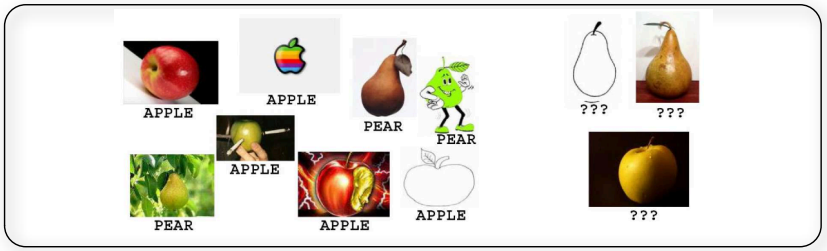
- **Recherche de régularités (modélisation du monde)**
- Capacité d'**adaptation** et d'**amélioration de performance**



1.2 Exemples d'applications

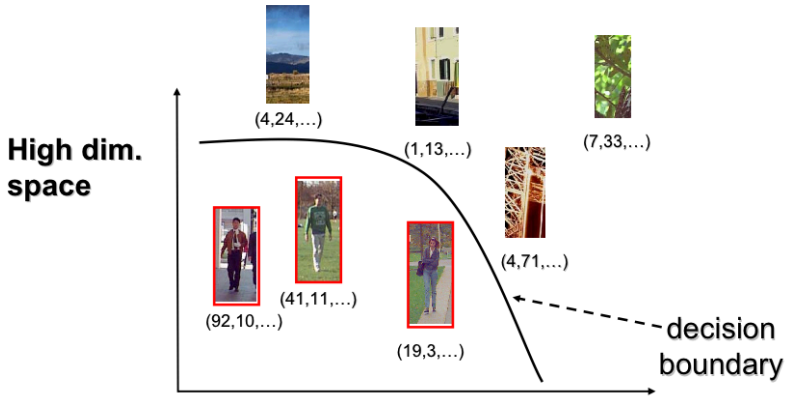
- **Reconnaissance de formes (apprentissage supervisé)**





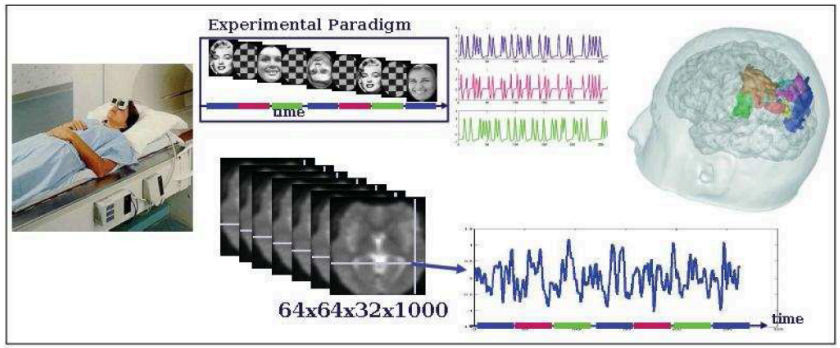
1.2 Apprentissage artificiel : exemples

- **Reconnaissance de formes (apprentissage supervisé)**



1.2 Apprentissage artificiel : exemples

- **Apprentissage supervisé : interprétation d'IRMf**

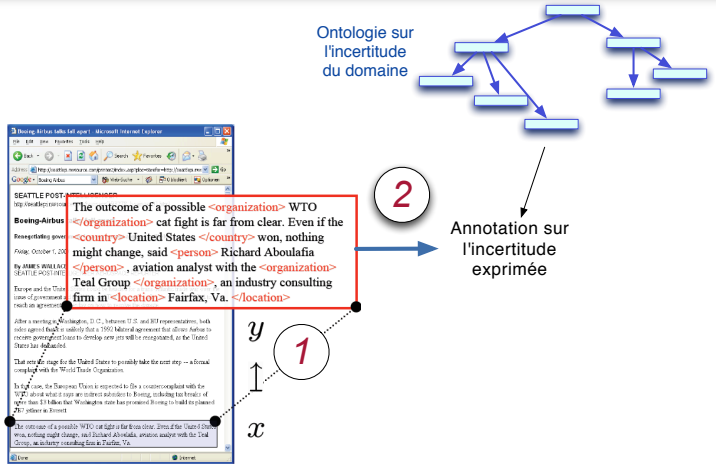


Trouble de la reconnaissance de visage ou *non*

1.2 Apprentissage artificiel : exemples

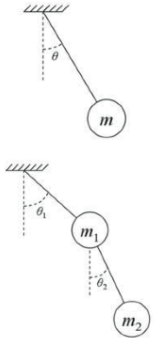
- **Apprentissage supervisé : Recherche d'information + annotation**

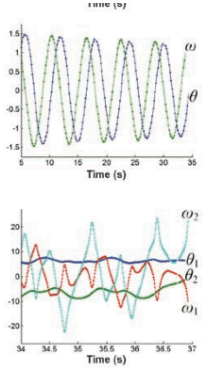
  Projet ANR Holyrisk (Met@risk + UMR MIA 518 + ...)



1.2 Apprentissage artificiel : exemples

- **Aide à la « découverte scientifique »**





$1.37 \cdot \omega^2 + 3.29 \cdot \cos(\theta)$
Lagrangian

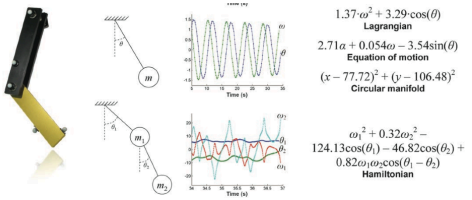
$2.71\alpha + 0.054\omega - 3.54\sin(\theta)$
Equation of motion

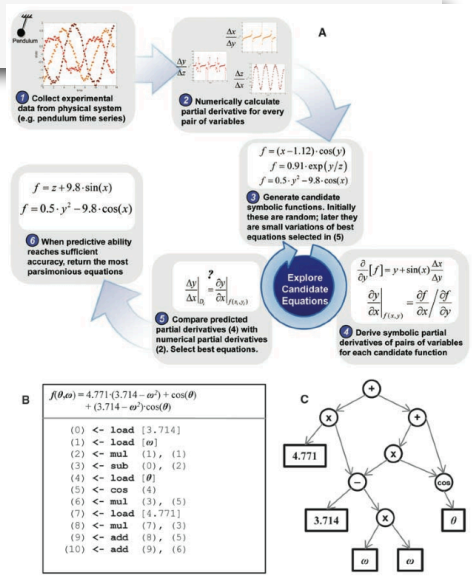
$(x - 77.72)^2 + (y - 106.48)^2$
Circular manifold

$\omega_1{}^2 + 0.32\omega_2{}^2 - 124.13\cos(\theta_1) - 46.82\cos(\theta_2) + 0.82\omega_1\omega_2\cos(\theta_1 - \theta_2)$
Hamiltonian

1.2 Apprentissage artificiel : exemples

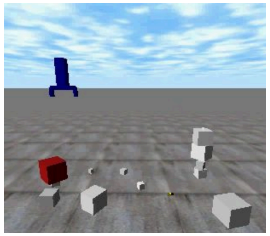
- **Aide à la « découverte scientifique »**





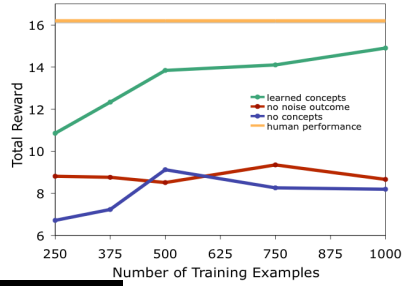
1.2 Apprentissage artificiel : exemples

- **Apprentissage par renforcement : planification pour robot**



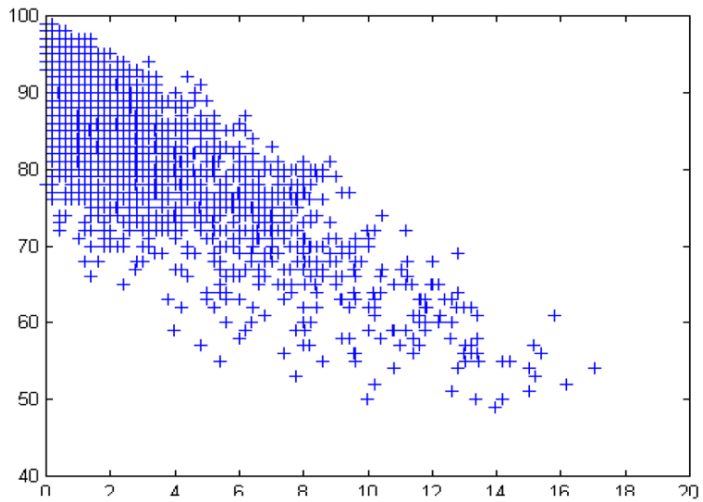
```
pickup(X): {Y: on(X,Y)}
    clear(X), inhand-nil, size(X)>2, size(X)<7→
      0.803 :¬on(X,Y)
      0.093 : no change
```

picking up middle-sized blocks usually works

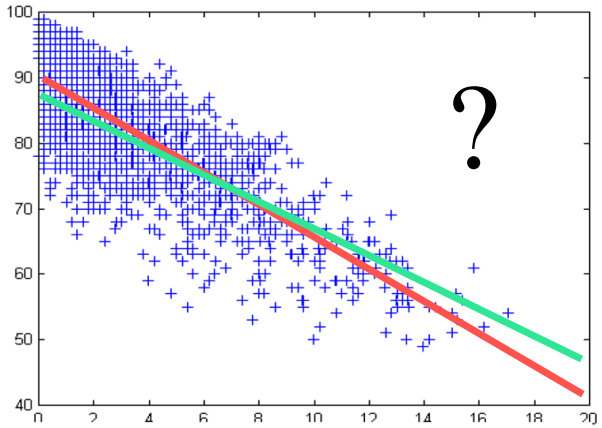


2. Modélisation du monde

2.1 Types de modèles : Comment résumer, représenter, modéliser ?



2.1 Types de modèles : Modèles linéaires



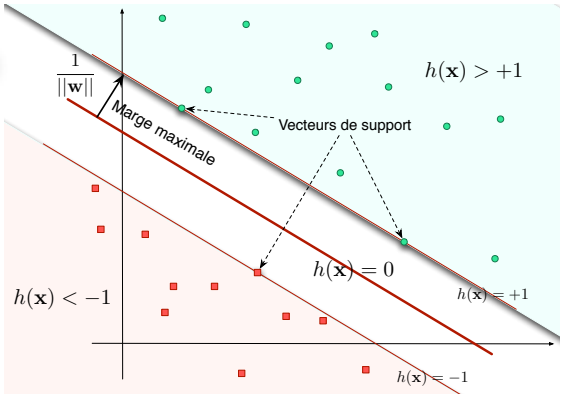
$$h(\mathbf{x}) = \sum_{i=1}^{n} w_i \ \phi_i(\mathbf{x})$$

2.1 Types de modèles : Modèles linéaires

SVM :

Séparateurs à Vastes Marges

(Support Vector Machines)

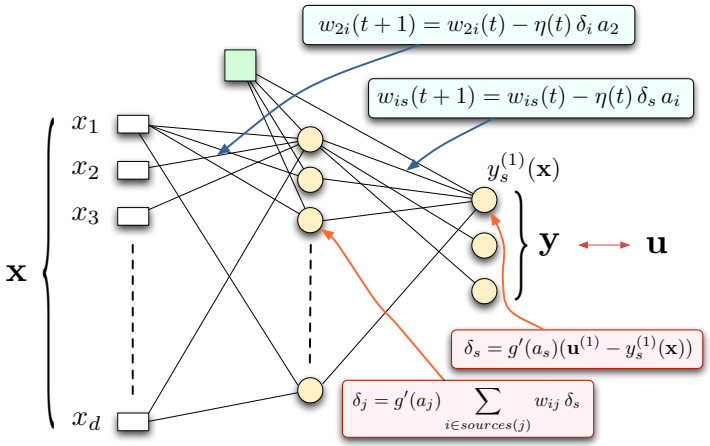


$$h(\mathbf{x}) = \sum_{i=1}^{n} w_i \, \phi_i(\mathbf{x})$$

$$h(\mathbf{x}) = \text{signe}\left\{ \sum_{i=1}^{m} \alpha_i^\star \, u_i \, . \, \langle \mathbf{x}_i \, , \, \mathbf{x} \rangle \; + \; w_0^\star \right\}$$

2.1 Types de modèles : Modèles non linéaires

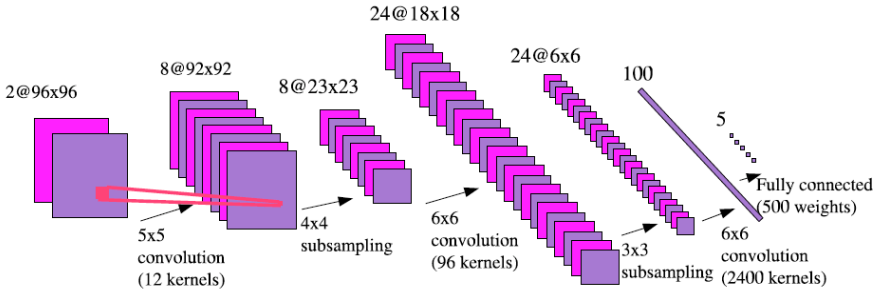
Réseaux connexionnistes



2.1 Types de modèles : Modèles non linéaires

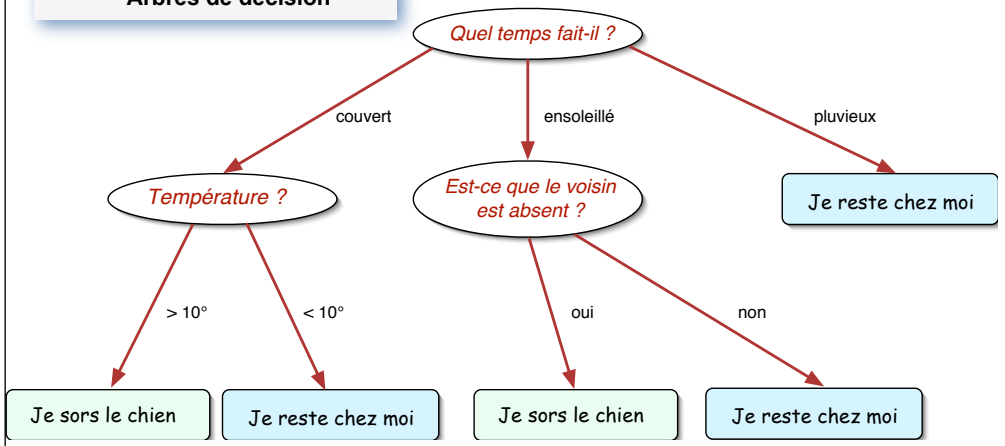
Réseaux connexionnistes

« profonds »

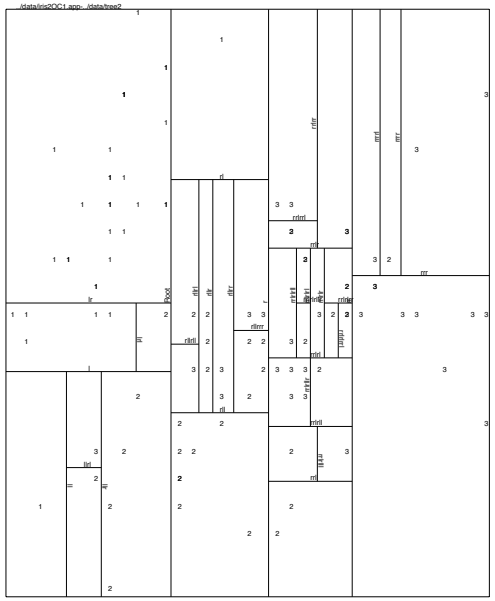


2.1 Types de modèles : Modèles constructifs

Arbres de décision

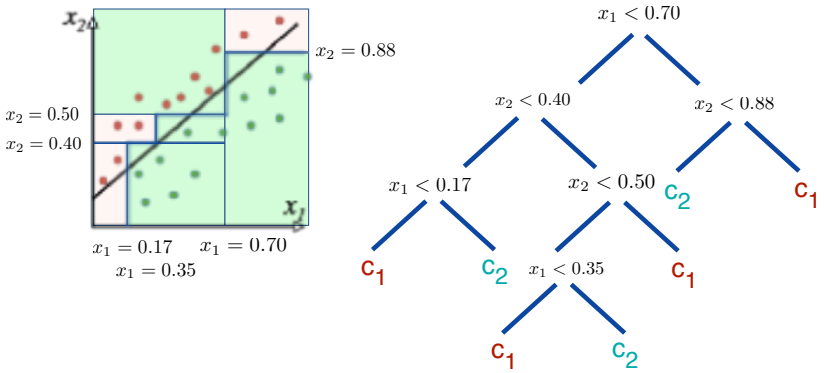


2.1 Types de modèles : Modèles constructifs



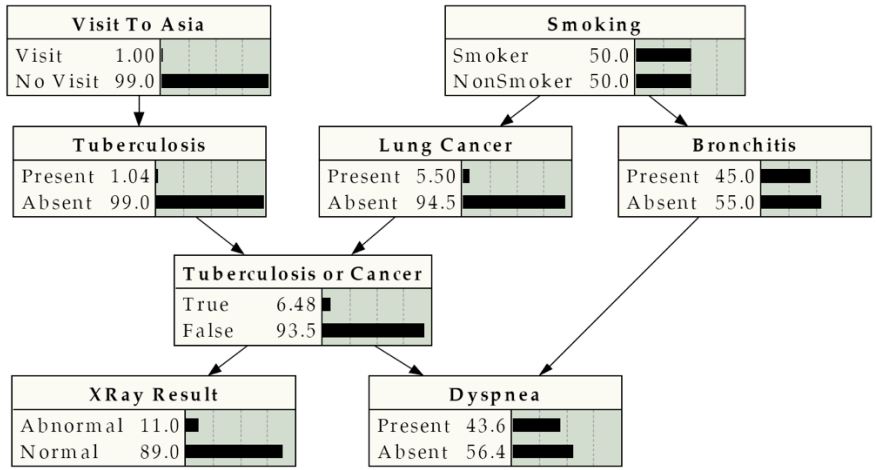
2.1 Types de modèles : Modèles constructifs

Arbres de décision

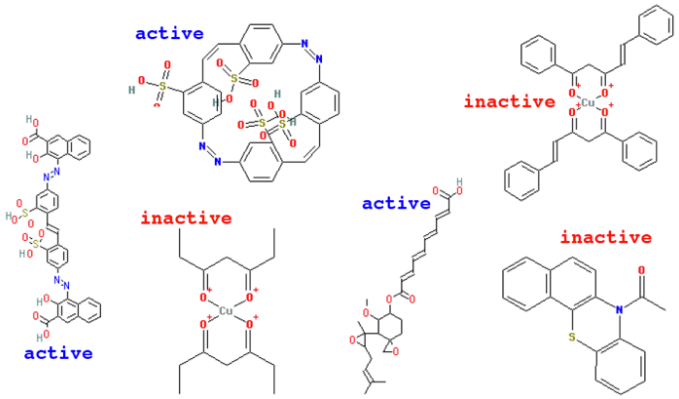


2.1 Types de modèles : Modèles constructifs

Modèles graphiques



2.1 Types de modèles : Modèles constructifs



NCI AIDS screen results (from http://cactus.nci.nih.gov).

2.1 Types de modèles : Modèles constructifs

φ_1 : $anm(x_3, [195, 22, 3, 27, 38, 40, 92]) \wedge \neg chrg(x_3, [-0.2, 0.2]) \wedge$
$anm(x_4, [195, 22, 3, 38, 40, 29, 92]) \wedge \neg type(x_4, [O]) \wedge \neg chrg(x_4, [-0.2]) \wedge$
$(x_1 < x_2) \wedge (x_1 < x_3) \wedge (x_1 < x_4) \wedge (x_2 < x_3) \wedge (x_2 < x_4) \wedge (x_3 < x_4) \wedge$
$bound(x_3, x_4) \rightarrow mutagenic,$

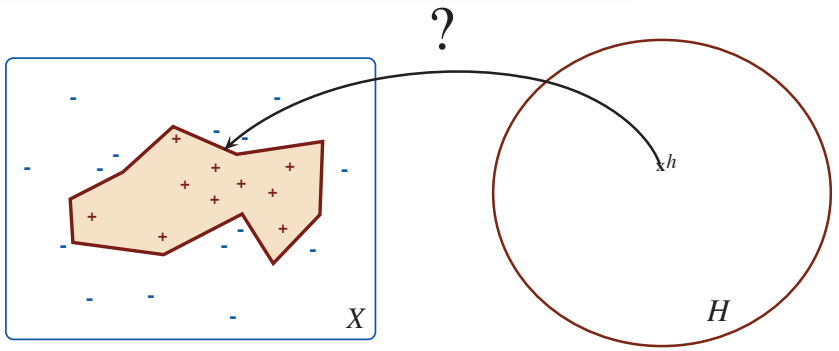
φ_2 : $\neg chrg(x_1, [-0.2]) \wedge \neg type(x_2, [N]) \wedge \neg anm(x_3, [22]) \wedge \neg chrg(x_3, [-0.6, -0.4]) \wedge$
$\neg type(x_4, [H, N, O]) \wedge (x_1 < x_2) \wedge (x_1 < x_3) \wedge (x_1 < x_4) \wedge (x_2 < x_3) \wedge$
$bound(x_2, x_3) \wedge (x_2 < x_4) \wedge (x_3 < x_4) \wedge bound(x_3, x_4) \rightarrow mutagenic,$

φ_3 : $anm(x_1, [195, 38, 29, 92]) \wedge chrg(x_1, [-0.8 \div 0.6]) \wedge \neg type(x_3, [C]) \wedge \neg chrg(x_3, [0.0]) \wedge$
$anm(x_4, [195, 22, 3, 27, 38, 29, 92]) \wedge \neg type(x_4, [N]) \wedge (x_1 < x_2) \wedge (x_1 < x_3) \wedge$
$(x_1 < x_4) \wedge (x_2 < x_3) \wedge (x_2 < x_4) \wedge (x_3 < x_4) \rightarrow mutagenic,$

φ_4 : $anm(x\ 1, [195, 3, 27, 38, 40, 29, 92]) \wedge \neg type(x_1, [H]) \wedge \neg chrg(x_1, [-0.2])$
$\neg anm(x_3, [40]) \wedge anm(x_4, [195, 22, 27, 38, 40, 29, 92]) \wedge \neg type(x_4, [H, N])$
$(x_1 < x_2) \wedge \neg bound(x_1, x2) \wedge (x_1 < x_3) \wedge (x_1 < x_4) \wedge (x_2 < x_3) \wedge (x_2 < x_4) \wedge$
$bound(x_{,} x_{)} \wedge (x3 < x4) \rightarrow mutagenic.$

2.2 Espace d'hypothèses et exploration

L'apprentissage (supervisé) : un jeu entre espaces



Espace des données

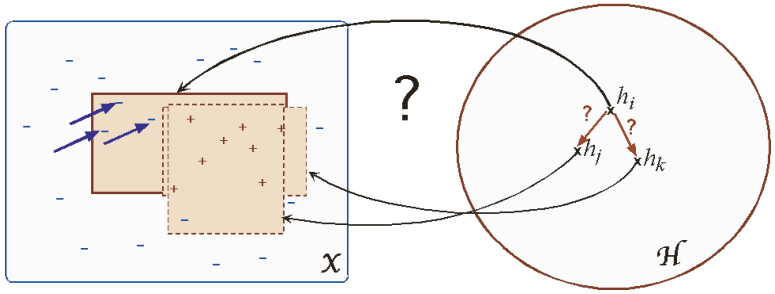
$\mathcal{X} \times \mathcal{Y}$

Espace des hypothèses

$\mathcal{H}$ $\mathcal{L}_{\mathcal{H}}$

2.2 Espace d'hypothèses et exploration

Question : **comment explorer *H* ?**



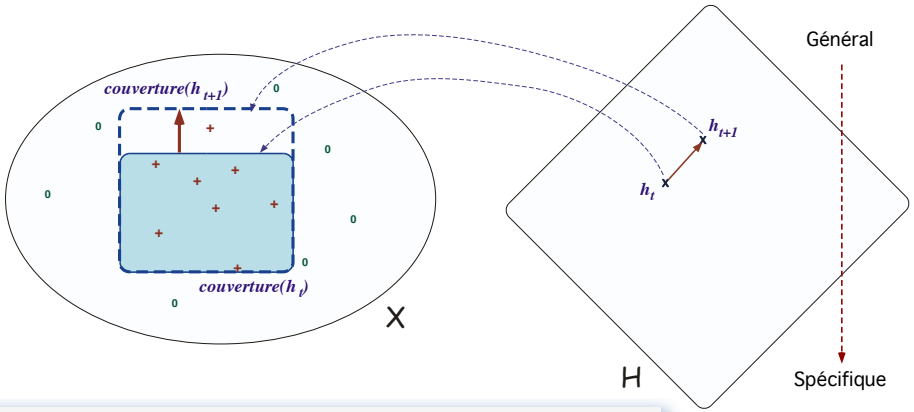
Espace des données

$\mathcal{X} \times \mathcal{Y}$

Espace des hypothèses

$\mathcal{H}$ $\mathcal{L}_{\mathcal{H}}$

2.2 Espace d'hypothèses et exploration

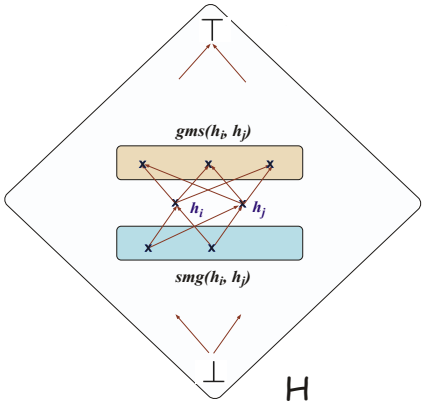


- **Opérateur de généralisation par abandon de conjonction**

$$A \wedge B \rightarrow C \overset{gen}{\Rightarrow} A \rightarrow C$$

$$Bec\ Aplati \wedge (Couleur = roux) \rightarrow \text{canard} \overset{gen}{\Rightarrow} Bec\ Aplati \rightarrow \text{canard}$$

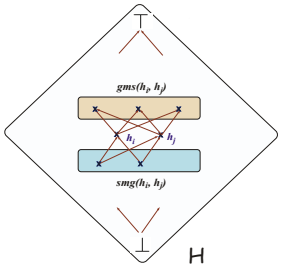
2.2 Espace d'hypothèses et exploration



Treillis de généralisation

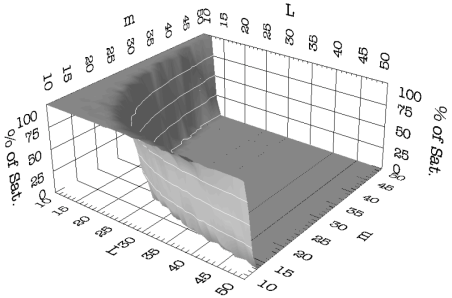
- **Représentation économique**
  des hypothèses cohérentes
- **Algorithmes efficaces**

2.2 Espace d'hypothèses et exploration



Langages de représentation expressifs

→ Phénomène de **transition de phase** !



Phase transitions in Machine Learning. Cambridge University Press, 400 p., April 2011.

2.3 Trois types de modèles

- **À base de dictionnaire de fonctions (combinaison de fonctions de base)**
  - Modèles linéaires
    - Discrimination logistique
    - Mélange de gaussiennes
  - **Modèles non linéaires**
    - Réseaux connexionnistes

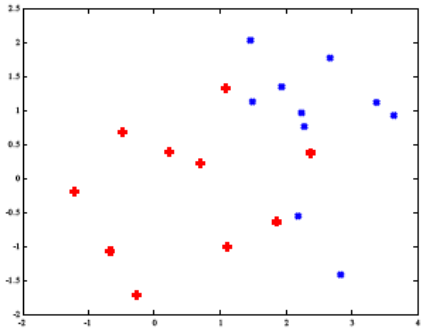
$$h(\mathbf{x}) = \sum_{i=1}^{n} w_i \, \phi_i(\mathbf{x})$$

- **Par comparaison et similarité à des exemples**
  - ***K*-ppv**
  - Méthodes à noyaux / SVM

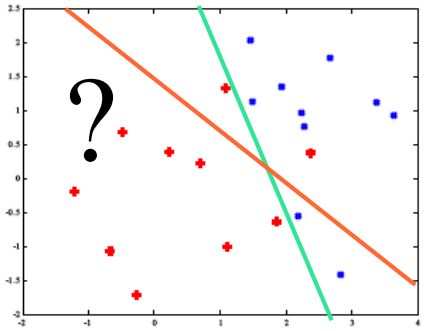
$$h(\mathbf{x}) \;=\; \text{signe}\Big\{\sum_{i=1}^{m} \alpha_i^\star \, u_i \, \kappa(\mathbf{x}, \mathbf{x}_i) \;+\; w_0^\star\Big\}$$

- **Modèles constructifs**
  - Arbres de décision
  - Modèles graphiques
  - Règles logiques
  - ...

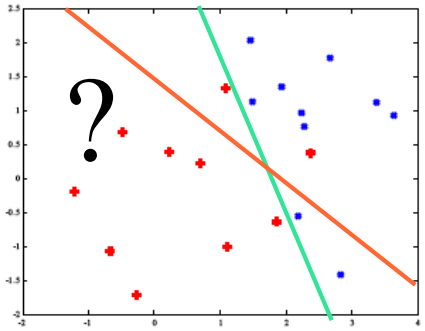
1.2 L'informatique : Modèles linéaires

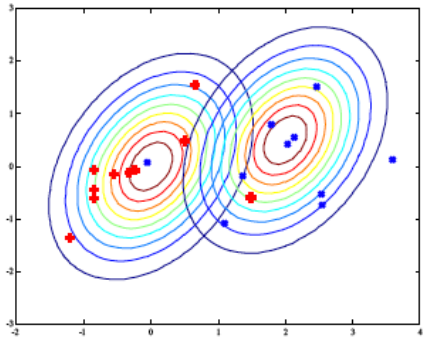


1.2 L'informatique : Modèles linéaires



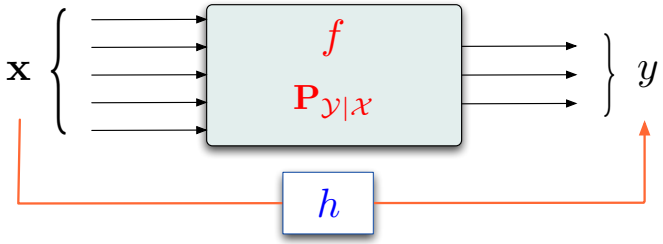
1.2 L'informatique : Modèles linéaires





3. Un problème inverse mal-posé

3.1 L'apprentissage : un problème inverse



À partir :

- d'un **échantillon d'apprentissage** $S_m = \langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \rangle$
- de **connaissances préalables** sur le type de dépendances sur $\mathcal{X} \times \mathcal{Y}$

Trouver :

- une **fonction** h
- permettant la **prédiction** de *y* pour une nouvelle entrée **x** $\quad h(\mathbf{x}) \approx y \ \big(= f(\mathbf{x})\big)$

3.1 Problème inverse mal posé

Problème **inverse :**

- À partir d'exemples $\langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\rangle$
- trouver f

Bien posé :

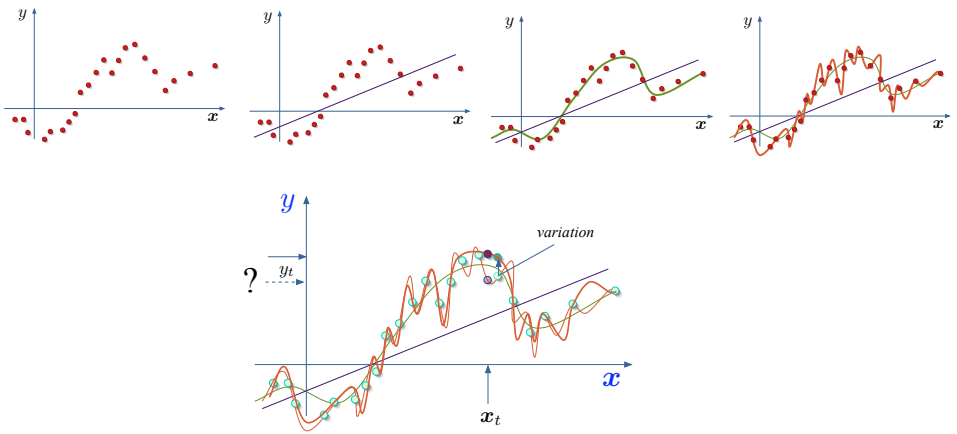
- une **solution existe**
- elle est **unique**
- elle est **stable** (dépend continûment des données)

J. S. Hadamard, 1865-1963

Le problème de l'induction est mal posé

3.1 L'apprentissage : un problème inverse

- Pas d'apprentissage par coeur, mais GÉNÉRALISATION



3.2 Quels *a priori* sur le monde ?

- Il faut contraindre l'espace des solutions

→ **Utilisation d'*a priori* sur le monde**

3.2 Quels *a priori* sur le monde ?

- **En statistique**
  - Hypothèses sur les distributions de probabilité sous-jacentes (e.g. distributions normales)

  → **Modèles *paramétriques***

- **En apprentissage artificiel**
  - **Classes d'hypothèses** (biais de langage de représentation)
  - Critère de **simplicité**
    - E.g. Peu de règles et règles simples
    - Parcimonie / économie de la solution

  → **Modèles *non paramétriques* mais exprimés dans un code**

3.2 Quels *a priori* sur le monde ?

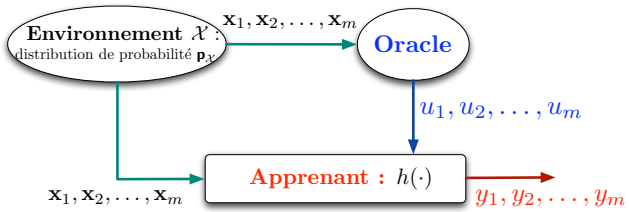
Induction = Problème inverse mal posé

=> **Il faut donc un *a priori* sur l'espace des solutions**

Problème de la sélection de modèle

3.3 Le paradigme classique

- **Objectif**
  - Formuler un modèle du monde **intelligible** et **performant**
  - Être **performant en prédiction à l'avenir**



$$L(h) = \begin{cases} \mathbf{P}_{\mathcal{X}\mathcal{Y}}\{h(\mathbf{x}) \neq y\} & (0\_1 \ \ loss) \\ \mathbb{E}[\ell(h(\mathbf{x}), y)] = \displaystyle\int_{\mathbf{x}\in\mathcal{X}, y\in\mathcal{Y}} \ell(h(\mathbf{x}), y)\ \mathbf{P}_{\mathcal{X}\mathcal{Y}}\ d(\mathbf{x}, y) \end{cases}$$

3.3 Le paradigme classique

$$L(h) = \begin{cases} \mathbf{P}_{\mathcal{XY}}\{h(\mathbf{x}) \neq y\} & (0\_1\ loss) \\ \mathbb{E}[\ell(h(\mathbf{x}), y)] = \int_{\mathbf{x}\in\mathcal{X}, y\in\mathcal{Y}} \ell(h(\mathbf{x}), y)\ \mathbf{P}_{\mathcal{XY}}\ d(\mathbf{x}, y) \end{cases}$$

- Monde supposé **stationnaire**
- Décrit par **distribution de probabilité** sur l'espace joint $\mathbf{P}_{\mathcal{XY}}$ inconnue !

Comment faire le lien entre le passé (les données d'observation) et le futur ?

- Comment choisir la fonction *h* sur la base du passé pour optimiser *L*(*h*) ?

Rôle du ***critère inductif***

3.3 Le paradigme classique

$$L(h) = \begin{cases} \mathbf{P}_{\mathcal{X}\mathcal{Y}}\{h(\mathbf{x}) \neq y\} & (0\_1\ loss) \\ \mathbb{E}[\ell(h(\mathbf{x}), y)] = \displaystyle\int_{\mathbf{x}\in\mathcal{X}, y\in\mathcal{Y}} \ell(h(\mathbf{x}), y)\ \mathbf{P}_{\mathcal{X}\mathcal{Y}}\ d(\mathbf{x}, y) \end{cases}$$

$$S_m = \langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \rangle$$

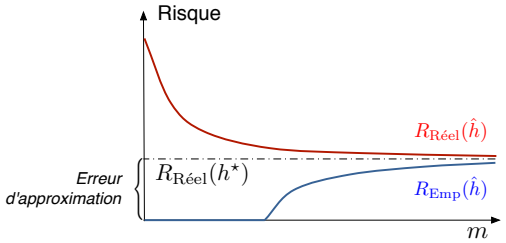
$$\hat{L}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)$$

Risque empirique

Principe de Minimisation du Risque Empirique (MRE)

$$\hat{h} = \underset{h\in\mathcal{H}}{\text{Argmin}} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i) \right\}$$

3.3 Le paradigme classique



La convergence requiert une **loi centrale limite**

Analyse statistique de l'Apprentissage

(Vapnik, et autres, 70s-80s)

Il faut contraindre l'espace des hypothèses

Principe de Minimisation du **Risque Empirique Régularisé**

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{Argmin}} \left\{ \underbrace{\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)}_{\text{Risque empirique}} + \underbrace{\lambda \cdot \Omega(h)}_{\text{Régularisation}} \right\}$$

3.3 Le paradigme classique

Régularisation

- **Sur l'espace des hypothèses** $\mathcal{H}$: $\Omega(\mathcal{H})$
  - Contrôle de la ***capacité*** de *H* (e.g. SRM)
  - Règle de l'arrêt prématuré

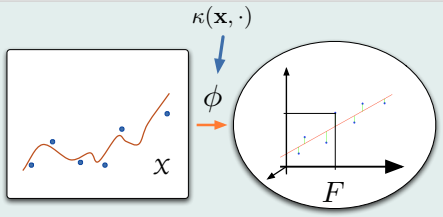
- **Sur l'hypothèse considérée** h : $\Omega(h)$
  - Norme L2, L1, ...
  - Élagage

Idée : augmenter la **stabilité**

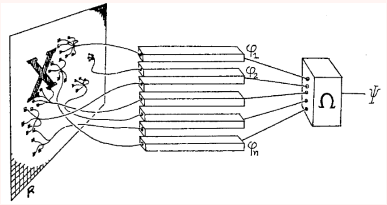
3.3 Le paradigme classique

Récemment : les méthodes à noyaux

Remise au goût du jour des
méthodes linéaires



$$h(\mathbf{x}) = \text{signe}\left\{\sum_{i=1}^{m} \alpha_i^\star \, u_i \, \kappa(\mathbf{x}, \mathbf{x}_i) + w_0^\star\right\}$$



$$h(\mathbf{x}) = \sum_{i=1}^{n} w_i \, \phi_i(\mathbf{x})$$

3.3 Le paradigme classique

Les méthodes à noyaux

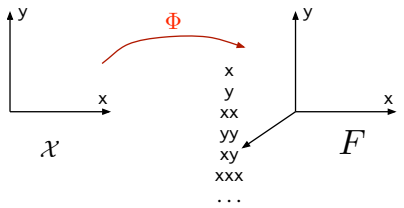
$$\kappa_s(\mathbf{x},\mathbf{x}') \;=\; (\langle \mathbf{x},\mathbf{x}'\rangle + R)^p \;=\; \sum_{s=0}^{p} \begin{pmatrix} p \\ s \end{pmatrix} R^{p-s} \langle \mathbf{x},\mathbf{x}'\rangle^s$$

Noyau polynomial

$$\begin{aligned}\kappa(\mathbf{x},\mathbf{x}') &= \big(1+(\mathbf{x}\cdot\mathbf{x}')\big)^2 \\ &= \big(1+(x_1x_1'+x_2x_2')\big)^2 \\ &= 1+2(x_1x_1'+x_2x_2')+(x_1x_1'+x_2x_2')^2 \\ &= 1+2\,x_1x_1'+2\,x_2x_2'+x_1^2{x_1'}^2+x_2^2{x_2'}^2+2\,x_1x_1'x_2x_2'\end{aligned}$$

Correspond à la projection :

$$(x_1,x_2)^\top \;\overset{\Phi}{\mapsto}\; (1,\sqrt{2}\,x_1,\;\sqrt{2}\,x_2,\;x_1^2,\;x_2^2,\;\sqrt{2}\,x_1x_2)^\top$$



3.3 Le paradigme classique

Nouveau biais vers la **parcimonie** **(sparseness)**

- Hypothèse plus **lisible** (intelligible)
- Plus **stable**

Mise en jeu d'une **norme L1** (*pénalise le nombre de paramètres non nuls*)

Nouveau **critère inductif** (parcimonieux)

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{Argmin}} \left\{ \underbrace{\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)}_{\text{Risque empirique}} + \underbrace{\lambda_1 \cdot \Omega_{L_2}(h)}_{\text{Régularisation}} + \underbrace{\lambda_2 \cdot \Omega_{L_1}(h)}_{\text{Parcimonie}} \right\}$$

3.3 Le paradigme classique

Nouveau biais vers la **parcimonie**
(sparseness)

Mise en jeu d'une **norme L1** (*pénalise le nombre de paramètres non nuls*)

- **Shrinkage**
- **LASSO**
- **LARS (Régression)**
- **...**

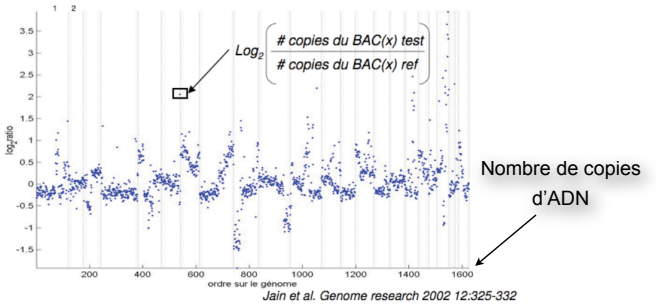
3.3 Le paradigme classique

- Des techniques puissantes
- Bien fondées mathématiquement
- Permettant d'exprimer des connaissances préalables (?)

3.4 Exemples d'application

Comparative Genomic Hybridization (CGH)

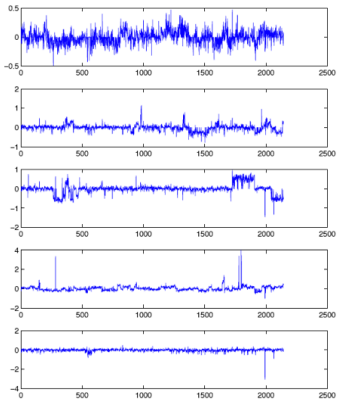
- **Motivations**
  - Les données CGH mesurent le nombre de copies d'ADN le long du génome
  - Très utile, en particulier dans la recherche sur le cancer
  - Question : peut-on classer les puces CGH pour du diagnostic ou du prognostic ?



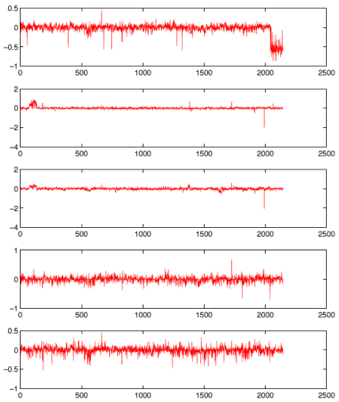
Jain et al. Genome research 2002 12:325-332

3.4 Exemples d'application

Comparative Genomic Hybridization (CGH)



Mélanome **aggressif**



Mélanome **non-aggressif**

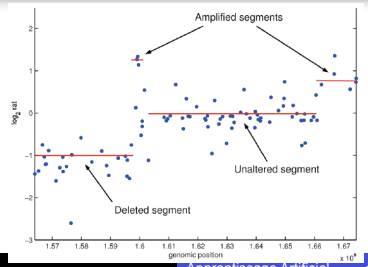
3.4 Exemples d'application

Classification de puces CGH

- **Connaissances préalables**
  - Pour un profil CGH $\mathbf{x} \in \mathbb{R}^p$ on se concentre sur les classifieurs linéaires, cad le signe de :

$$h_\beta(\mathbf{x}) = \beta^\top \cdot \mathbf{x}$$

  - On s'attend à ce que β soit :
    - **clairsemé** : toutes les positions ne devraient pas être discriminatives
    - **Constante par morceaux** : dans une région, toutes les sondes devraient contribuer de manière égale



3.4 Exemples d'application

Classification de puces CGH

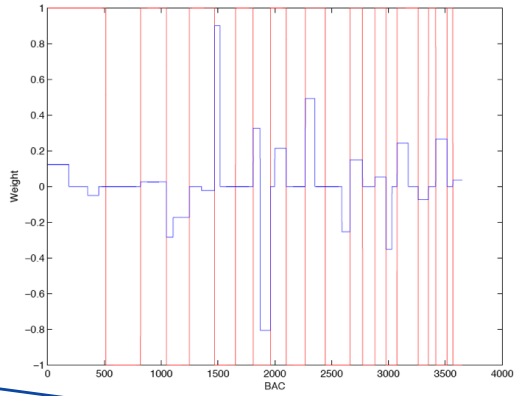
- **Connaissances préalables**
  - Introduire une pénalité favorisant les morceaux constants longs (on pénalise les sauts)
  - « Fused LASSO penalty» [Tibshirani et al., 2005]

$$\hat{h} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \left\{ \underbrace{L(h)}_{\text{Risque empirique}} + \underbrace{\lambda \cdot \sum_{i}^{p-1} |\beta_{i+1} - \beta_i|}_{\text{Régularisation}} \right\}$$

3.4 Exemples d'application

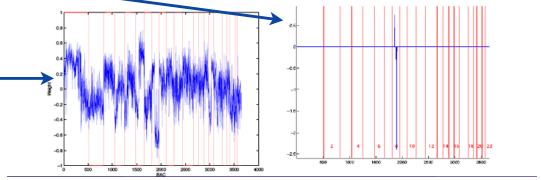
Classification de puces CGH

LASSO pénalisé →



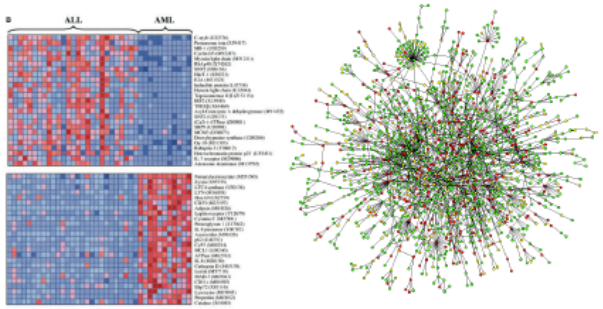
LASSO non pénalisé →

SVM sans pénalité →



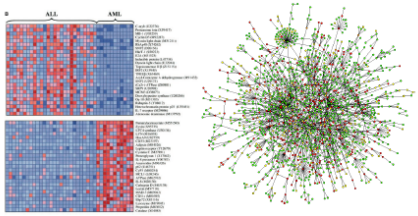
3.4 Exemples d'application

Souvent, on connaît (partiellement) des réseaux de régulations entre gènes : peut-on en tenir compte ?



3.4 Exemples d'application

- **Connaissances préalables**
  - Les fonctions biologiques impliquent généralement l'action coordonnée de plusieurs protéines (réseaux de régulation, voies métaboliques)
  - De nombreuses voies sont déjà connues
  - Hypothèse : les poids du modèle linéaire devraient être compatibles avec cette connaissance.



3.4 Exemples d'application

Prior hypothesis

Genes near each other on the graph should have similar weigths.

An idea (Rapaport et al., 2007)

$$\Omega_{spectral}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 ,$$

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2 .$$

- Gene selection + Piecewise constant on the graph

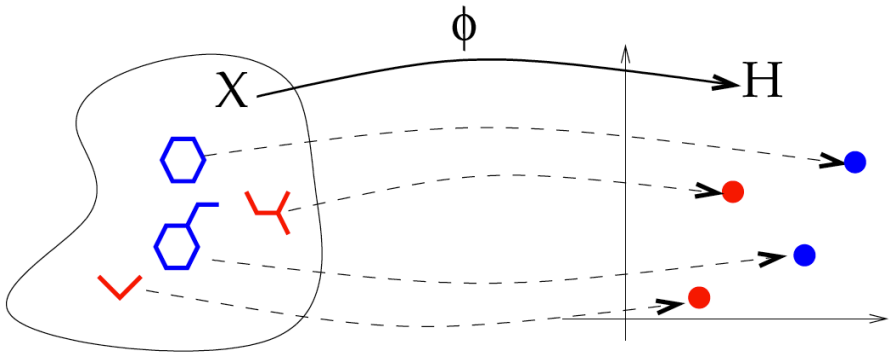
$$\Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^{p} |\beta_i|$$

- Gene selection + smooth on the graph

$$\Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^{p} |\beta_i|$$

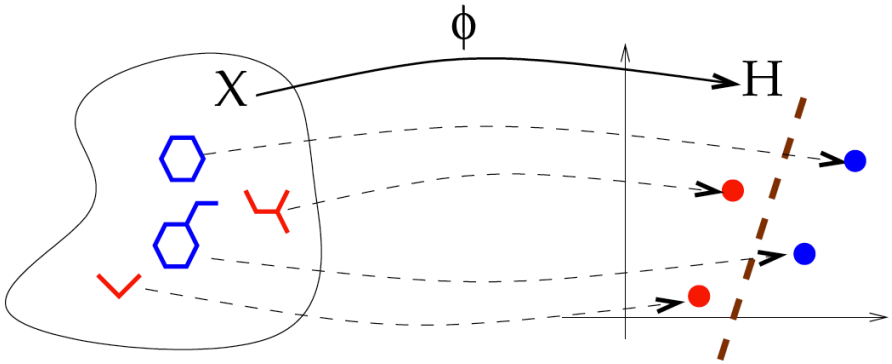
3.4 Exemples d'application

Description de molécules



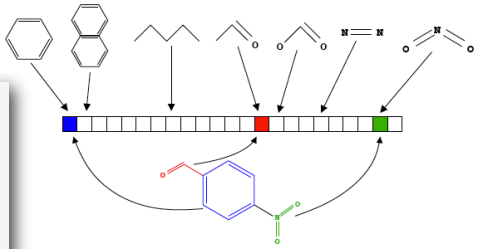
3.4 Exemples d'application

Description de molécules



3.4 Exemples d'application

Description de molécules



- **Du non vectoriel au vectoriel**
  - Les sous-structures jouent sans doute un rôle
  - Dictionnaire de sous-structures
  - Vecteur booléen sur ce dictionnaire

Problème : **comment identifier ces sous-structures ?**

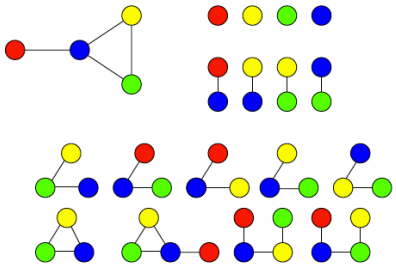
Problème 2 : **Attention, détecter la présence de ces sous-structures peut être très coûteuse**

3.4 Exemples d'application

Description de molécules

Definition

A subgraph of a graph (V, E) is a connected graph (V', E') with $V' \subset V$ and $E' \subset E$.



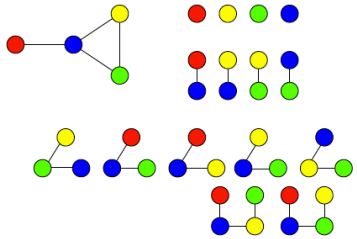
Attention, calculer toutes les occurrences de sous-graphes est un **problème NP-difficile !**

3.4 Exemples d'application

Description de molécules

Definition

- A path of a graph (V, E) is sequence of distinct vertices $v_1, \ldots, v_n \in V$ $(i \neq j \implies v_i \neq v_j)$ such that $(v_i, v_{i+1}) \in E$ for $i = 1, \ldots, n-1$.
- Equivalently the paths are the linear subgraphs.



Attention, calculer toutes les occurrences de chemins est un **problème NP-difficile !**

3.4 Exemples d'application

Description de molécules

Substructure selection

We can imagine more limited sets of substuctures that lead to more computationnally efficient indexing (non-exhaustive list)

- substructures selected by domain knowledge (MDL fingerprint)
- all path up to length k (Openeye fingerprint, Nicholls 2005)
- all shortest paths (Borgwardt and Kriegel, 2005)
- all subgraphs up to k vertices (graphlet kernel, Sherashidze et al., 2009)
- all frequent subgraphs in the database (Helma et al., 2004)

Permet de réduire / contrôler la complexité de la recherche et des calculs

3.4 Exemples d'application

Definition: Complete graph kernels

A graph kernel is complete if it separates non-isomorphic graphs, i.e.:

$$\forall G_1, G_2 \in \mathcal{X}, \quad d_K(G_1, G_2) = 0 \implies G_1 \simeq G_2 .$$

Equivalently, $\Phi(G_1) \neq \Phi(G_1)$ if G_1 and G_2 are not isomorphic.

- **Si** un noyau de graphe n'est **pas complet**, il est impossible de représenter toutes les fonctions sur $\mathcal{X}$: le noyau n'est **pas assez expressif**
- Mais on veut des **calculs réalisables**
- **Peut-on trouver des noyaux assez expressifs, et effectivement calculables ?**

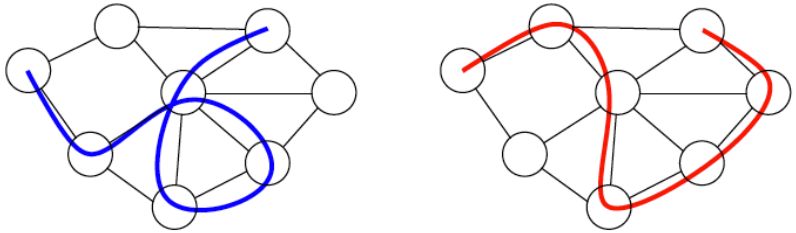
3.4 Exemples d'application

Théorème : Calculer n'importe quel noyau de graphe complet est au moins aussi difficile que le problème de trouver les sous-graphes [Gärtner et al., 2003]

Développements scientifiques

Noyau de marches aléatoires

3.4 Exemples d'application



3.4 Exemples d'application

- The nth-order walk kernel is the walk kernel with $\lambda_G(w) = 1$ if the length of w is n, 0 otherwise. It compares two graphs through their common walks of length n.
- The random walk kernel is obtained with $\lambda_G(w) = P_G(w)$, where P_G is a Markov random walk on G. In that case we have:

$$K(G_1, G_2) = P(label(W_1) = label(W_2)),$$

  where W_1 and W_2 are two independant random walks on G_1 and G_2, respectively (Kashima et al., 2003).
- The geometric walk kernel is obtained (when it converges) with $\lambda_G(w) = \beta^{length(w)}$, for $\beta > 0$. In that case the feature space is of infinite dimension (Gärtner et al., 2003).

3.4 Exemples d'application

- For the nth-order walk kernel we have $\lambda_{G_1 \times G_2}(w) = 1$ if the length of w is n, 0 otherwise.
- Therefore:

$$K_{nth-order}(G_1, G_2) = \sum_{w \in \mathcal{W}_n(G_1 \times G_2)} 1 .$$

- Let A be the adjacency matrix of $G_1 \times G_2$. Then we get:

$$K_{nth-order}(G_1, G_2) = \sum_{i,j} [A^n]_{i,j} = \mathbf{1}^\top A^n \mathbf{1} .$$

- Computation in $O(n|G_1||G_2|d_1 d_2)$, where d_i is the maximum degree of G_i.

4. Défis et avenir

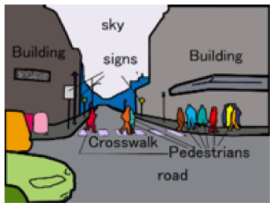
4.1 Le paradigme actuel

- Centré sur le « batch learning »
- Monde stationnaire
- Données i.i.d. (indépendantes et identiquement distribuées)

4.2 De nouveaux besoins

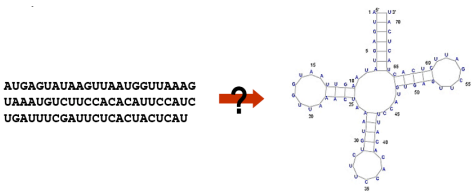
- Applications en vision
  - Très haute dimensionalité
  - Dépendances particulières (spatiales et temporelles)
  - **Flux de données** et **Temps réel**

Bileschi, Wolf, Serre, Poggio, 2007



4.2 De nouveaux besoins

- Des **espaces d'entrée et de sortie** plus complexes
  - De données **vectorielles** ...
  - ... à des données **non vectorielles**
    - *Séquences (génomiques, ...)*
    - *Textes*
  - Sorties **structurées**
    - *Arbre de dérivation*
    - *Graphe d'interaction*
  - Sorties multi-valuées
    - *Apprentissage multi-objectifs*



4.2 De nouveaux besoins

- Des **fonctions** plus complexes (e.g. multi-objectif, tri, recommandations, ...)
  - Nouvelles mesures de **similarité**
  - Nouvelles **mesures de coût**
  - **Nouveaux algorithmes** d'exploration de l'espace des hypothèses
- Des **masses de données** gigantesques et **en flux**
  - Nouveaux **algorithmes** de **complexité en *O*(*1*)**
  - Savoir quoi **oublier** et comment

4.2 De nouveaux besoins

- Des **données non i.i.d.**
  - Nouveaux critères inductifs
  - Savoir quoi **oublier** et comment
  - **tenir compte des informations de séquence** ou de spatialité
    (ou tout autre type de dépendance)

$$\hat{h} = \underset{h \in \mathcal{H}}{\text{Argmin}} \Big\{ \underbrace{\frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), y_i)}_{\text{Risque empirique}} + \underbrace{\lambda \cdot \Omega(h)}_{\text{Régularisation}} \Big\}$$

$$\begin{aligned} \hat{h}_t &= r(h_{t-1}, \mathbf{x}_t, y_t) \\ &= \text{Argmin}_{h \in \mathcal{H}} \Big\{ \frac{1}{m} \sum_{\tau=1}^{\tau=t} \ell(h_\tau(\mathbf{x}_\tau, y_\tau) \\ &\quad + \lambda \cdot \Omega(\mathcal{H}) \\ &\quad + \mu \cdot \Omega_2(r) \Big\} \end{aligned}$$

- **Apprentissage à longue vie** :
  multi-tâches, transfert, apprentissage en-ligne, apprentissage incrémental, dérive de concept

4.2 De nouveaux besoins

- Apprentissage en **interaction** (avec le monde ou entre systèmes)
  - Apprentissage par renforcement
- Apprentissages **multi-stratégies**
  - **Combinaisons** de méthodes d'apprentissage
  - Vers l'**intelligence ambiante**

4.3 Conclusions

- L'étude de **l'induction** est centrale
- Accent sur la **représentation des connaissances** et les règles d'inférences (le **raisonnement**)
- Souci de **calculabilité effective**
- Vers un **renouvellement paradigmatique** important ?
  - Théorie de l'information ?
  - Trajectoires de systèmes dynamiques (hors d'équilibre) ?
  - ...

Remerciements

- Certains transparents doivent **beaucoup** à :
  - **Jean-Philippe Vert**

Merci !!