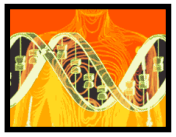
MASTER IBM 2010

Introduction à la fouille de données BioMédicales

Cours MASTER IBM 2010



Jean-Daniel Zucker

DR à l'IRD UR UMMISCO
(Modèlisation Mathématiques Et Informatiques des Systèmes Complexes)
UMR U755 INSERM/Paris 6 et LIM&BIO/Paris 13

UPMC PARIS UNIVERSITAS — IRD Institut de recherche pour le développement — Inserm Institut national de la santé et de la recherche médicale — Lim&Bio

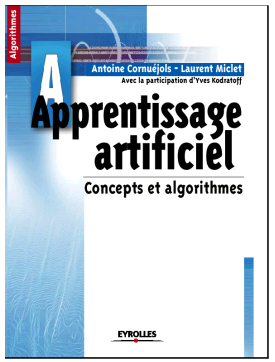
Module Fouille de Données : didactique

- **Objectifs:**
  - Comprendre le domaine de la Fouille de Données
  - Le rôle des analyses prédictives dans l'informatique médicale et la bioinfo.
  - Comprendre les algorithmes de base de la fouille de données biomédicales
  - Utiliser un environnement de Fouille de données : Clementine (et R)
- **Examen: projet personnel + soutenance**
- **Lecture: articles clefs à lire.**

Administratif: Module IF-FD Master IBM

- **Vendredi 13 Novembre 2009 –** Intro Générale / Analyse Prédictive
  - **La fouille de données**
  - **Les données biomédicales**
  - **Algorithmes pour la classification supervisée : Arbre de décision, SVM, Réseaux de Neurones, etc.**
  - **Evaluation**
  - **Exemples**
- **Lundi 16 Novembre –** fouille de données (fin)
  - **Le clustering**
  - **Les modèles graphiques (réseaux Bayésiens)**
  - **Les graphes et les réseaux d'interactions**

- **Crédits**
  - **Antoine Cornuejols**



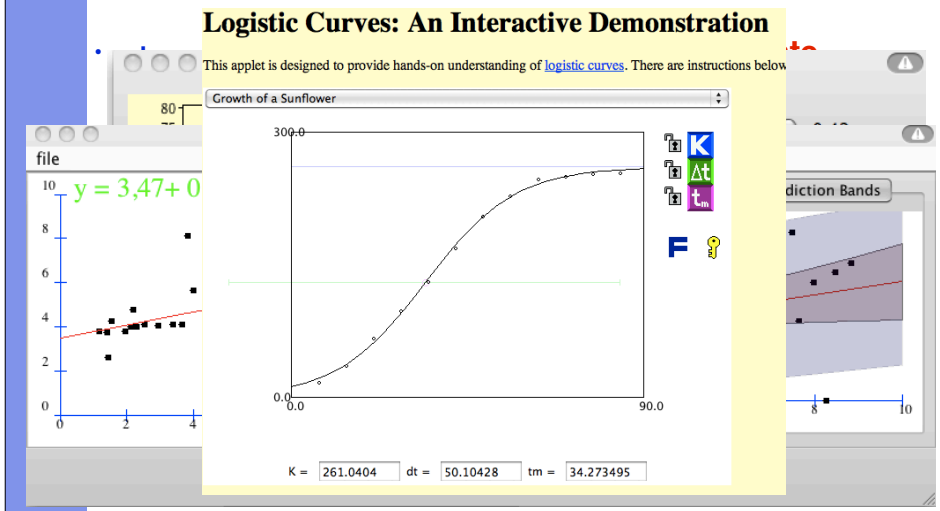
PLAN

TÂCHES DE LA FOUILLE DE DONNÉES

- ***Définition:*** *"L'exploration et l'analyse de grandes quantité de données afin de découvrir des formes et des règles significatives en utilisant des moyens automatique ou semi-automatique."*
- **Classification (valeurs discrètes):** réponse **qualitative** à un médicament, classification de demandeurs de crédits, détermination des numéros de fax, dépistage de demandes d'assurances frauduleuses, etc.
- **L'estimation (valeurs continues):** réponse **quantitative** à un médicament, du nombre d'enfants d'une famille, revenu total par ménage, probabilité de réponse à une demande, etc.
- **La prédiction (pour vérifier il faut attendre):** **durée** de vie d'un patient, des clients qui vont disparaître, des abonnés qui vont prendre un service, etc..
- **Regroupement par similitudes:** des patients qui ont telles mutations génétiques développent tel type d'obésité, etc.

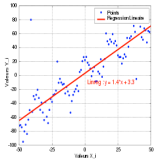
LIEN AVEC LES ANALYSES STATISTIQUES CONNUES ?

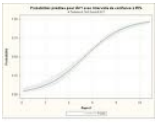
- ***Oui !***

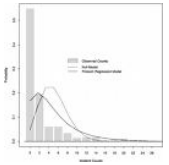


EXEMPLE 2 : COMPARAISON, ENTRE LES COMMUNAUTÉS « RICHE » ET « PAUVRE ». RÉGRESSION

- Tension artérielle moyennes :
  - Régression
- Proportion d'adultes hypertendus :
  - Régression LOGISTIQUE
- Nombre d'œufs de parasites dans les selles
  - Régression de POISSON







Exemple 1 (suite) : Expression des résultats. Régression

- Tension artérielle moyennes : Régression LINEAIRE : **la tension artérielle systolique des pauvres des environ 30% plus élevée que celle des riches***
- Proportion d'adultes hypertendus : Régression LOGISTIQUE : **la proportion d'hypertendu est 1,5 plus grande chez les pauvres que chez les riches**
- Nombre d'Œufs de parasites dans les selles : Régression de POISSON : **Le nombre d'œufs de parasites dans les selles est en moyenne 12 fois plus grande chez les riches que chez les pauvres**

* Toute choses étant « égales par ailleurs »

Dans la fouille: aspect «prédictif»

Repose sur l'induction: Proposer des lois générales à partir de l'observation de cas particuliers

Problème

Quel est le nombre *a* qui prolonge la séquence :

1 2 3 5 … *a* ?

…

- **Solution(s).** Quelques réponses valides :
  - $a = 6$. Argument : c'est la suite des entiers sauf 4.
  - $a = 7$. Argument : c'est la suite des nombres premiers.
  - $a = 8$. Argument : c'est la suite de Fibonacci
  - $a = 2\pi$. (a peut être n'importe quel nombre réel supérieur ou égal à 5)
    Argument : la séquence présentée est la liste ordonnée des racines du polynôme :
    $$P = x^5 - (11 + a)x^4 + (41 + 11a)x^3 - (61 - 41a)x^2 + (30 + 61a)x - 30a$$
    qui est le développement de : $(x - 1) . (x - 2) . (x - 3) . (x - 5) . (x - a)$
- **Généralisation**
  Il est facile de démontrer ainsi que n'importe quel nombre est une prolongation correcte de n'importe quelle suite de nombre

Mais alors … comment faire de l'induction ?

et que peut-être une science de l'induction ?

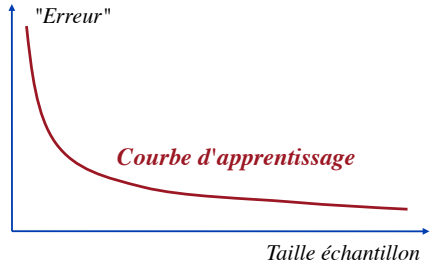
Apprendre par coeur ? IMPOSSIBLE

Représenter

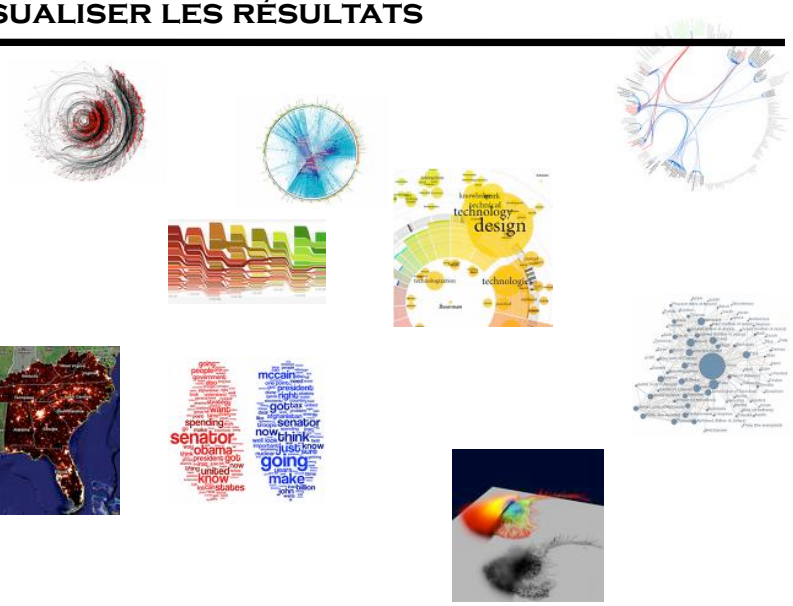
- **Extraction de caractéristiques (descripteurs, attributs)**
  - ***Eliminer*** **les descripteurs non pertinents**
  - ***Introduction*** **de nouveaux descripteurs**
    - **Utilisation de connaissances a priori**
      - **Invariance par translation**
      - **Invariance par changement d'échelle**
    - **Histogrammes**
    - **Combinaisons de descripteurs**
  - ***Ajouter*** **des descripteurs (beaucoup) !!**

Valider les résultats

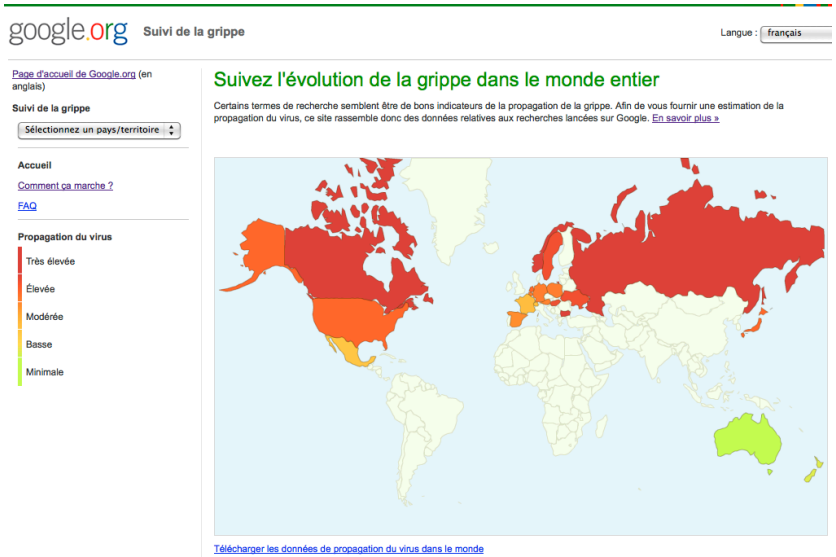
- **Quel *critère de performance* (de succès) ?**
  - **Probabilité de misclassification**
  - **Risque**
  - **Nombre d'erreurs**
- **Apprentissage sur un *échantillon d'apprentissage***
- **Test sur une *base de test***



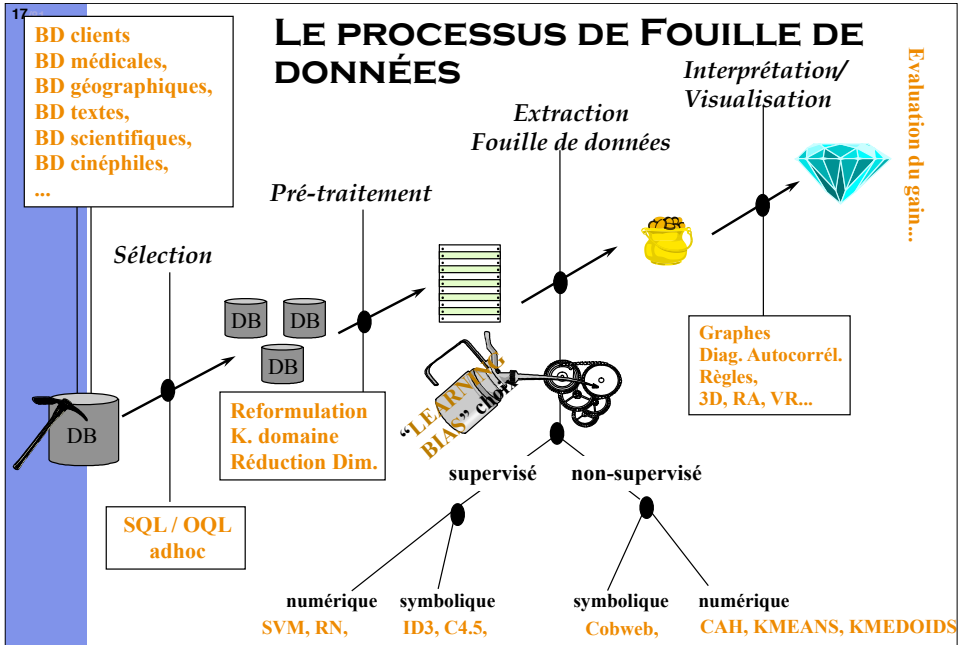
Visualiser les résultats



http://www.google.org/flutrends/

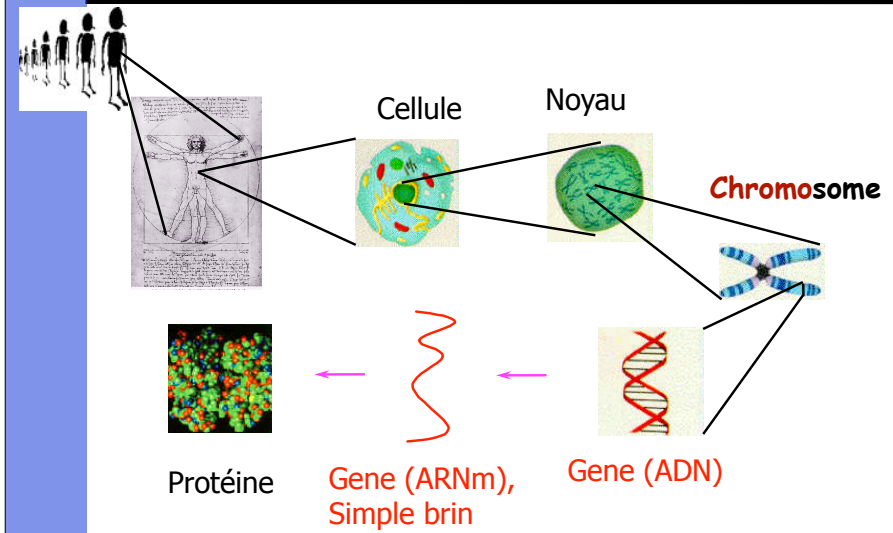


Le processus de Fouille de données



PLAN

Niveaux de l'information biologique



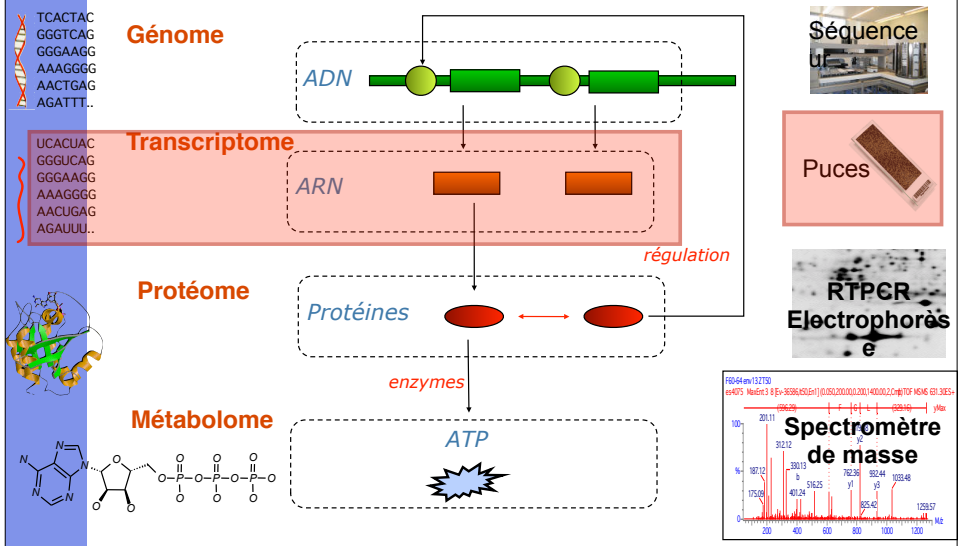
National Human Genome Research Institute

l'étymologie de **chromosome**: *chromo* = couleur et *soma* = corps: des corps colorables

Données Biomédicales

- Données écologiques
- Données épidémiologiques
- Données démographique
- Données d'analyses radiographiques
- Données d'analyse clinique
- Données d'analyse anthropomorphique
- Données d'analyse sanguines
- Données d'analyse psychologique
- Données d'analyse d'effort
- Données ...
- ....
- Données 'omiques

Les types d'information biologique : les "omes" et outils



Quelles Types d'informations: «omes» ?

Génome (l'ensemble du matériel génétique d'un individu ou d'une espèce.)

Transcriptome (l'ensemble des ARN messagers transcrits à partir du génome)

Protéome (l'ensemble des protéines exprimés à partir du génome)

Métabolome (l'ensemble des composés organiques (sucres, lipides, amino-acides, …)

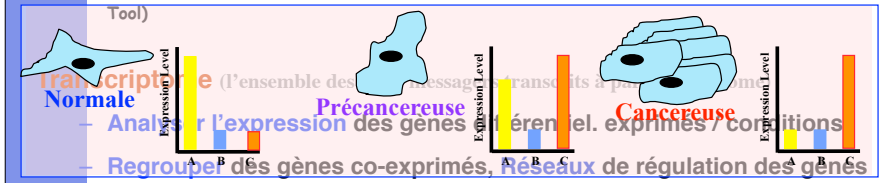
Intéractome (l'ensemble des interactions protéine-protéine)…

Bioinformatics in the Post-Genomic Era: Genome, Transcriptome, Proteome, and Information-Based Medicine

Problèmes algorithmiques et "omes" : le rôle de

Génome (l'ensemble du matériel génétique d'un individu ou d'une espèce.)

- **Identifier, prédire** les gènes dans une séquence (HMM)
- **Aligner** et comparer de **séquences** ex: BLAST (Basic Local Alignment Search Tool)



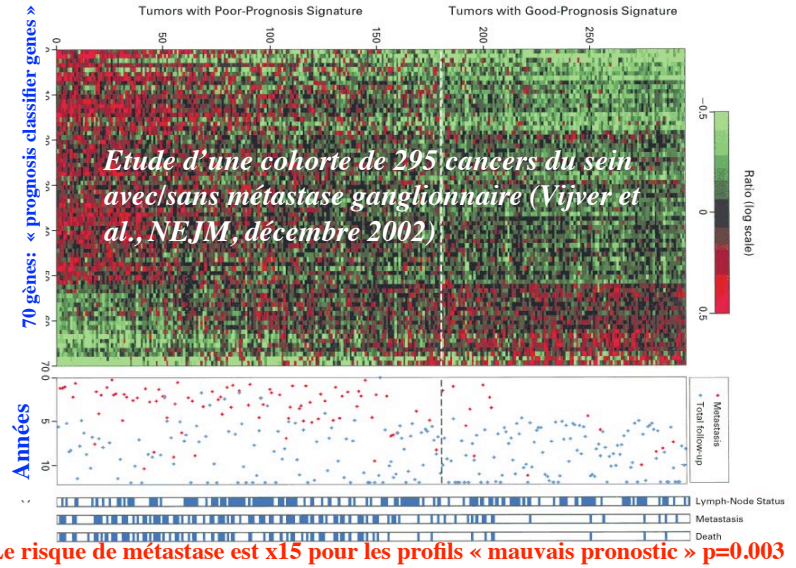
Transcriptome (l'ensemble des messagers transcrits à partir du génome)

- **Analyser l'expression** des gènes différentiel. exprimés / conditions
- **Regrouper** des gènes co-exprimés, **Réseaux** de régulation des gènes
- **Identifier** la fonction de gènes.

Protéome (l'ensemble des protéines exprimés à partir du génome)

- **Prédire** de la **structure** secondaire, la **fonction** des protéines, …
- **Analyser, mesurer l'expression** en fonction des organes

Preuve de concept



Le risque de métastase est x15 pour les profils « mauvais pronostic » p=0.003

Niveau de l'information Biologique

DNA
mRNA
Proteins
Informational Pathways
Informational Networks
Cells
Organs
Individuals
Populations
Ecologies

Traditional Biology

'omics

Genomics
Functional Genomics
Proteomics
Metabolomics
Systems Biology
Cellular Biology
Medicine
Medicine
Genetics
Ecology

Des BD et encore des BD…

- Base de Données ADN
  - GenBank, DDBJ, EMBL,…
- Base de Données Protéines
  - PIR, Swiss-Prot, PRF, GenPept, TrEMBL, PDB,…
- Base de Données EST
  - dbEST, DOTS, UniGene, Gls, STACK,…
- Base de Données Structure
  - MMDB, PDB, Swiss-3DIMAGE,…
- Base de Données voies métabol.
  - KEGG, BRITE, TRANSPATH,…
- Base de Données intégrées
  - SRS

Base de Données de Motifs
- Prosite, Pfam, BLOCKS, TransFac, PRINTS, URLs,…

Base de Données sur les maladies
- GeneCards, OMIM, OMIA,…

Base de Données taxonomique

Base de données littérature scient.
- PubMed, Medline,…

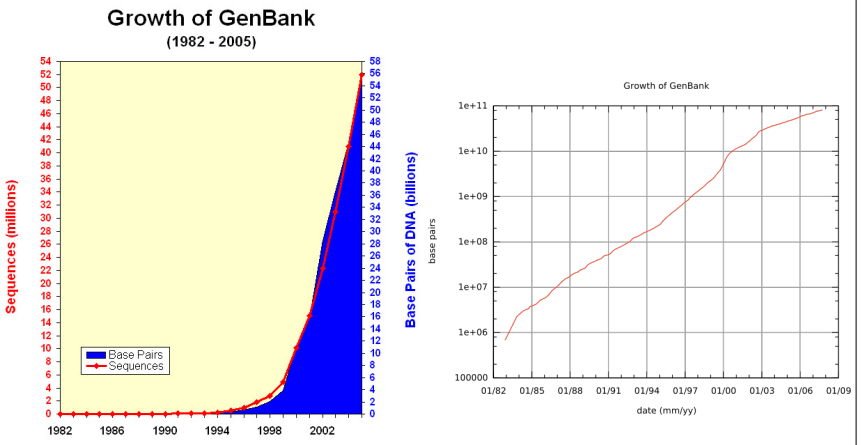
Base de données de brevets
- Apipa, CA-STN, IPN, USPTO, EPO, Beilstein,…

Autres…
- RNA databases, QTL…

Données en bioinformatique

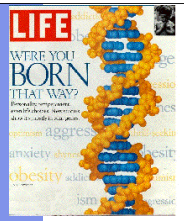
- Explosion de la quantité de données (ADN 73 Gb, arrivée des données biopuces, voies métaboliques, …)
- Croissance exponentielle des données (11-15% tous les 3 mois), plus traitable localement
- Données hétérogènes dans leur structure et leur sémantique
- Systèmes d'information hétérogènes
- Beaucoup de connaissances cachées, privées ou inconnues.
- …

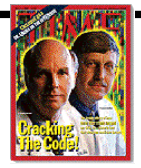
Croissance des données dans GenBank



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html
http://commons.wikimedia.org/wiki/Image:Growth_of_Genbank.svg

Les Défis



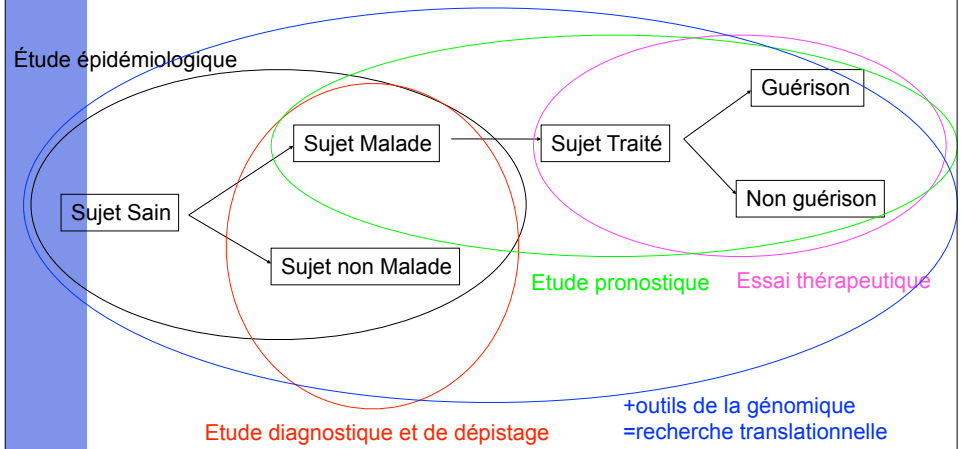



End of the beginning

Just over three years ago, it was announced that a first draft of the human genome sequence had been completed. Gaps and errors remained, but the job of fixing those problems is now largely done.

- **Décoder** l'information contenue dans les séquences d'ADN et de protéines
  - Trouver les gènes
  - Différencier entre introns et exons
  - Analyser les répétitions dans l'ADN
  - Identifier les sites des facteurs de transcription
  - Étudier l'évolution des génomes
- **Génomique Comparative**
  - Construire les relations de parenté entre organismes
- **Génomique fonctionnelle**
  - Étudier l'expression des gènes
  - Étudier la régulation des gènes
  - Déterminer les réseaux d'interaction entre les protéines
- **Génomique structurale**:
  - Modéliser les structures 3D des protéines et des ARN structurels
  - Déterminer la relation entre structure et fonction
- **Pharmacogénomique**

Une révolution pour la recherche clinique



Plan

I. La fouille de données

II. Les données biomédicales

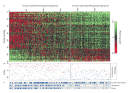
III. Algorithmes pour la classification supervisée : Arbre de décision, SVM, Réseaux de Neurones, etc.

IV. Réduction de dimmension et Evaluation

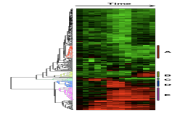
V. Exemples

Types d'apprentissages

1. Apprentissage *supervisé*
   - prédiction/régression/classification
   - apprentissage de paramètres



2. Apprentissage non-*supervisé*
   - regroupement (clustering)



3. Apprentissage *par renforcement*
   - planification dans monde inconnu
   - supervision par l'environnement

Formulation du problème d'apprentissage supervisé

Le 'Monde':

Donnees: $\{(\mathbf{x}_n, y_n)\}_{n=1}^N, \mathbf{x}_n \in \mathbf{R}^d, y_n \in \{\pm 1\}$

Fonction cible inconnue: $y = f(\mathbf{x})$ (or $y \sim P(y|\mathbf{x})$)

Distribution inconnue: $\mathbf{x} \sim p(\mathbf{x})$

But: Etant donne n nouvel $\mathbf{x}$, predire y

Probleme: $P(\mathbf{x}, y)$ est inconnu!

Formulation du problème d'apprentissage supervisé

- **Si *f* est une *fonction continue***
  - **Régression (ex: durée de vie d'un malade)**
  - **Estimation de densité**
- **Si *f* est une *fonction discrète***
  - **Classification (ex: niveau de gravité)**
- **Si *f* est une *fonction binaire* (booléenne)**
  - **Apprentissage de concept (ex: rechute oui/non plastic: oui/non)**

Formulation du problème d'apprentissage supervisé

Le 'Modèle'

Espace des Hypotheses: $\mathcal{H} = \{h \mid h : \mathbf{R}^d \to \{\pm 1\}\}$

Fonction de perte: $l(y, h(\mathbf{x}))$ (par ex. $\mathbf{I}[y \neq h(\mathbf{x})]$)

But: Minimiser la vraie (attendue) perte – "erreur en generalisation"

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ L(h) \text{ with } \boxed{L(h) := \mathbf{E}_{\mathbb{X} \times \mathbb{Y}}\, l(\mathbb{Y}, h(\mathbb{X}))}$$

Probleme: on a qu'un echantillon de donnees disponible, $P(\mathbf{x}, y)$ inconnue!

Solution: Trouver un minimiseur empirique

$$\hat{h}_N = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ \frac{1}{N} \sum_{n=1}^{N} l(y_n, h(\mathbf{x}_n))$$

Comment construire efficacement des hypotheses complexes de faible erreur en generalisation ?

Exemples de fonctions de perte

- **Discrimination**

$$l\,(h(\boldsymbol{x}_i), y_i) = \begin{cases} 0 & \text{si } y_i = h(\boldsymbol{x}_i) \\ 1 & \text{si } y_i \neq h(\boldsymbol{x}_i) \end{cases}$$

- **Régression**

$$l\,(h(\boldsymbol{x}_i), y_i) = \left[h(\boldsymbol{x}_i) - y_i\right]^2$$

- **Estimation de densité**

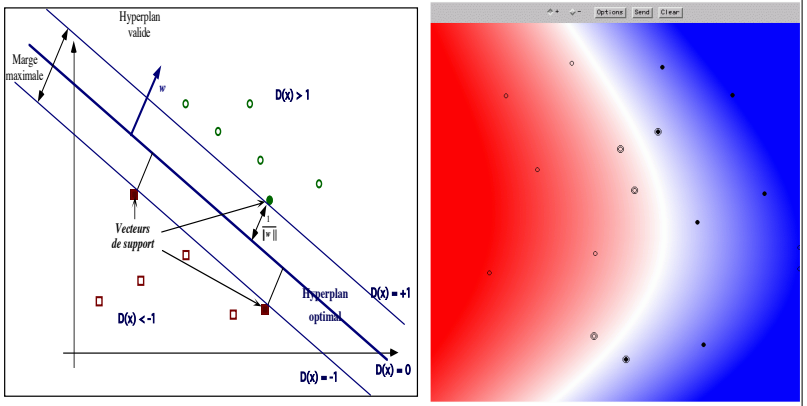
$$l\,(h(\boldsymbol{x}_i)) = -\ln h(\boldsymbol{x}_i)$$

A) Les algorithmes interpretables (regles, arbres)

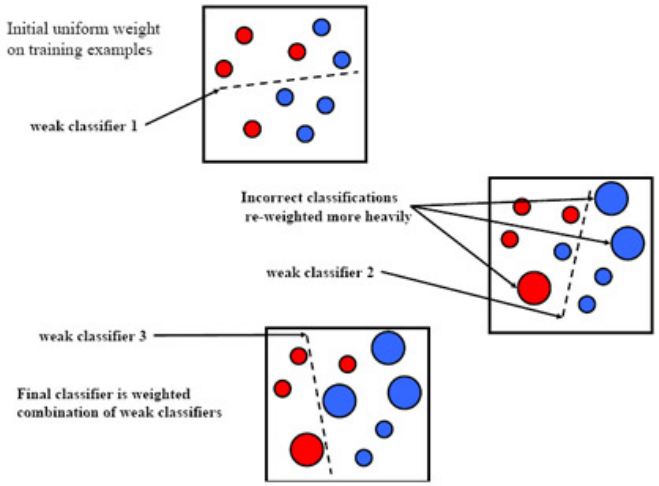
SI BMI.J1 < 40.46 (17/22)
- ET Insul.J1 < 3.745 ALORS C(+) (4/17)
- ET Insul.J1 >= 3.745 (13/17)
  - ET MG.J1 < 41.405 (8/13)
    - ET Poids.J1 < 88.5 ALORS C(-) (3/8)
    - ET Poids.J1 >= 88.5 (5/8)
      - ET Poids.J1 < 101.9 ALORS C(+) (4/5)
      - ET Poids.J1 >= 101.9 ALORS C(-) (1/5)
  - ET MG.J1 >= 41.405 ALORS C(-) (5/13)

SI BMI.J1 >= 40.46 ALORS C(+) (5/22)

B) Les algorithmes numériques (black-box)

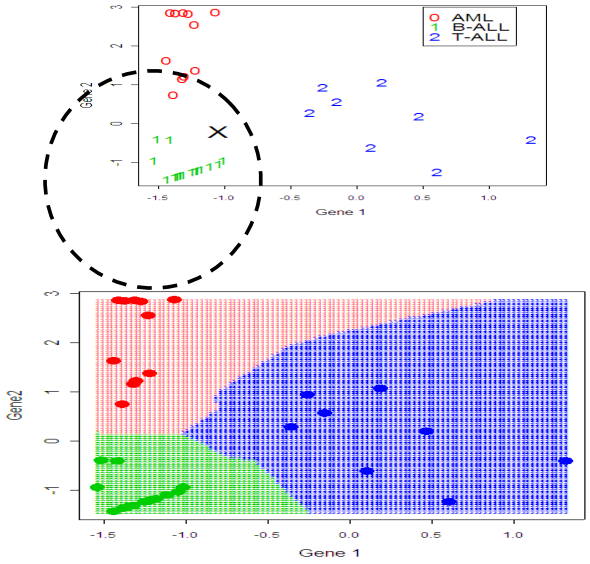


C) Les algorithmes ensemblistes



$$H(x) = sign(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x))$$

D) Les algorithmes "paresseux" k-PPV



PLAN

Algorithme de construction d'arbre de décision

- Une première approche pourrait consister à générer tous les arbres possibles et à choisir le meilleur

  ... trop couteux sauf si l'on a très peu d'attributs et très peu d'exemples.

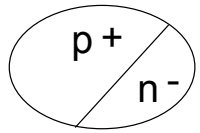
- Il faut un biais
  - exploration ascendante ou descendante
  - forme des arbres de décision
- Le problème principal: choisir à chaque étape le bon attribut sur lequel tester ...

c) Biais de la famille ID3 (Induction of Decision tree≈3)

- Approche descendante: on part d'une racine de l'arbre et on raffine.
  Famille TDIDT (Top Down Induction of Decision Trees).
  Recherche en meilleur d'abord (avec une fonction d'évaluation) sans retour arrière.
- ID3 a été conçu pour prendre en compte de nombreux attributs et de nombreux exemples
- ID3 cherche à construire des arbres relativement simple mais ne garantit pas de produire le plus simple
  (qu'est-ce que la simplicité ?)

II) L'algorithme ID3. A) Hypothèses générales

Soitp = le nombre d'exemples positifs
n = le nombre d'exemples négatifs



Dans ID3, les hypothèses de bases sont:

- (H1: exemples représentatifs)
  Un arbre de décision approprié classera des objets inconnus dans la même proportion que celle des exemples d'apprentissage

  Un objet arbitraire sera donc assigné
  à la classe P avec la probabilité de p/ (p+n) et
  à la classe N avec la probabilité de n/ (p+n)

- (H2: simplicité inhérente du monde)

  Parmi les arbres solutions, l'arbre le plus simple est préférable

A) Exemple illustratif

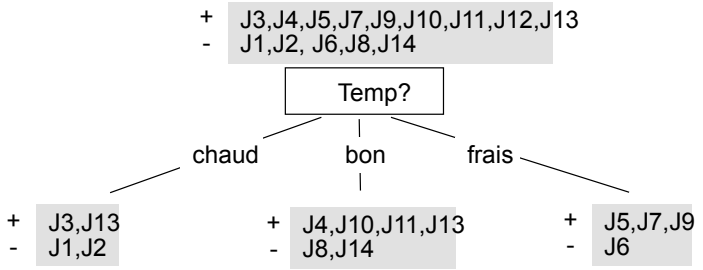
Attributs	Pif	Temp	Humid	Vent
Valeurs possibles	soleil,couvert,pluie	chaud,bon,frais	normale,haute	vrai,faux

N°	Pif	Temp	Humid	Vent	Golf
1	soleil	chaud	haute	faux	NePasJouer
2	soleil	chaud	haute	vrai	NePasJouer
3	couvert	chaud	haute	faux	Jouer
4	pluie	bon	haute	faux	Jouer
5	pluie	frais	normale	faux	Jouer
6	pluie	frais	normale	vrai	NePasJouer
7	couvert	frais	normale	vrai	Jouer
8	soleil	bon	haute	faux	NePasJouer
9	soleil	frais	normale	faux	Jouer
10	pluie	bon	normale	faux	Jouer
11	soleil	bon	normale	vrai	Jouer
12	couvert	bon	haute	vrai	Jouer
13	couvert	chaud	normale	faux	Jouer
14	pluie	bon	haute	vrai	NePasJouer

la classe

B) L'algorithme d'apprentissage d'arbres de décision (AAD)

- On choisit le premier attribut à utiliser pour l'arbre.



• Après ce choix, on se trouve fasse au problème initial sur des sous-ensembles d'exemples.

• D'où l'idée d'un algorithme récursif.

B) L'algorithme (récursif) AAD: principe

PRCEDURE *AAD*(T,N)

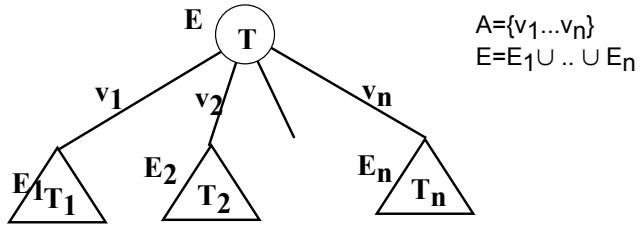
SI tous les exemples de N sont dans la même classe Ci

ALORS affecter l'étiquette Ci au noeud courant FIN

SINON *sélectionner* un attribut A avec les valeurs $v_1...v_n$

Partitionner N selon $v_1...v_n$ en N_1, ..., N_n

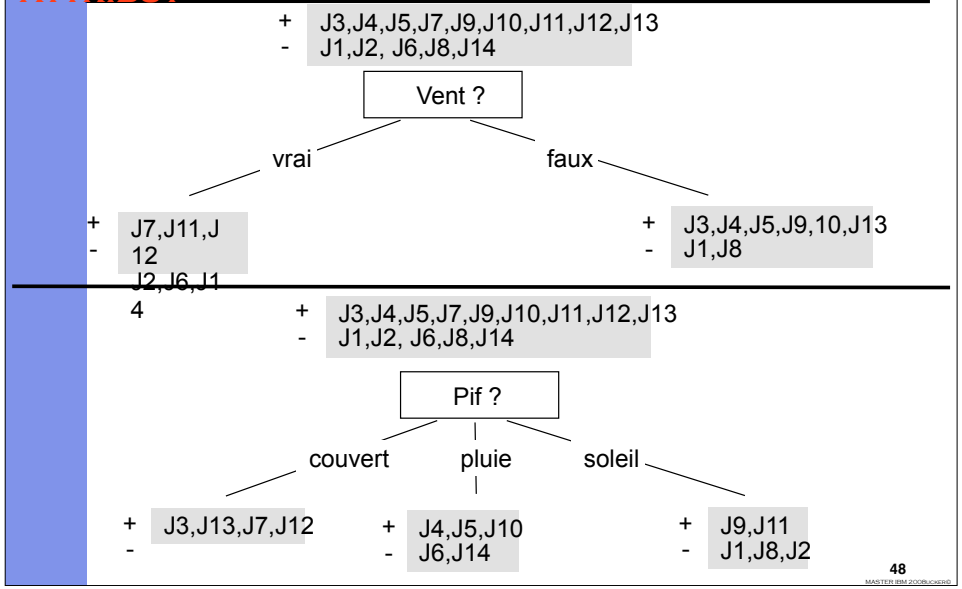
Pour j=1 à n *AAD*(T_j, N_j).



$A=\{v_1...v_n\}$

$E=E_1 \cup .. \cup E_n$

B) Sélectionner le bon attribut



B) Sur la simplicité d'un arbre de décision

- L'arbre le plus simple est celui qui permet de minimiser l'espérance du nombre de questions nécessaires à la classification d'un exemple d'apprentissage.
- Quelle fonction d'évaluation locale de l'importance d'un attribut peut correspondre à la mesure de simplicité globale ?

B) Construction d'arbres décision

- ... personne ne le sait !
- Plusieurs critères locaux ont été proposés (cf. cours suivant)
- L'entropie est une mesure du désordre règnant dans une collection d'objet. Si tous les objets appartiennent à la même classe, il n'y a pas de désordre.
- Quinlan a proposé de choisir l'attribut qui minimise le désordre de la partition résultante.

C) La mesure d'information

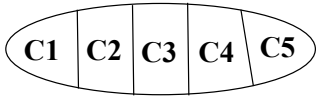
- Shannon en 1949 a proposé une mesure d'entropie valable pour les distributions discrètes de probabilité.
- Elle exprime la quantité d'information, c'est à dire le nombre de bits nécessaire pour spécifier la contribution
- L'entropie d'information est:

$$I = -\sum_{i=1..k} p_i \times \log_2(p_i)$$

où p_i est la probabilité de la classe C_i

C) L'entropie, mesure d'information

Entropie d'information de N objets:



k classes équiprobables: **$I = \lg_2(k)$**

1 seule classe: **EI=0**

- Est nulle quand il n'y a qu'une classe
- D'autant plus grande que les classes sont équiprobables
- Vaut $\log_2(k)$ quand les k classes sont équiprobables
- Unité: le bit d'information

C) L'entropie dans le cas de deux classes

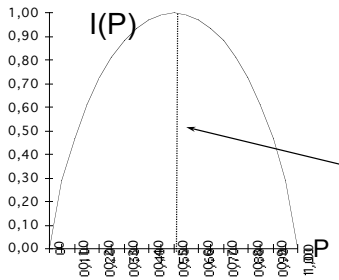
- Pour k=2 on a $I(p,n) = -p_+ \times \log_2(p_+) - p_- \times \log_2(p_-)$

  D'après l'hypothèse H1 on a $p_+ = p/(p+n)$ et $p_- = n/(p+n)$

d'où $I(p,n) = -\frac{p}{(p+n)} \log\left(\frac{p}{(p+n)}\right) - \frac{n}{(p+n)} \log\left(\frac{n}{(p+n)}\right)$

et

$I(P) = -P \log P - (1-P) \log(1-P)$



P=p/(p+n)=n/(n+p)=0.5
équiprobable

C) Exemple de calcul d'entropie d'arbre

Pour les exemples initiaux
$I(p,n) = -\ 9/14 \log_2 (9/14) - 5/14 \log_2 (5/14) = 0.940$ bits

Sous arbres de Pif?

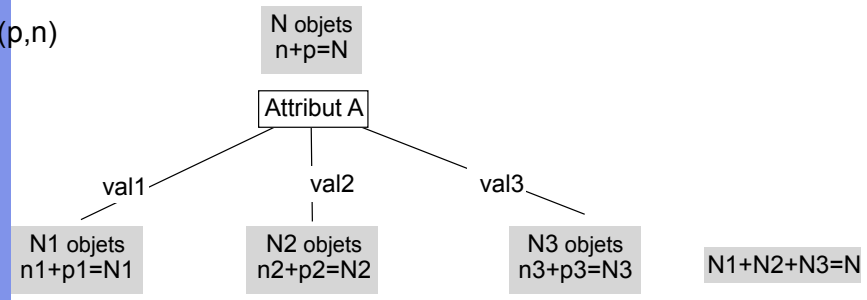
p1=4 n1=0 I(p1,n1)=0
p2=2 n2=3 I(p2,n2)=0.971
p3=3 n3=2 I(p3,n3)=0.971

Sous arbres de Temp?

p1=2 n1=2 I(p1,n1)=1
p2=4 n2=2 I(p2,n2)=0.918
p3=3 n3=1 I(p3,n3)=0.811

C) Gain d'entropie d'un arbre de décision

I(p,n)



$E(N,A) = N1/N \times I(p1,n1) + N2/N \times I(p2,n2) + N3/N \times I(p3,n3)$

Le gain d'entropie de A vaut: GAIN(A)= I(p,n)-E(N,A)

C) Exemple de calcul de gain

Pour les exemples initiaux
$I(p,n) = -\ 9/14 \log_2 (9/14) - 5/14 \log_2 (5/14) = 0.940$ bits

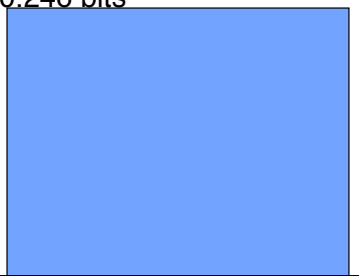
Information du sous-arbre en testant sur Pif
E(Pif) = 4/14 I(p1,n1) + 5/14 I(p2,n2) + 5/14 I(p3,n3)
= 0.694 bits

Gain(Pif)= 0.940 - 0.694= 0.246 bits

Gain(Temp)= 0.029 bits

Gain(Humid)= 0.151 bits

Gain(Vent)= 0.048 bits



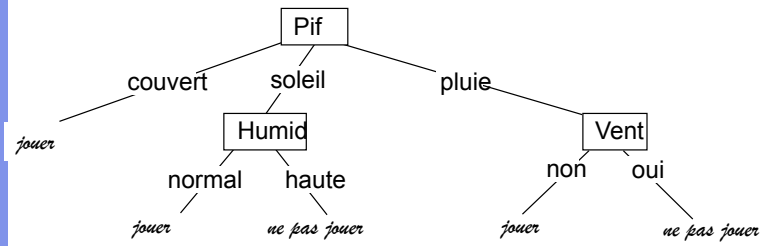
D) AAD: pseudo-code

function AAD(*exemples*, *attributs*, *défaut*) return un arbre de décision
input *exemples*, un ensemble d'exemples d'apprentissage
attributs, un ensemble d'attributs
défaut, la valeur par défaut du concept à apprendre

if *exemples* is empty then return défaut
else if (all *exemples* have same class C) then return C
else if *attributes* is empty then return VAL-MAJ(exemples)
else
best ← CHOIX-ATRIBUT(*attributes*,*exemples*)
tree ← a new decision tree with root test *best*
for each value v_i of *best* do
exemples$_i$ ← {elements of *exemples* with *best* =v_i }
subtree ← AAD(exemples$_i$, attributes-best, VAL-MAJ(exemples))
add a branch to *tree* with label v_i and subtree *subtree*
end

D) Exemple

- Arbre obtenu pour les 14 exemples du cours



III. Systèmes TDIDT

Source: vecteur d'attributs valués associés à chaque exemple
Cible: arbre de décision

- CLS (Hunt, 1966) [analyse de données]
- ID3 (Quinlan 1979)
- ACLS (Paterson & Niblett 1983)
- ASSISTANT (Bratko 1984)
- C4.5 (Quinlan 1986) puis C5 et See 5
- CART (Breiman, Friedman, Ohlson, Stone, 1984)
- CHAID,QUEST,

Plan

Les SVMs (Séparateurs à Vastes Marges)

- **Tâche de discrimination** (entre deux classes)
  - **Cas de la** séparation linéaire

- On cherche h sous forme d'une fonction linéaire : $h(\boldsymbol{x}) = w.\boldsymbol{x} + b$
- La ***surface de séparation*** est donc l'hyperplan :

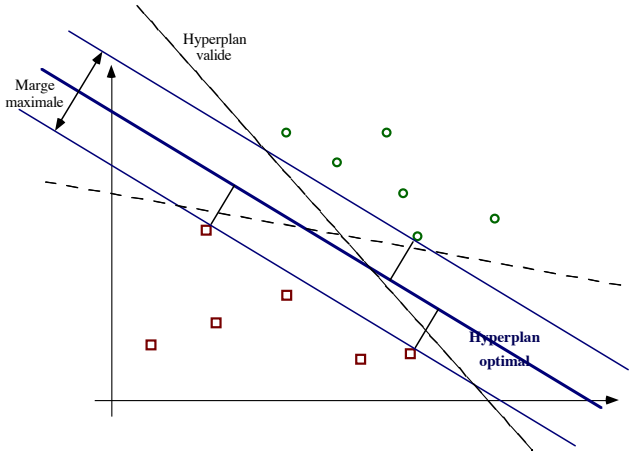
$$w.\boldsymbol{x} + b = 0$$

- Elle est valide si $\forall i \quad u_i\, h(\boldsymbol{x}_i) \geq 0$
- L'hyperplan est dit sous forme canonique lorsque $\min_i |w.\boldsymbol{x} + b| = 1$

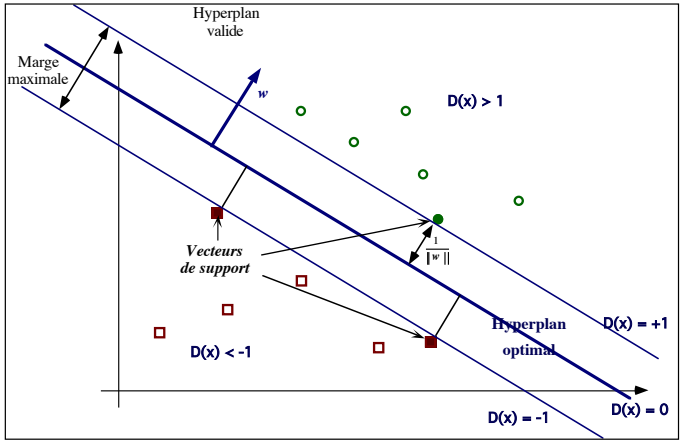
  ou encore

$$\forall i \quad u_i (w.\boldsymbol{x}_i + b) \geq 1$$

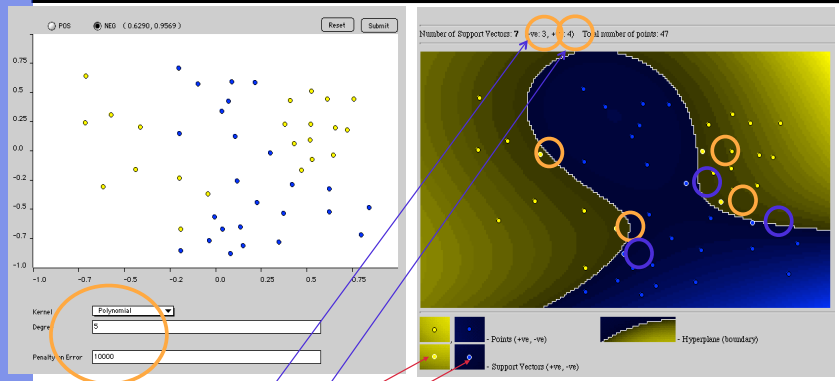
Hyperplan de plus vaste marge



Optimisation de la marge

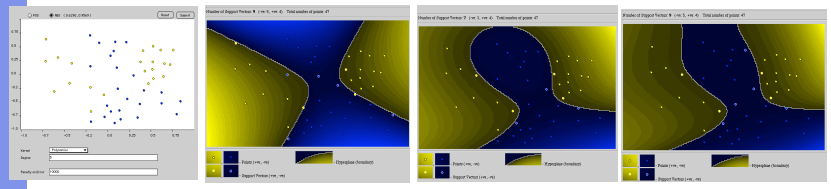


Paramètres de contrôle : les fonctions noyau



- http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml
- **47 exemples** (22 +, 25 -)
- *Exemples critiques* : 4 + et 3 -
- Ici *fonction polynomiale* de *degré 5* et *C = 10000*

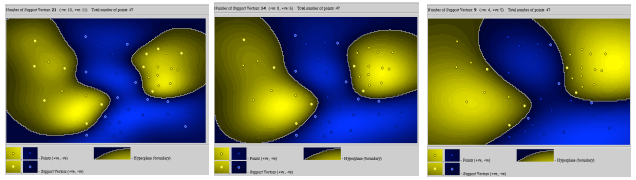
Paramètres de contrôle : les fonctions noyau



- 47 exemples (22 +, 25 -)
- *Exemples critiques* : 4 + et 3 -

(5-, 4+) (3-, 4+) (5-, 4+)

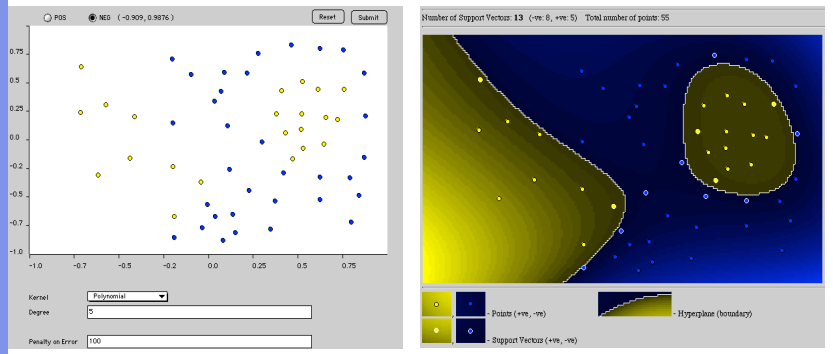
Ici *fonction polynomiale* de *degré 2, 5,* 8 et $C = 10000$



(10-, 11+) (8-, 6+) (4-, 5+)

Ici *fonction Gaussienne* de $\sigma = 2, 5,$ 10 et $C = 10000$

Ajout de quelques points ...



- `http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml`
- 47 + 8 exemples (30 +, 25 -)
- *Exemples critiques* : 5 + et 8 -
- Ici ***fonction polynomiale*** de ***degré 5*** et ***C = 10000***

Estimation de la performance

- ***Empiriquement*** : par **validation croisée**
- ***Heuristiquement*** (mais théoriquement fondé)
  - Nombre de points de supports
    - » **Moins il y en a, mieux c'est**
  - Caractéristiques de la matrice noyau
    - » **Si pas de structure dans K, aucune régularité ne peut-être trouvée**
    - » **E.g.**
      - Si les termes hors diagonale sont très petits : sur-adaptation
      - Si matrice uniforme : sous-apprentissage : tous les points sont attribués à la même classe

Construction de fonctions noyau

- Construction à partir de fonctions noyau de base (Propriétés de clôture)
  - » $K(x,z) = K_1(x,z) + K_2(x,z)$
  - » $K(x,z) = a\, K_1(x,z)$
  - » $K(x,z) = K_1(x,z)\ .\ K_2(x,z)$
  - »…
- Construction de fonctions noyau dédiées
  - **Splines B_m**
  - **Expansion de Fourrier**
  - **Ondelettes**
  - **...**

Implémentation des SVMs

- Minimisation de fonctions différentiables convexes à plusieurs variables
  - Pas d'optima locaux
  - Mais :
    - » Problèmes de stockage de la matrice noyau (si milliers d'exemples)
    - » Long dans ce cas
  - D'où mise au point de méthodes spécifiques
    - » Gradient sophistiqué
    - » Méthodes itératives, optimisation par morceaux
  - Plusieurs packages publics disponibles
    - » SVMTorch
    - » SVM^{Light}
    - » SMO
    - » …

Bilan : état des recherches

- Deux tâches évidentes
  - *Conception de noyaux*
    - » Commence à être bien étudié
    - » Encore des recherches pour certains types de données
  - *Noyautiser* les algorithmes classiques (« kernelization »)
    - » SVM
    - » Kernel Régression
    - » Kernel PCA
    - » Clustering (K-means, …)
    - » Estimation de densité, détection de nouveauté
    - » Tri (ranking)
    - » …
- Recherche sur la sélection automatique des modèles (choix des paramètres)

Plan

Prédiction de courses hippiques

Comment gagner aux courses ?

- On interroge des parieurs professionnels
- Supposons:
  - Que les professionnels ne puissent pas fournir une règle de pari simple et performante
  - Mais que face à des cas de courses, ils puissent toujours produire des règles un peu meilleures que le hasard
- Pouvons-nous devenir riche?

Idée

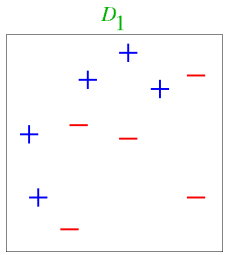
- Demander à l'expert des heuristiques
- Recueillir un ensemble de cas pour lesquels ces heuristiques échouent (cas difficiles)
- Ré-interroger l'expert pour qu'il fournisse des heuristiques pour les cas difficiles
- Et ainsi de suite…

- Combiner toutes ces heuristiques
- Un expert peut aussi bien être un algorithme d'apprentissage peu performant (*weak learner*)

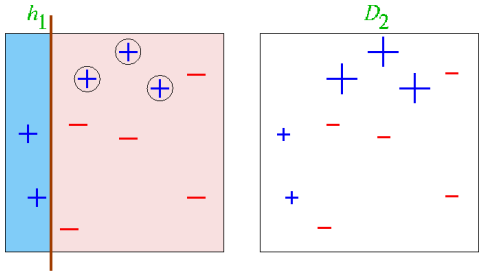
Boosting

- boosting = méthode générale pour convertir des règles de prédiction peu performantes en une règle de prédiction (très) performante
- Plus précisément :
  - Étant donné un algorithme d'apprentissage "faible" qui peut toujours retourner une hypothèse de taux d'erreur $\leq 1/2-\gamma$
  - Un algorithme de boosting peut construire (de manière prouvée) une règle de décision (hypothèse) de taux d'erreur $\leq \varepsilon$

Exemple jouet



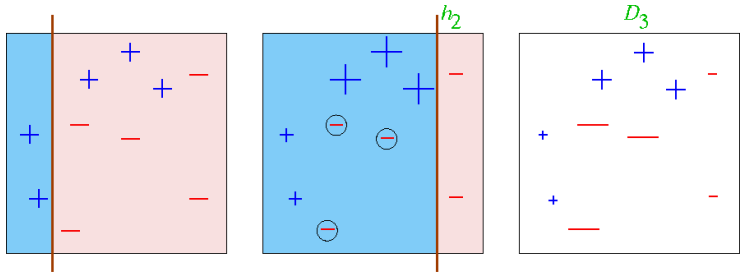
Étape 1



$\varepsilon_1 = 0.30$

$\alpha_1 = 0.42$

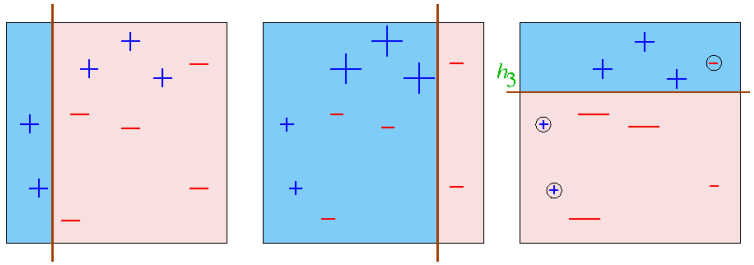
Étape 2



$\varepsilon_2 = 0.21$

$\alpha_2 = 0.65$

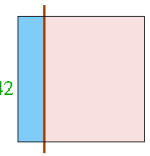
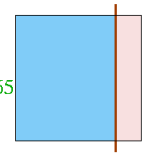
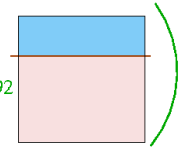
Étape 3

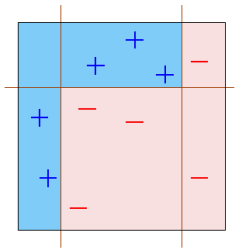


$\varepsilon_3 = 0.14$

$\alpha_3 = 0.92$

Hypothèse finale

$H_{\text{final}} = \text{sign}\left(0.42 \right.$  $+ 0.65$  $+ 0.92$  $\left. \right)$

$=$ 

Boosting : résumé

- La prédiction finale est issue d'une combinaison (vote pondéré) de plusieurs prédictions
- Méthode :
  - Itérative
  - Chaque classifieur dépend des précédents
    (les classifieurs ne sont donc pas indépendants comme dans d'autres méthodes de vote)
  - Les exemples sont pondérés différemment
  - Le poids des exemples reflète la difficulté des classifieurs précédents à les apprendre

Plan

Pourquoi la sélection d'attributs

- Facteurs sans influence ou peu influents
- Facteurs redondants
- Dimension des entrées telle que coût de l'apprentissage trop grand

- Apprentissage moins coûteux
- Faciliter l'apprentissage
  - Meilleure performance en classification
  - Meilleure compréhensibilité de l'hypothèse
- Identifier les facteurs pertinents
  - Génomique
  - Vision

La sélection d'attributs

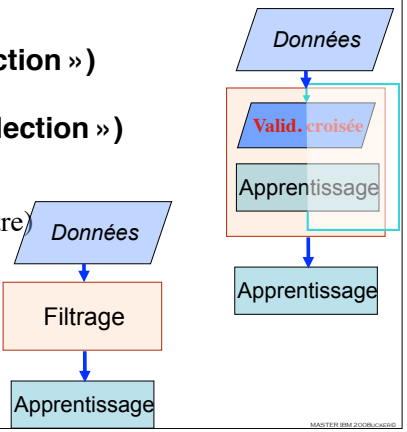
- **Idéalement**
  - Identifier le **sous-ensemble d'attributs de taille minimale nécessaire et suffisant pour définir le concept cible**
- **Classiquement**
  - Sélectionner un **sous-ensemble d'attributs de taille $n < d$**, tel qu'**un critère soit optimisé** par rapport à tous les sous-ensembles de taille n.
- **Amélioration de l'erreur en classification**
  - Apprentissage **supervisé**
- **Rester proche de la distribution originale des classes**
  - Apprentissage **non supervisé**

Pertinence d'un attribut

- **Non pertinent ou redondant**
  - **Si sa présence n'améliore pas**
    - **» L'erreur en classification (supervisé)**
    - **» La proximité à la distribution originale des classes (non supervisé)**

Les approches

1. **Approche intégrée** (« embedded »)
2. « **Wrapper methods** » (approche symbiose) [Blum & Langley, 97] [Guyon & Elisseeff, 03]
   - **Utilisent la performance en aval pour sélectionner les attributs**
   - **Deux stratégies**
     - *Ascendante* **(« forward selection »)**
       - **Par ajouts successifs d'attributs**
     - *Descendante* **(« backward selection »)**
       - **Par retraits successifs d'attributs**
3. « **Filter methods** » (approche par filtre)
   - **Indépendantes des traitements aval**

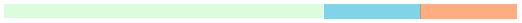


RELIEF (1)

- [Kira & Rendell,92], [Kononenko,94]
- **Les attributs les plus pertinents sont ceux qui varient plus lorsque l'exemple considéré change de classe que lorsqu'il ne change pas**
  - **Complexité faible**
  - **Grande résistance au bruit**

Ensembles de données (collections)

Toutes les données disponibles

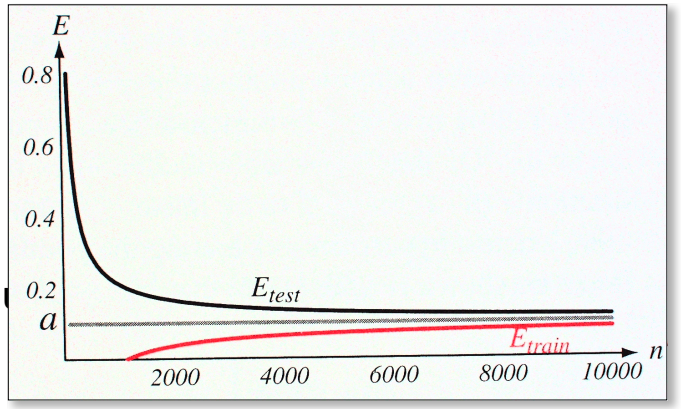


*Ensemble d'**apprentissage***

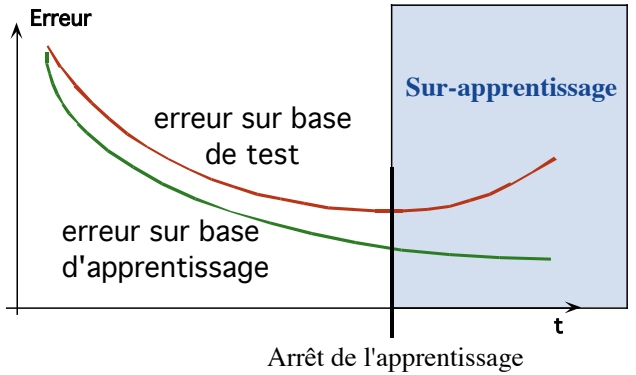
*Ensemble de **test***

*Ensemble de **validation***

Prédiction asymptotique (le cas idéal)



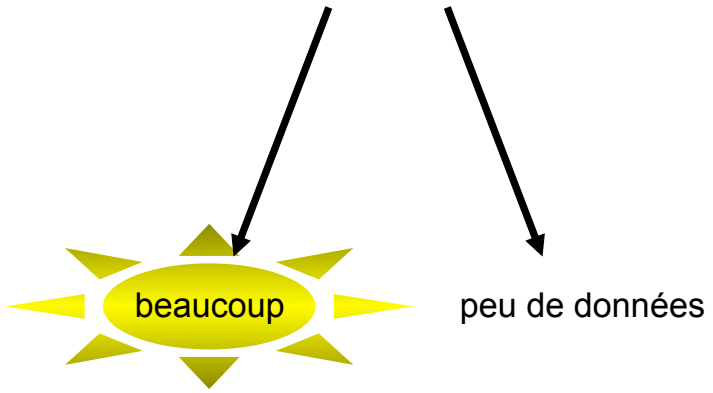
Le sur-apprentissage (*over-learning*)



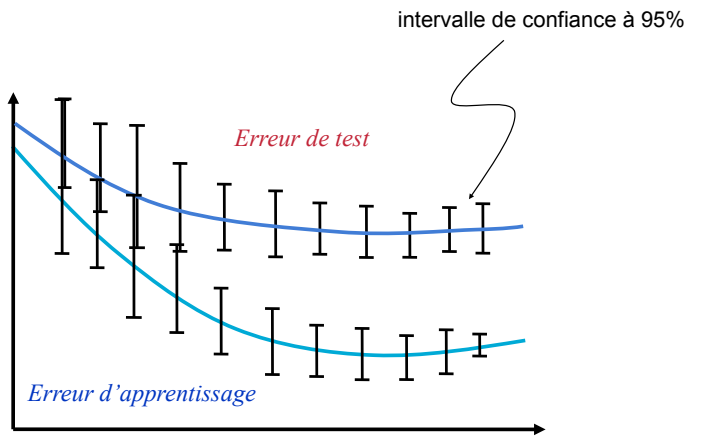
Utilisation de l'ensemble de validation

- **On règle les paramètres de l'algorithme d'apprentissage**
  - » **E.g. : nb de couches cachées, nb de neurones, ...**
  - **– en essayant de réduire l'erreur de test**
- **Pour avoir une estimation non optimiste de l'erreur, il faut recourir à une base d'exemples non encore vus : la *base de validation***

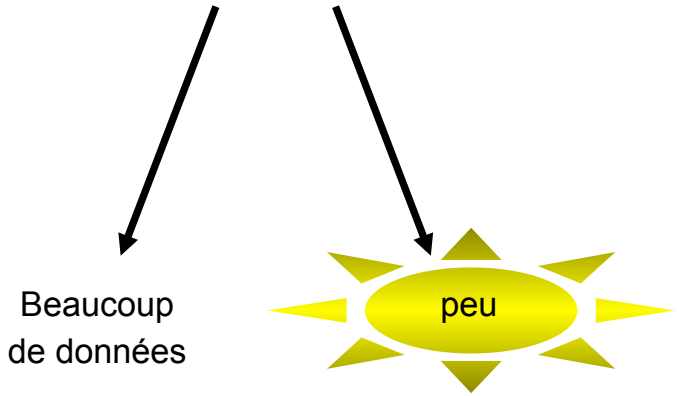
Évaluation des hypothèses produites



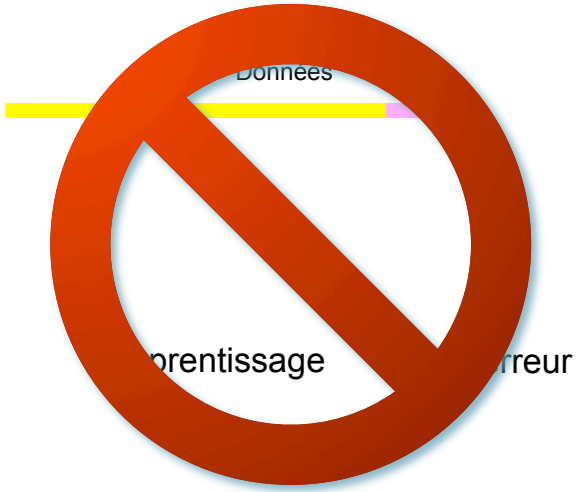
Courbes de performance



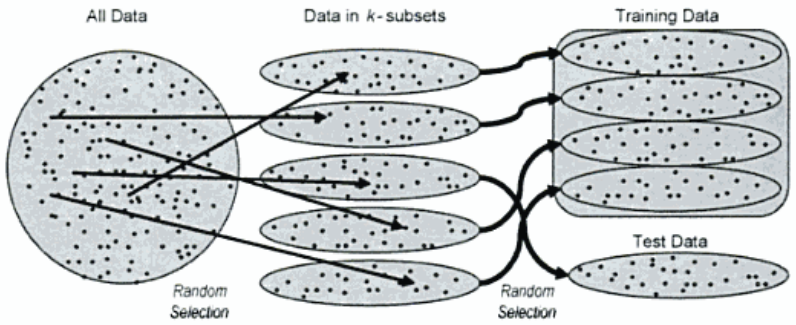
Évaluation des hypothèses produites



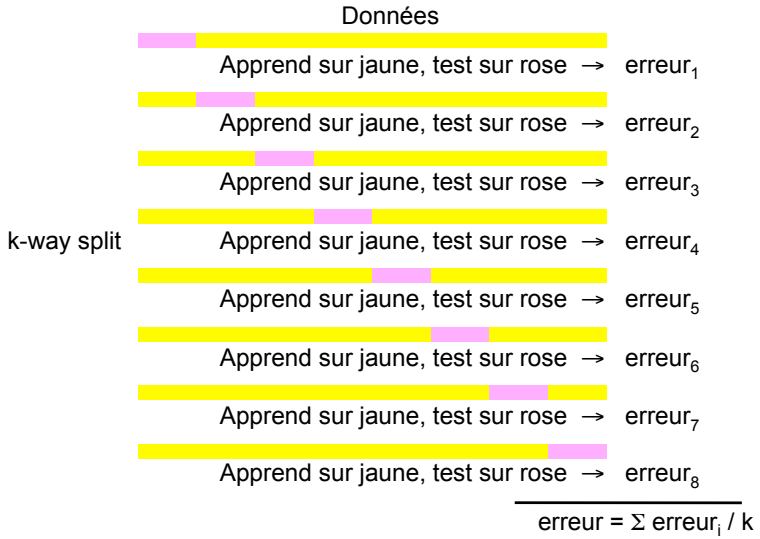
Différents ensembles



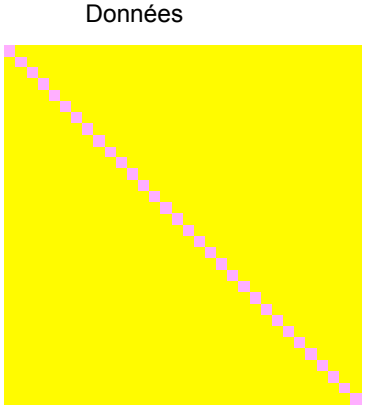
Validation croisée à k plis (k-fold)



Validation croisée à k plis (k-fold)



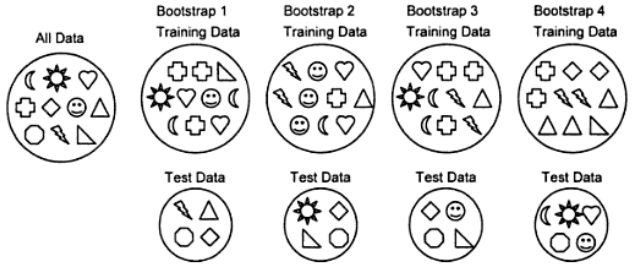
Procédure "leave-one-out"



- Faible biais
- Haute variance
- Tend à sous-estimer l'erreur si les données ne sont pas vraiment i.i.d.

[Guyon & Elisseeff, jMLR, 03]

Le Bootstrap



Le bootstrap est biaisé

Le bootstrap est biaisé (son estimation du biais est biaisée vers zéro), car certaines observations **sont utilisées à la fois dans l'échantillon pour construire le modèle et dans l'échantillon pour le valider.** Le bootstrap "hors du sac" (out-of-the-bag) et le bootstrap .632 tentent de corriger ce biais.

Le Bootstrap

Out-of-the-bag bootstrap

Le bootstrap "hors du sac" consiste à ne pas utiliser toutes les observations pour valider le modèle mais uniquement celles qui ne figurent pas déjà dans l'échantillon ayant servi à le construire (c'est d'ailleurs ce qu'on faisait pour la validation croisée).

Bootstrap .632

En fait, le bootstrap "out-of-the-bag" est quand-même biaisé, mais dans l'autre sens. Pour tenter de corriger ce biais, on peut faire une moyenne pondérée du bootstrap initial et du bootstrap oob.

```
.368 * (biais estimé par le bootstrap) +
.632 * (biais estimé par le bootstrap oob)
```

(le coefficient .632 s'interprète ainsi : pour n grand, les échantillons de bootstrap contiennent en moyenne 63,2% des observations initiales).

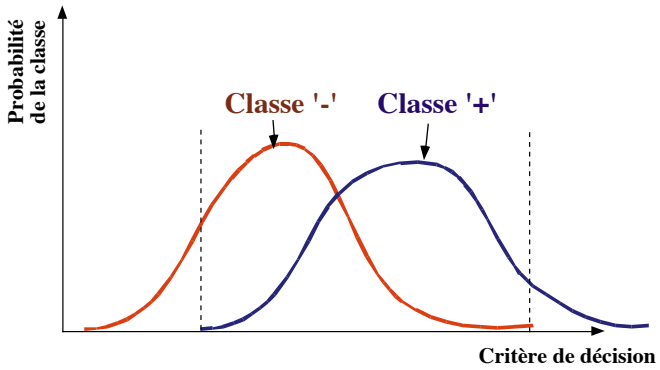
Types d'erreurs

- Erreur de type 1 (alpha) : *faux positifs*
  - Probabilité d'accepter l'hypothèse alors qu'elle est fausse
- Erreur de type 2 (beta) : *faux négatifs*
  - Probabilité de rejeter l'hypothèse alors qu'elle est vraie

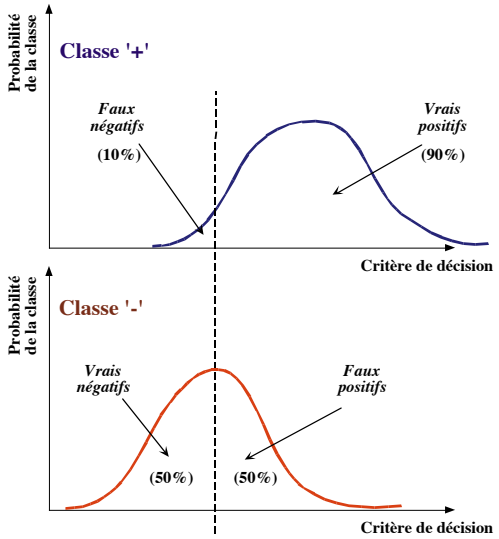
Comment arbitrer entre ces types d'erreurs ?

Courbe ROC

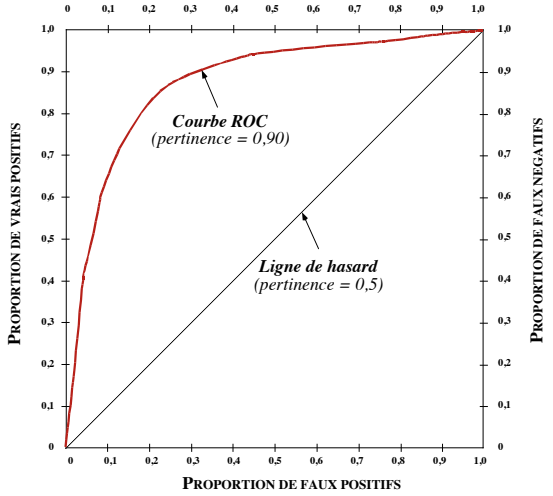
ROC = Receiver Operating Characteristic



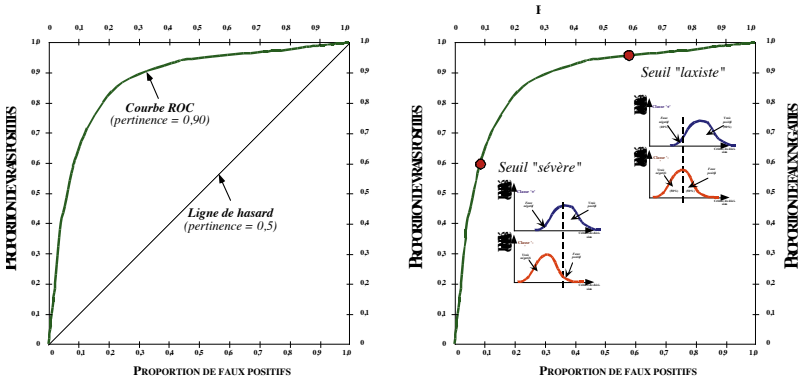
La courbe ROC



La courbe ROC



La courbe ROC



Courbe ROC

- *Spécificité* $\frac{VP}{VP + FN}$
- *Sensibilité* $\frac{VN}{FP + VN}$
- *Rappel* $\frac{VP}{VP + FN}$
- *Précision* $\frac{VP}{VP + FP}$

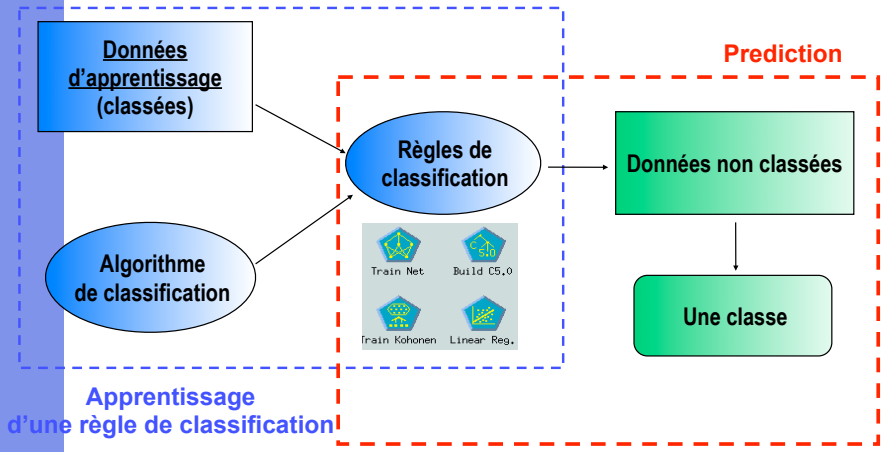
Réel / Estimé	+	-
+	VP	FP
-	FN	VN

Résumé

- **Attention à votre fonction de coût :**
  - **qu'est-ce qui importe pour la mesure de performance ?**
- **Données en nombre fini:**
  - **calculez les intervalles de confiance**
- **Données rares :**
  - **Attention à la répartition entre données d'apprentissage et données test. Validation croisée.**
- **N'oubliez pas l'ensemble de validation**
- **Mesure de la précision (accuracy) 100-erreur%**

- **L'évaluation est très importante**
  - **Ayez l'esprit critique**
  - **Convainquez-vous vous même !**

Construction automatique de classeurs

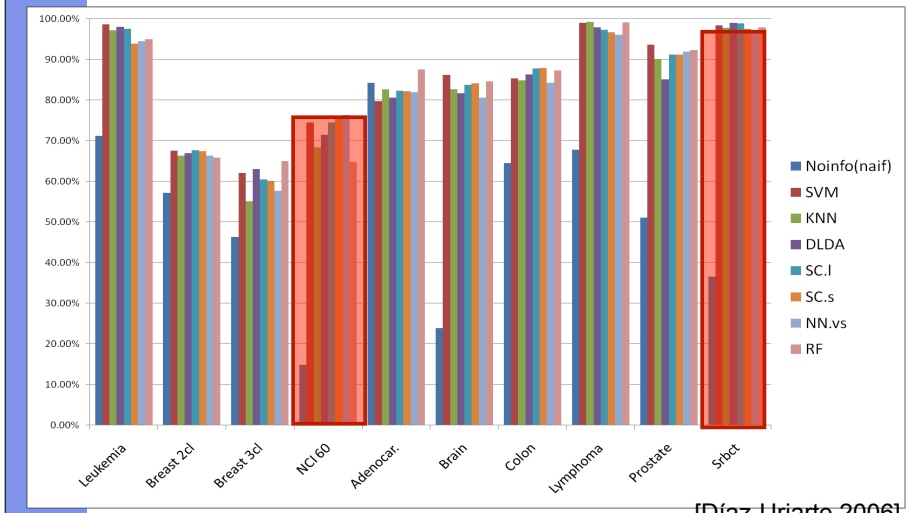


Bases d'apprentissage de la littérature

Nombre d'attributs = O([NB exemples]^2)

Dataset	Original Ref.	AttributsGenes	Exemples Patients	Classes
Leukaemia	[44]	3051	38	2
Breast	[9]	4869	78	2
Breast	[9]	4869	96	3
NCI 60	[61]	5244	61	8
Adenocarcinoma	[62]	9868	76	2
Brain	[63]	5597	42	5
Colon	[64]	2000	62	2
Lymphoma	[65]	4026	62	3
Prostate	[66]	6033	102	2
Srbct	[67]	2308	63	4

Résultats d'apprentissage sur ces bases



[Díaz-Uriarte,2006]

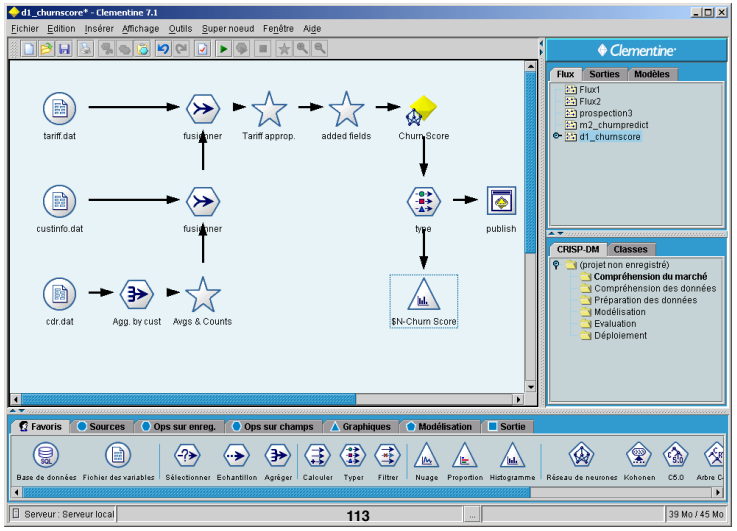
Plan

Démarche pour la prédiction et/ou la classification

Choisir une tâche de prédiction (ex: prédire la réponse au VLCD)

1. **Définir les classes** (ex: répondeur)
2. **Choisir un ou des algorithmes à utiliser**
   (ex: un SVM, Réseau de neurones, k-plus-proche-voisins, arbre de décision, etc.)
3. **Construire le classeur à partir des données** (ex: un réseau)
4. **Evaluer le classeur** (ex: validation)
5. **Analyser et améliorer les résultats**

Clementine (SPSS, IBM)



Clementine, un atelier de data mining

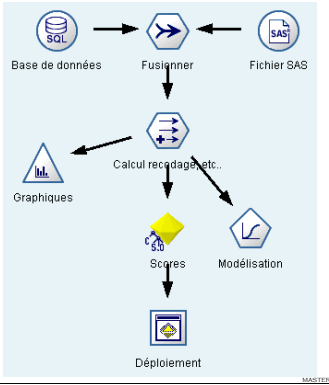
Une solution intégrée (un outil unique) pour :

- **l'accès aux données dans des sources multiples :**
  - **SGBD/R (Oracle9i, DB/2, Teradata, etc.)**
  - **Fichiers ASCII et propriétaires (SAS)**
- **la préparation des données (fusion, agrégation, calcul, etc.)**
- **l'analyse (exploration statistique et visuelle)**
- **la modélisation**
- **l'industrialisation (ou déploiement)**

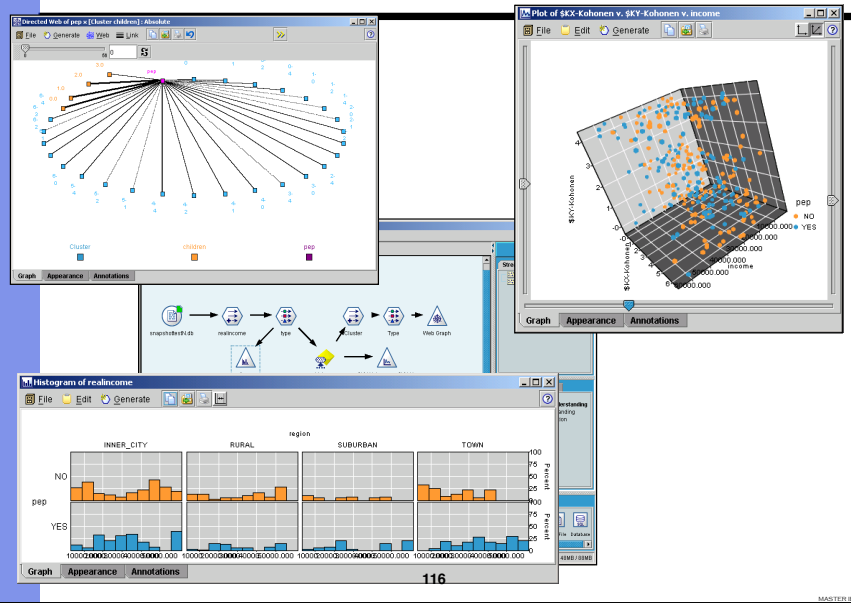
Clementine : programmation visuelle

La construction de flux de traitement des données à l'aide d'icones :

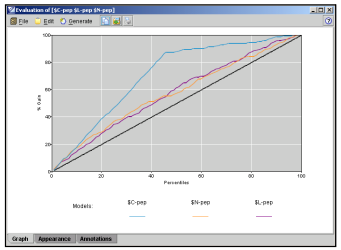
- Accès
- Préparation
- Exploration graphique
- Modélisation
- Déploiement

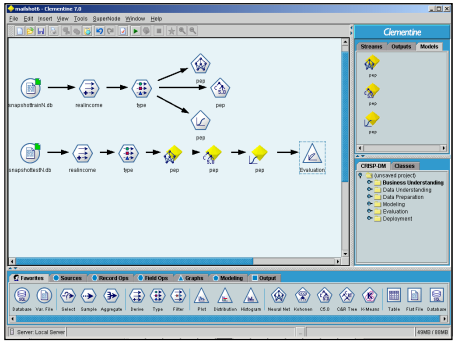


Visualisation

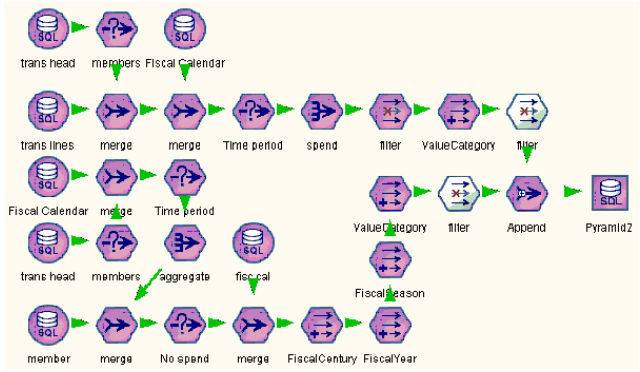


Différentes techniques





Fortes volumétries



- Maintenant en TP...
  ouebe.org ou
- http://zucker.limbio-paris13.org/COURS/M2IBM

Bibliographie