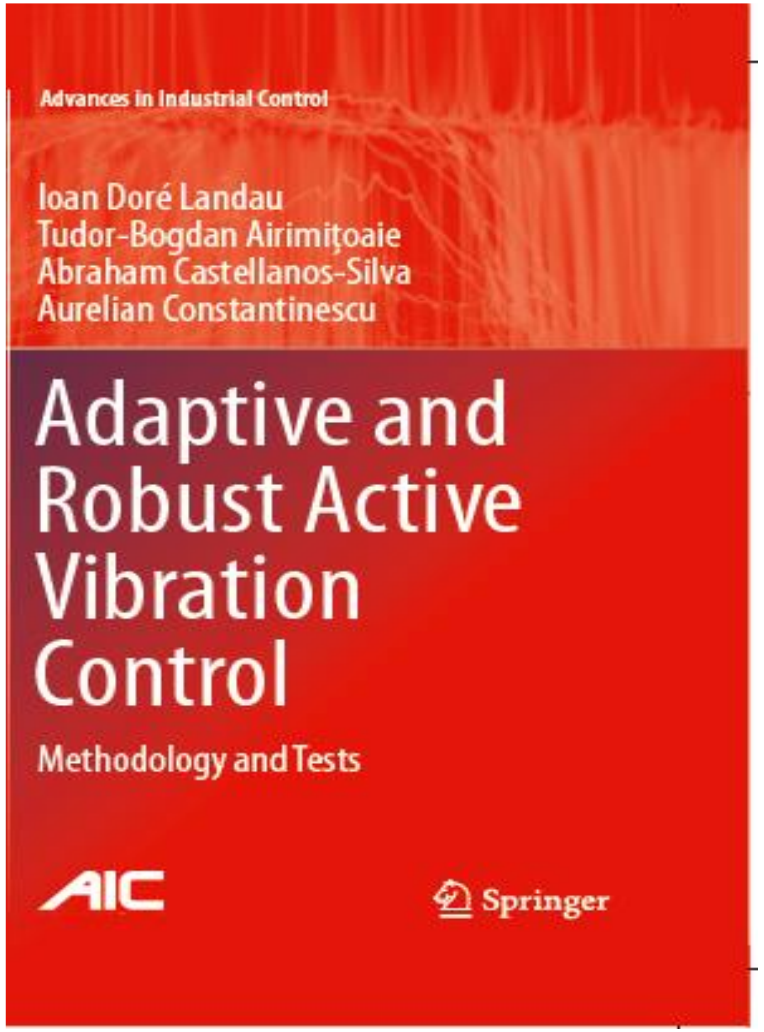
Data Mining et Machine Learning dans les Big Data

Une tentative de démystification

Ioan Landau
Vlad Landau

GIPSA –LAB, Grenoble, 3 Novembre 2016

New!
2016



Il y a du « data mining »
et du « machine learning »

http://www.landau-adaptivecontrol.org/

Autres titres possibles (sous réserve de quelques modifications)

Techniques d'identification et d'adaptation pour le data mining dans les big data

Autres titres possibles (sous réserve de quelques modifications)

Techniques d'identification et d'adaptation pour le data mining dans les big data

ou

Comment gagner des millions avec des Séquences Binaires Pseudo Aléatoires (SBPA)

Autres titres possibles (sous réserve de quelques modifications)

Techniques d'identification et d'adaptation pour le data mining dans les big data

ou

Comment gagner des millions avec des Séquences Binaires Pseudo Aléatoires (SBPA)

ou

Comment lever des capitaux avec l'algorithme d'adaptation paramétrique du gradient

Autres titres possibles (sous réserve de quelques modifications)

Techniques d'identification et d'adaptation pour le data mining dans les big data

ou

Comment gagner des millions avec des Séquences Binaires Pseudo Aléatoires (SBPA)

ou

Comment lever des capitaux avec l'algorithme d'adaptation paramétrique du gradient

ou

Data Science – un nouveau Eldorado

Quelques éléments de langage

J'ai un modèle.
Il me faut des données pour le vérifier

Ancien paradigme

J'ai des données.
Quel est le modèle générateur?

Nouveau paradigme

Quelques éléments de langage

J'ai un modèle.
Il me faut des données pour le vérifier

Ancien paradigme

J'ai des données.
Quel est le modèle générateur?

Nouveau paradigme

Les « **données** » sont un « **bien** » (asset) qu'il faut valoriser

Valorisation des données → Vente des fichiers
Valorisation des données → **Data Mining**

Data Mining, c'est un investissement, qui doit rapporter!

Quelques éléments de langage

J'ai un modèle.
Il me faut des données pour le vérifier

Ancien paradigme

J'ai des données.
Quel est le modèle générateur?

Nouveau paradigme

Les « données » sont un « bien » (asset) qu'il faut valoriser

Valorisation des données → Vente des fichiers
Valorisation des données → **Data Mining**

Data Mining, c'est un investissement, qui doit rapporter!

Données = *Matière première*, Algorithmes= *Machine Outils*, Big Data Centers = *Nouvelles usines*

Data Mining

Définition 1

Le Data Mining (exploration des données) a comme objectif de bâtir un modèle de prédiction à partir des données.

Data Mining

Définition 1

Le Data Mining (exploration des données) a comme objectif de bâtir un ***modèle de prédiction*** à partir des données.

Définition 2

Le Data Mining a comme objectifs de trouver dans les données des variables (prédictives) significatives (***attributs***) et de construire un ***modèle de prédiction*** à partir de ces variables.

Data Mining

Définition 1

Le Data Mining (exploration des données) a comme objectif de bâtir un modèle de prédiction à partir des données.

Définition 2

Le Data Mining a comme objectifs de trouver dans les données des variables (prédictives) significatives (***attributs***) et de construire un ***modèle de prédiction*** à partir de ces variables

Le **Data Mining** comporte plusieurs étapes:

- Préparation des données
- Choix des variables significatives (attributs)
- Construction du Modèle de Prédiction
- Validation du Modèle de Prédiction
- Déploiement du Modèle de Prédiction

Il comporte en amont la compréhension du problème à résoudre (connaissance métier) et la compréhension des données

Forecasting Sales (Dataiku)

April 01, 2016 (http://www.dataiku.com/learn/guide/gallery/forecasting-sales.html)

Many retail businesses need accurate forecasting of the revenue produced by each of their stores.

……

In this project, we use historical data from the Rossman pharmacy chain **to build a predictive model** to forecast the revenue of each of their stores. This model can be run weekly or monthly and provide business actors with accurate predictions about the revenue for coming days or weeks. This information can then be used to optimize business practices and streamline operations.

Business Goal

We want to build a project to answer the following questions:

•What is the expected revenue for each store on each day?

•What factors influence the revenue of a store most?

Non! « To Validate »

How Do We Do This?

We start with 2 different data sources: - two datasets with the revenue per store per day, split between our historical data (**used to train the model**), and our forecasting data (used to deploy our model) - a dataset with information about each store.

Like many data projects, we then proceed with three steps: 1. **Data Cleaning:** we clean our data and build our features 2. **Predictive Modeling:** we build and deploy ***a predictive model*** 3. **Visualization:** we create a useful visualization of our predicted data

Un constat

- Les automaticiens spécialistes en identification de systèmes font sans le savoir du « data mining » et du « machine learning »
- Idem pour ceux qui font de la commande ou régulation adaptative
- Les traiteurs de signaux font du « data mining » et du « machine learning »

Data Mining, Machine Learning, Big Data

1) **Concepts**

2) **Algorithmes**

3) Outils

L'accent sera mis sur les concepts et les algorithmes

Big Data

- Les « Big Data » sont des bases de données (relationnelles)
- Les « bases de données » sont un objet d'étude en informatique
- Les structures des bases de données et les outils de manipulation évoluent sans cesse.
- Les outils « à la mode »: langage NoSQL, MapReduce, Hadoop, ….

Big Data:
Volume de la base de données à partir duquel les données ne peuvent plus être traitées en un *temps raisonnable (vitesse de traitement)* par des systèmes constitués d'un seul nœud.

C'est le rapport *volume/vitesse* qui définit (d'une façon floue) la frontière des Big Data.

Big Data

- Les « Big Data » sont des bases de données (relationnelles)
- Les « bases de données » sont un objet d'étude en informatique
- Les structures des bases de données et les outils de manipulation évoluent sans cesse.
- Les outils « à la mode »: langage NoSQL, MapReduce, Hadoop, ….

Big Data:
Volume de la base de données à partir duquel les données ne peuvent plus être traitées en un *temps raisonnable (vitesse de traitement)* par des systèmes constitués d'un seul nœud.

C'est le rapport *volume/vitesse* qui définit (d'une façon floue) la frontière des Big Data.

Remarques:
1) Des problèmes de *data mining* se posent souvent pour des BD qui n'entrent pas dans la définition ci-dessus
2) Ne pas confondre « data mining » avec « data processing » (manipulation des big data)
3) Les travaux de M. Berekmi (Thèse Gipsa-lab dir.: B. Robu, N. Marchand) ont montré qu'on peut interfacer les structures modernes de traitement des BD pour faire des « mesures » et appliquer une « commande » pour optimiser le fonctionnement des cloud computer centers.

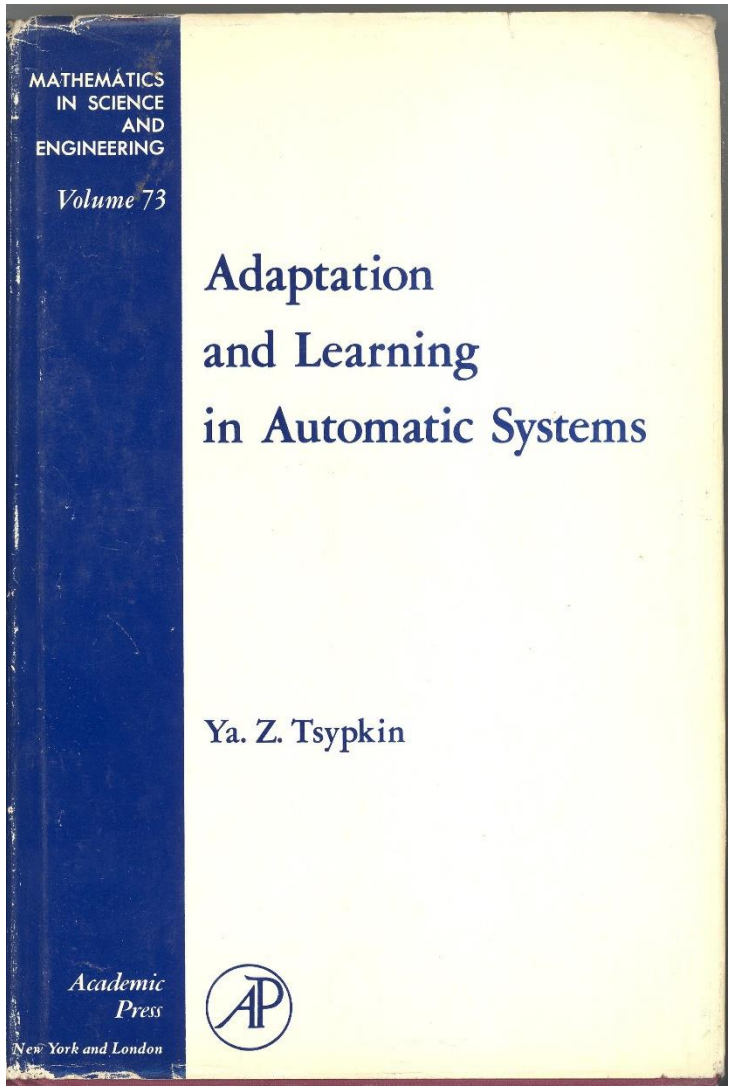
Techniques (Algorithmes) d'Apprentissage Automatique (Automatic (Machine) Learning Systems/Algorithms)

- Ce n'est pas les informaticiens qui ont inventé les algorithmes d'apprentissage (la bible du « machine learning » (Mitchell) date de 1997)

- La communauté Automatique s'est intéressée depuis la fin des années 60 aux techniques d'apprentissage

- Un groupe de travail IEEE a été constitué dans les années 1970 (G. Saridis)

- Une série de symposiums IFAC « Adaptive and Learning in Control Systems and Signal Processing » a été mise en place en 1983
  (IFAC Technical Commitee « Adaptive and Learning Systems »)



Version en russe 1968, version anglaise 1971

Machine Learning ?

La définition du « machine learning » (Mitchell « Machine Learning » 1997) est une « imposture » scientifique (*how to make computers learn*)
mais c'est une expression « vendeuse »

La pseudo définition est contredite par l'auteur même (lire la préface)

Les ordinateurs n'apprennent pas!

Machine Learning ?

La défintion du « machine learning » (Mitchell « Machine Learning » 1997) est une « imposture » scientifique (*how to make computers learn*)
mais c'est une expression « vendeuse »

La pseudo définition est contredite par l'auteur même (lire la préface)

Les ordinateurs n'apprennent pas!

On programme des algorithmes d'apprentissage sur un ordinateur

Le domaine de l'apprentissage comporte des paradigmes et des algorithmes.

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. (Wikipedia)

Machine Learning = Apprentissage Automatique

Quelques termes utilisés en apprentissage

Apprentissage non - supervisé (unsupervised learning)

- Recherche des attributs
- Classification

Apprentissage supervisé (supervised learning)

- Définition de la variable cible (target variable)
- Définition des attributs (variables prédictives)
- Modèle de prédiction

Apprentissage supervisé **renforcé** (reinforced supervised learning)
(quand on applique une excitation pour mieux apprendre)

Terme alternatif: ***active*** *(supervised) learning* (apprentissage (supervisé) ***actif***)

Modèles de prédiction exogènes et endogènes

Modèle endogène

Relation entre les données acquises sur un site de e-commerce et le chiffre d'affaire du site

Modèle exogène (un entrée externe = signal de commande)

Relation entre la politique de promotion (taux de réduction, type de promotion) et le nombre de ventes (ou le chiffre d'affaires)

Relation dynamique entre le couple de commande d'un bateau et l'évolution de la position du bateau dans le temps

Types de modèles de prédiction

Modèle → Phénoménologique (de connaissance)

Modèle → Comportement

Modèle phénoménologique: à partir d'une compréhension approfondie des phénomènes (ex: physique – équation du mouvement)

Pas de modèles phénoménologiques fiables pour beaucoup de situations (motivations d'achats?)

Mêmes pour les systèmes physiques réels, les modèles de connaissance ne sont pas complets et leur paramètres sont de toute façon à estimer

Types de modèles de prédiction

Modèle → Phénoménologique (de connaissance)

Modèle → Comportement

Modèle phénoménologique: à partir d'une compréhension approfondie des phénomènes (ex: physique – équation du mouvement)

Pas de modèle phénoménologique fiable pour beaucoup de situations (motivations d'achats?)

Mêmes pour les systèmes physiques réels, les modèles de connaissance ne sont pas complets et leur paramètres sont de toute façon à estimer

Modèle de comportement (endogène ou exogène) : établit des relations entre des données (souvent appelées entrées ou variables prédictives ou attributs) et une ou plusieurs variables objectifs (appelées "sorties" ou "variables cibles") .

Un MODELE DE COMPORTEMENT permet de faire de la prédiction ou calculer une action optimale

Un **modèle de comportement** (prédiction) peut être obtenu à partir de données (à condition qu'elles soient significatives)

Modélisation et Identification

Modélisation:

Détermine un modèle de prédiction à partir de l'écriture des equations qui décrivent les phénomènes et en supposant connaitre certaines constantes physiques, chimiques, etc.

Identification:

Utilisation exclusive des données pour extraire un modèle de prédiction (appelé aussi « Modèle de Comportement »).

Question: *Ya-t-il des équations qui gouvernent le fonctionnement d'un site de vente par internet?*

Modélisation et Identification

Modélisation:

Détermine un modèle de prédiction à partir de l'écriture des equations qui décrivent les phénomènes et en supposant connaitre certaines constantes physiques, chimiques, etc.

Identification:

Utilisation exclusive des données pour extraire un modèle de prédiction (appelé aussi « Modèle de Comportement »).

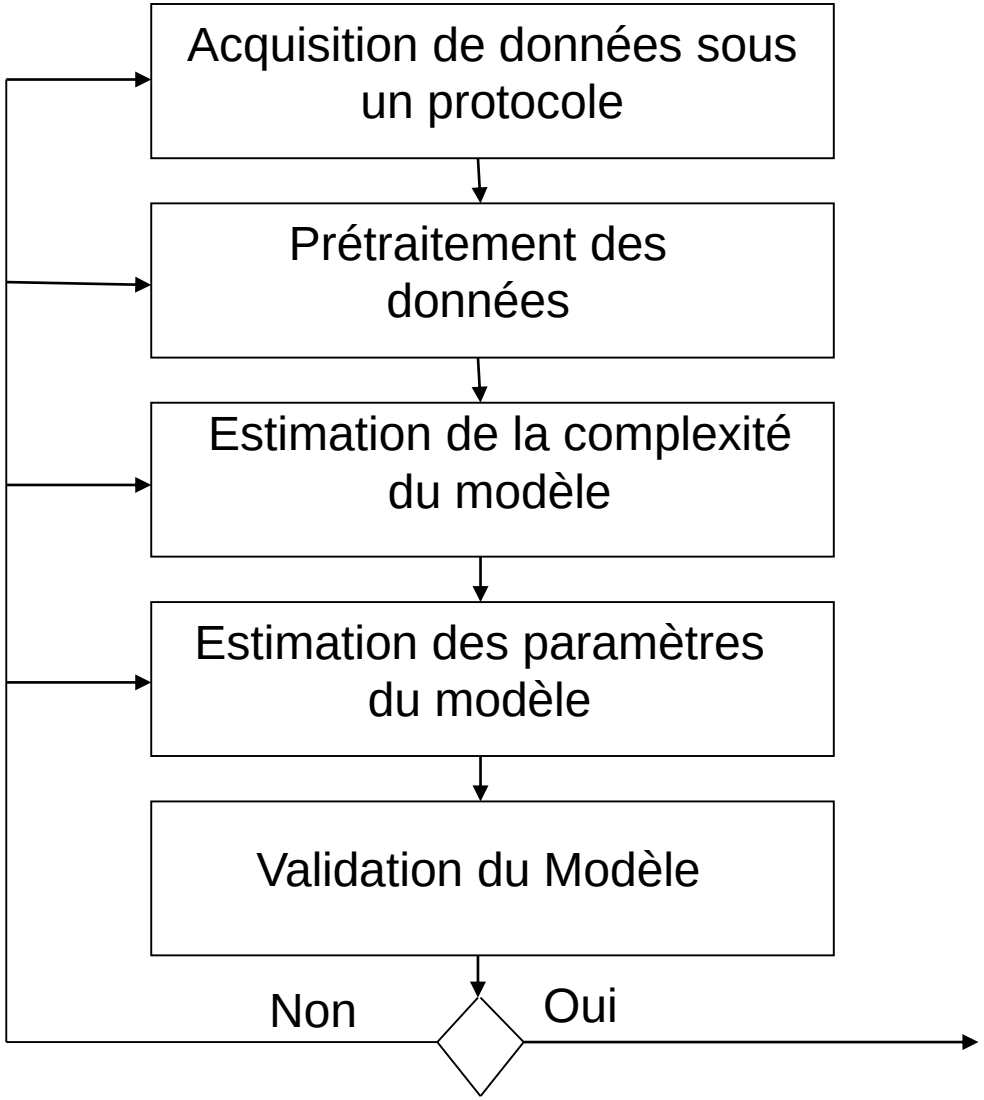
Question: *Ya-t-il des équations qui gouvernent le fonctionnement d'un site de vente par internet?*

Modèle « explicatif »:

- La modélisation phénoménologique (à partir des équations) produit un modèle « explicatif »;
- La connaissance « métier » permet dans certains cas d'établir un modèle qualitatif de type *explicatif*;
- Les modèles de comportement peuvent être rendus « explicatifs ».

Comment obtenir un modèle de prédiction (identification) ?

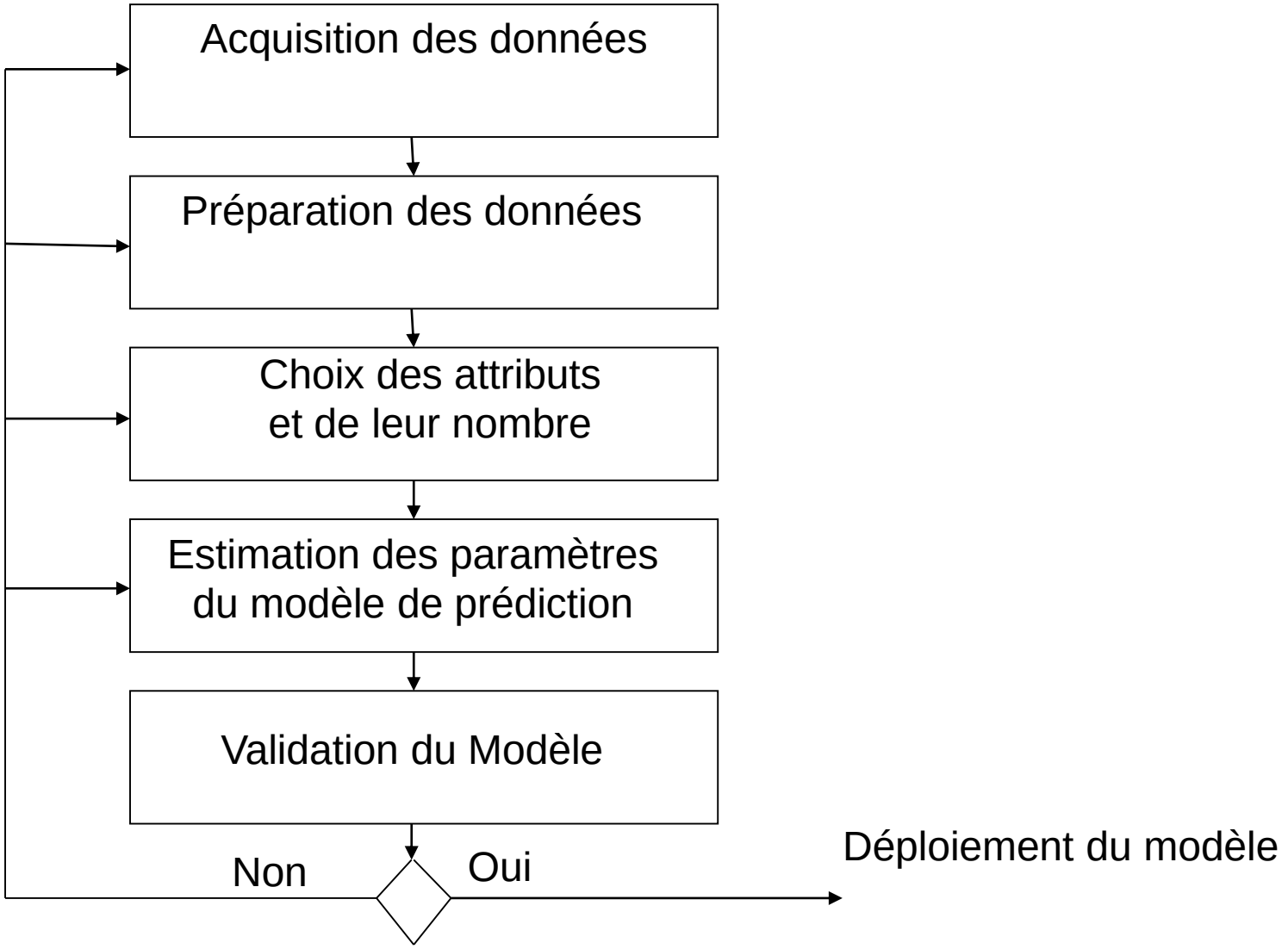
Cette procédure date des années 75



Utilisation pour la prédiction ou le calcul d'une commande

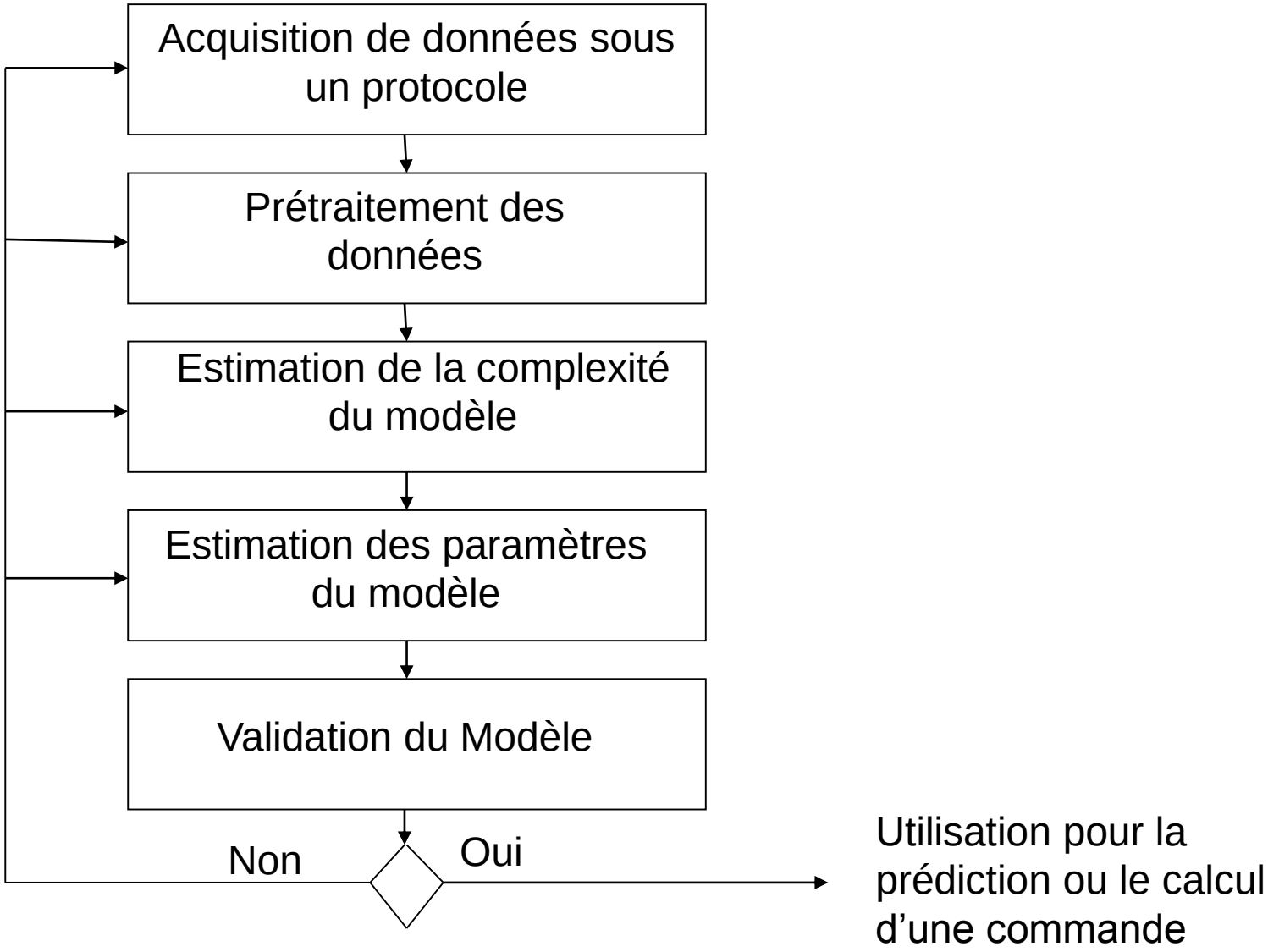
Comment obtenir un modèle de prédiction (identification) ?

L'identification avec le vocabulaire « data mining »

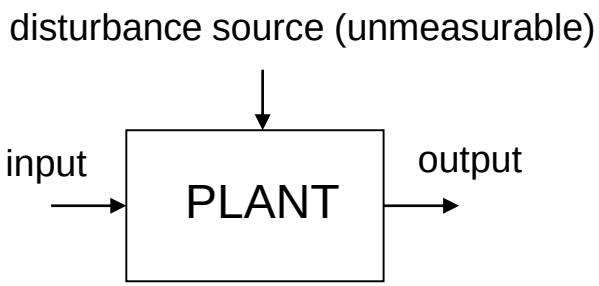


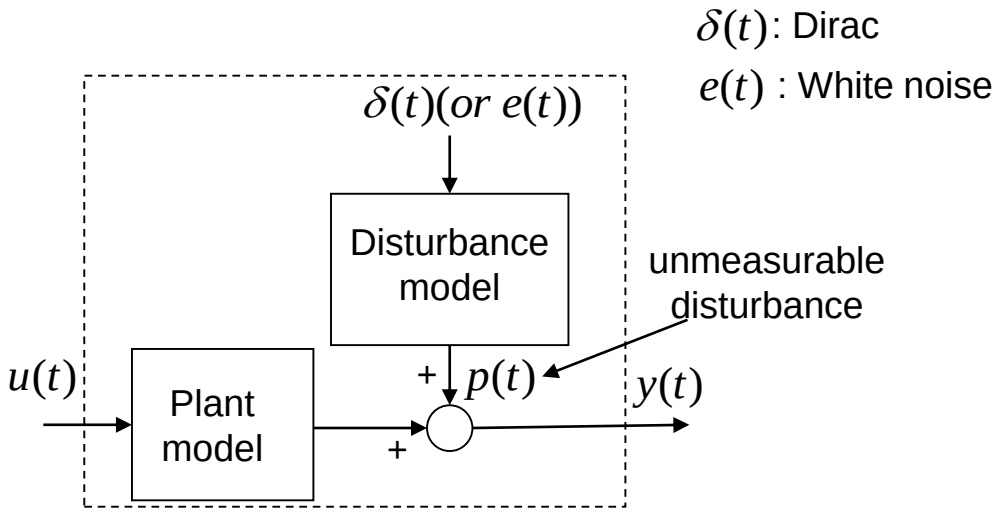
Standard CRISP-DM (Cross Industry Standard Process for Data Mining) - 2004

Comment obtenir un modèle de prédiction (identification) ?



Data Mining et Machine Learning en Automatique = *Identification des systèmes*





Identification du modèle du procédé = *Reinforced Supervised Learning*

Identification du modèle de la perturbation = *Supervised Learning*

Type de modèle pour le procédé : FIR ou IIR

$$G(q^{-1}) = \frac{q^{-d}B(q^{-1})}{A(q^{-1})} = \frac{q^{-d}(b_1 q^{-1} + \ldots + b_{n_B} q^{-n_B})}{1 + a_1 q^{-1} + \ldots + a_{n_A} q^{-n_A}}$$

$$FIR : a_i = 0$$

$$n = \max(n_a, n_b + d) \longleftarrow \text{Ordre du modèle}$$

Type de modèle pour la perturbation : MA ou ARMA

$$H(q^{-1}) = \frac{C(q^{-1})}{D(q^{-1})}$$

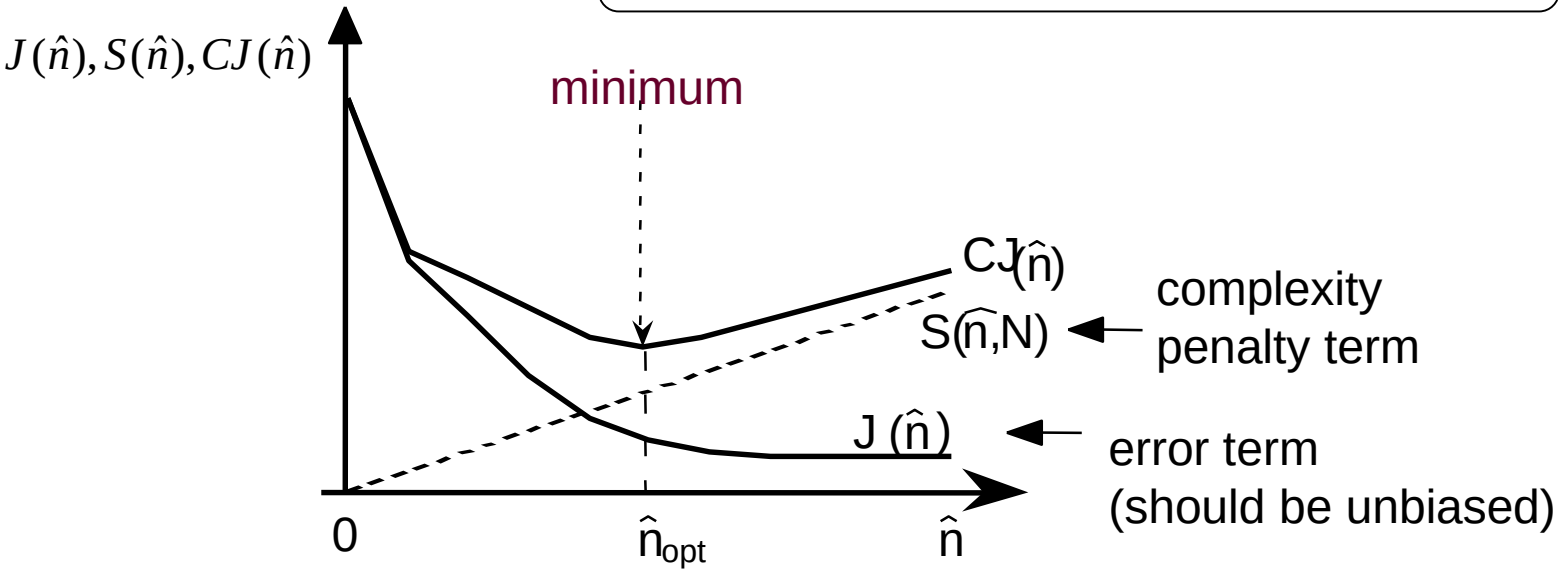
Complexity Estimation from I/O Data

Objective :

To get a good estimation of the model complexity (n_A, n_B, d) directly from noisy data

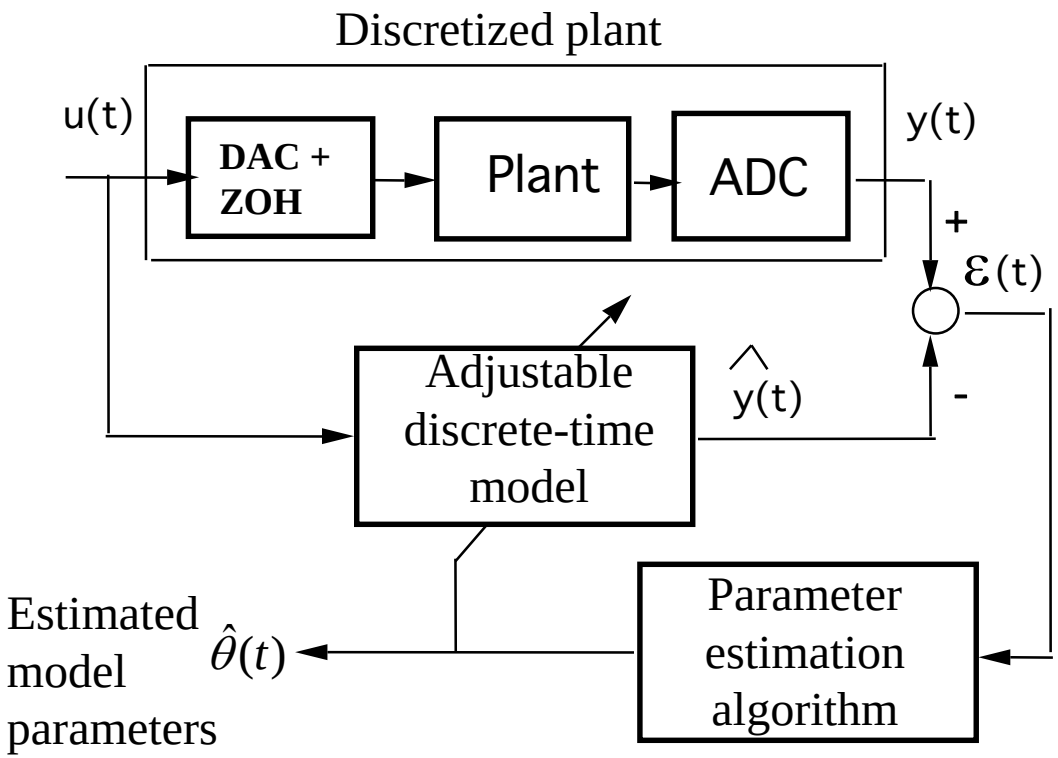
$$n = \max(n_A, n_B + d)$$

$$\boxed{\hat{n}_{opt} = \min_{\hat{n}} CJ = \min_{\hat{n}} \left[J(\hat{n}) + S(\hat{n}, N) \right]}$$



To get a good order estimation, J should tend to the value for noisy free data when $N \to \infty$ *(use of instrumental variables)*

Parameter Estimation



Parameter estimation alg.: → batch (non recursive)
→ recursive

In the presence of measurement noise, it does not exists a unique algorithm providing good results in all the situations encountered in practice

https://www.youtube.com/watch?v=5u4G23_OohI&index=2&list=PLA89DCFA6ADACE599

Lecture 2 cours de Machine Learning à Stanford: *Algorithme du gradient*

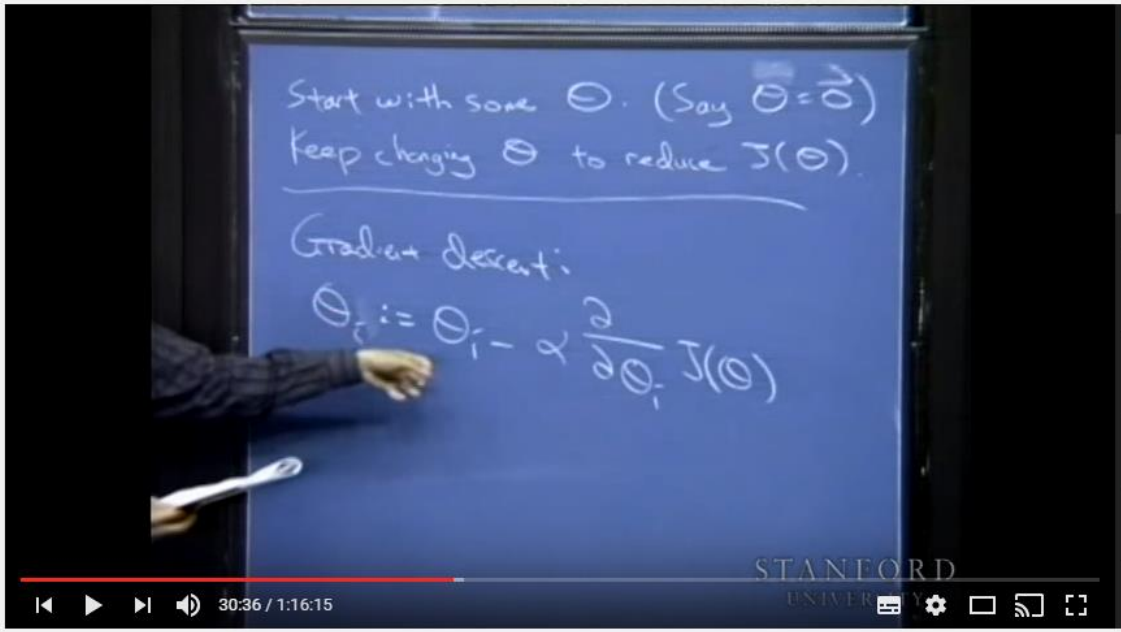
Temps : 30:52

Lecture 2 | Machine

https://www.youtube.com/watch?v=5u4G23_OohI&index=2&list=PLA89DCFA6ADACE599

Apps | Bookmarks | Box | Atout jazz | Box |Jujuers en lic | Importés depuis I | Discount Marine | Voyages-sncf.com | easyJet | Impôts | Other bookmarks

YouTube FR — Rechercher — Mettre en ligne



Lecture Collection | Machine Learning
Stanford • 2/20 vidéos

1. Lecture 2 | Machine Learning (Stanford) — Stanford
3. Lecture 3 | Machine Learning (Stanford) — Stanford
4. Lecture 4 | Machine Learning (Stanford) — Stanford
5. Lecture 5 | Machine Learning (Stanford) — Stanford
6. Lecture 6 | Machine Learning (Stanford) — Stanford
7. Lecture 7 | Machine Learning (Stanford) — Stanford

Lecture 2 | Machine Learning (Stanford)

Stanford — S'abonner 566 598

433 682 vues

Ajouter à — Partager — Plus — 1 290 — 41

Mise en ligne le 22 juil. 2008

Lecture by Professor Andrew Ng for Machine Learning (CS 229) in the Stanford Computer Science department. Professor Ng lectures on linear regression, gradient descent, and normal equations and discusses how they relate to machine learning.

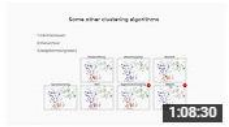
PLUS

COMMENTAIRES • 242

Ajouter un commentaire public...

Lecture 3 | Machine Learning (Stanford)
Stanford
238 983 vues

Lecture 1 | Machine Learning (Stanford)
Stanford
1 181 441 vues



Introduction to Data Analysis using Machine Learning
David Taylor
36 408 vues

Deep Learning: Intelligence from Big Data
vlabvideos

FR — 18:14 — 21/10/2016

Algorithms for parameter estimation

Discrete time plant model (unknown parameters))

$$y(t+1) = -a_1 y(t) + b_1 u(t) = \theta^T \phi(t)$$

Measurement vector

$\theta^T = [a_1, b_1]$ ← Parameter vector ; $\phi(t)^T = [-y(t), u(t)]$

Vector of attributes

Adjustable prediction model (à priori)

$$\hat{y}^o(t+1) = \hat{y}(t+1|\hat{\theta}(t)) = -\hat{a}_1(t) y(t) + \hat{b}_1(t) u(t) = \hat{\theta}(t)^T \phi(t)$$

$\theta(t)^T = [\hat{a}_1(t), \hat{b}_1(t)]$ ← Vector of adjustable parameters

Prediction error (à priori)

$$\varepsilon^o(t+1) = y(t+1) - \hat{y}^o(t+1) = \varepsilon^o(t+1, \hat{\theta}(t))$$

Criterion to be minimized (objective):

$$J(t+1) = \left[\varepsilon^o(t+1)\right]^2 = \left[\varepsilon^o(t+1, \hat{\theta}(t))\right]^2$$

?

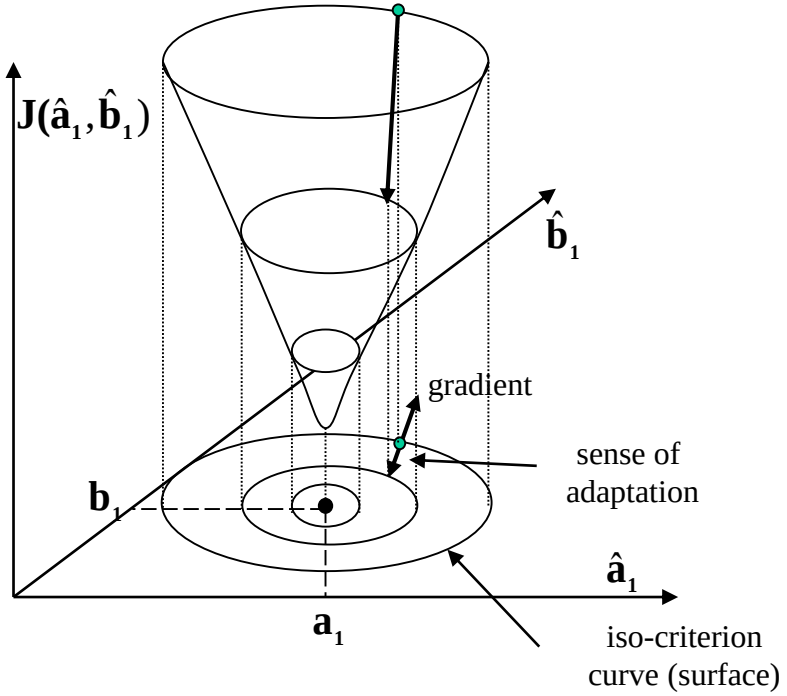
Parameter adaptation algorithm

$$\hat{\theta}(t+1) = \hat{\theta}(t) + \Delta\hat{\theta}(t+1) = \hat{\theta}(t) + f\left(\hat{\theta}(t), \phi(t), \varepsilon^o(t+1)\right)$$

PAA – Gradient algorithm

Criterion to be minimized (objective):

$$\min_{\hat{\theta}(t)} J(t+1) = \left[\varepsilon^o(t+1)\right]^2$$



gradient of the criterion

$$\hat{\theta}(t+1) = \hat{\theta}(t) - F\,\frac{\partial J(t+1)}{\partial \hat{\theta}(t)}$$

$F = \alpha I \quad (\alpha > 0)$ **(*I* = unit matrix)**

$$\frac{1}{2}\frac{\partial J(t+1)}{\partial \hat{\theta}(t)} = \frac{\partial \varepsilon^o(t+1)}{\partial \hat{\theta}(t)}\,\varepsilon^o(t+1)$$

$$\varepsilon^o(t+1) = y(t+1) - \hat{y}^o(t+1) = y(t+1) - \hat{\theta}(t)^T\phi(t) \quad \Rightarrow \quad \frac{\partial \varepsilon^o(t+1)}{\partial \hat{\theta}(t)} = -\phi(t)$$

$$\hat{\theta}(t+1) = \hat{\theta}(t) + F\phi(t)\varepsilon^o(t+1)$$

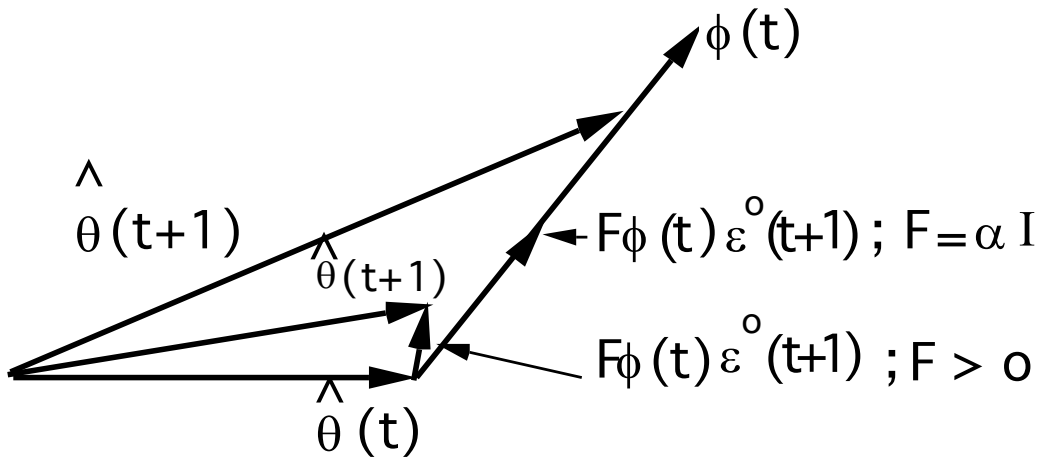
PAA – Gradient algorithm

$$\boxed{\hat{\theta}(t+1) = \hat{\theta}(t) + F\phi(t)\varepsilon^{o}(t+1)}$$

Adaptation gain

$$\begin{cases} F = \alpha I & (\alpha > 0) \\ F > 0 & \textit{Positive definite matrix} \end{cases}$$

Geometrical interpretation



Attention: Instability risk if $F\ (\alpha)$ is large !!

(see book Landau-Zito, pg. 213 – 214 for details)

PAA – Improved gradient algorithm

a posteriori output of the adjustable predictor

$$\hat{y}(t+1) = \hat{y}(t+1|\hat{\theta}(t+1)) = -\hat{a}_1(t+1)y(t) + \hat{b}_1(t+1)u(t) = \hat{\theta}(t+1)^T \phi(t)$$

Prediction error (a posteriori): $\varepsilon(t+1) = y(t+1) - \hat{y}(t+1)$

Criterion to be minimized (objective): $\boxed{\min_{\hat{\theta}(t+1)} J(t+1) = \left[\varepsilon(t+1)\right]^2}$

$$\boxed{\hat{\theta}(t+1) = \hat{\theta}(t) + F\phi(t)\varepsilon(t+1)}$$

$$\varepsilon(t+1) = \frac{\varepsilon^o(t+1)}{1+\phi(t)^T F\phi(t)}$$

$$\boxed{\hat{\theta}(t+1) = \hat{\theta}(t) + \frac{F\phi(t)\varepsilon^o(t+1)}{1+\phi(t)^T F\phi(t)}}$$

Stable for any *F* > *0*

PAA - (Recursive) Least Squares Algorithm

The Least Squares Criterion:

$$\min_{\hat{\theta}(t)} J(t) = \frac{1}{t}\sum_{i=1}^{t}\left[y(i) - \hat{\theta}(t)^T \phi(i-1)\right]^2 = \frac{1}{t}\sum_{i=1}^{t}\varepsilon^2\left(i, \hat{\theta}(t)\right)$$

$$\boxed{\begin{aligned} \hat{\theta}(t+1) &= \hat{\theta}(t) + F(t+1)\Phi(t)\varepsilon^o(t+1) \\ F^{-1}(t+1) &= F^{-1}(t) + \Phi(t)\Phi(t)^T \end{aligned}}$$

$$F^{-1}(t+1) = \lambda_1(t)F^{-1}(t) + \lambda_2(t)\Phi(t)\Phi(t)^T$$

$$0 < \lambda_1(t) \le 1 \;;\; 0 \le \lambda_2(t) < 2 \;;\; F(0) > 0$$

General form for the adaptation gain updating (real-time)

$$F(t+1) = \frac{1}{\lambda_1(t)}\left[F(t) - \frac{F(t)\Phi(t)\Phi(t)^T F(t)}{\dfrac{\lambda_1(t)}{\lambda_2(t)} + \Phi(t)^T F(t)\Phi(t)}\right]$$

Validation du « modèle de prédiction » estimé

C'est sans aucun doute l'étape la plus importante !

Il faut être sur que le modèle estimé (quelle que soit la structure et la méthode d'apprentissage utilisée) garantit une bonne prédiction sur les données futures.

Validation du « modèle de prédiction » estimé

C'est sans aucun doute l'étape la plus importante !
Il faut être sur que le modèle estimé (quelle que soit la structure et la méthode d'apprentissage utilisée) garantit une bonne prédiction sur les données futures

Validation du modèle de prédiction sur la base de données d'apprentissage

Il faut s'assurer qu'on a extrait le modèle le plus performant à partir des données utilisées.

On utilise des tests statistiques.
Le plus courant est le ***test de blancheur*** sur l'erreur de prédiction résiduelle

Erreur de prédiction résiduelle « blanche » = Le modèle de prédiction a capturé tout ce qui peut être prédit

(on utilise aussi le *test de non corrélation* entre entrée et erreur de prédiction résiduelle)

Validation du « modèle de prédiction » estimé

C'est sans aucun doute l'étape la plus importante !
Il faut être sur que le modèle estimé (quelle que soit la structure et la méthode d'apprentissage utilisée) garantit une bonne prédiction sur les données futures

Validation du modèle de prédiction sur la base de données d'apprentissage

Il faut s'assurer qu'on a extrait le modèle le plus performant à partir des données utilisées ou qu'au moins l'erreur qu'on commet est très faible

On utilise des tests statistiques.
Le plus courant est le ***test de blancheur*** sur l'erreur de prédiction résiduelle

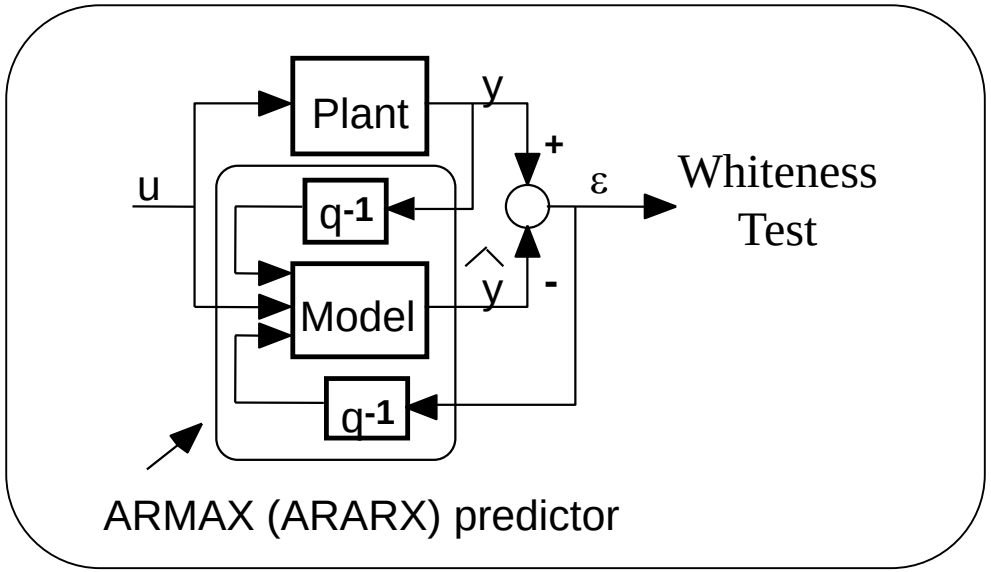
Erreur de prédiction résiduelle « blanche » = Le modèle de prédiction a capturé tout ce qui peut être prédit

(on utilise aussi le *test de non corrélation* entre entrée et erreur de prédiction résiduelle)

Remarque : *à ma grande surprise ces tests qui sont classiques en automatique ne sont pas utilisés par ceux qui font de l'apprentissage ou du data mining!*

Validation of Identified Models

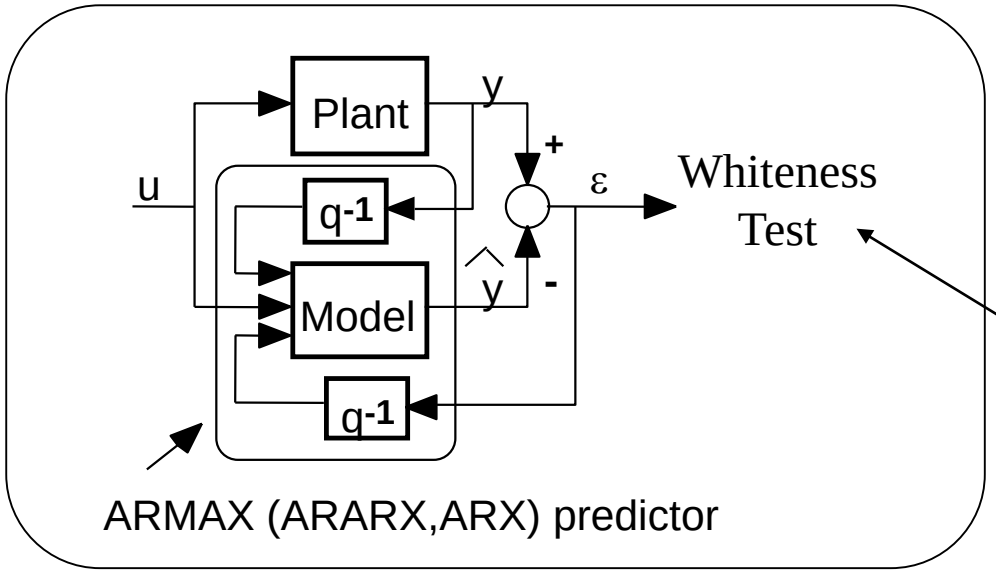
Statistical Validation (Whiteness Test)



Erreur de prédiction résiduelle « blanche » = Le modèle de prédiction a capturé tout ce qui peut être prédit

Validation of Identified Models

Statistical Validation (Whiteness Test)



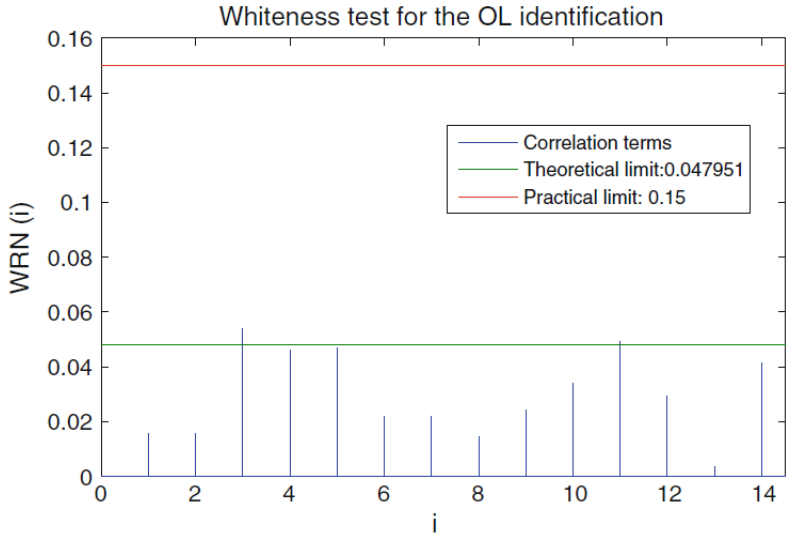
$$\left|RN(i)\right| \le \frac{2.17}{\sqrt{N}} \; ; \; i \ge 1$$

Normalized autocorrelations

number of data

$$R(0) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t) \quad ; \quad RN(0) = \frac{R(0)}{R(0)} = 1$$

$$R(i) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)\varepsilon(t-i) \quad ; \quad RN(i) = \frac{R(i)}{R(0)} \quad ; \quad i = 1,2,3..i_{\max}$$



- L'algorithme des Moindres Carrés ne passe que très rarement le test de blancheur !
- Il y a d'autre algorithmes bien plus performants

Identification dans l'ère « préhistorique » (avant 1973)

- Modèle : FIR (peu de IIR)
- Algorithmes d'estimation paramétrique : gradient, moindres carrés
- Données séparées en deux groupes (apprentissage, test)
- La validation porte uniquement sur l'estimation de l'ordre du modèle
- *Critère* : minimisation de l'erreur quadratique sur les données de test

Pas de test statistiques!

C'est la technique utilisée dans le data mining

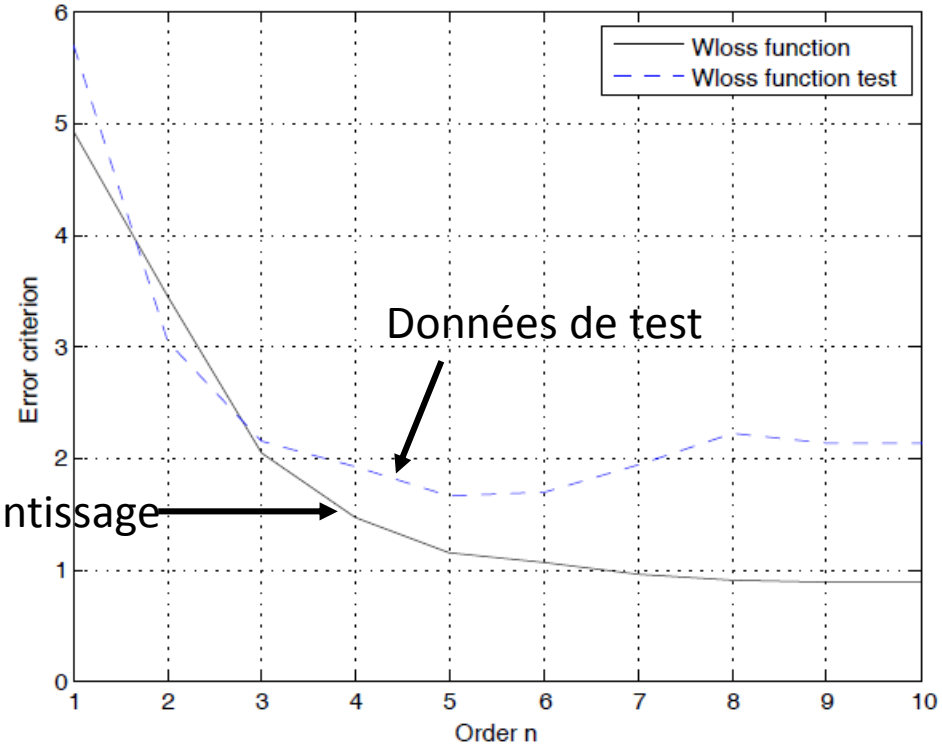
Identification dans l'ère « préhistorique » (avant 1973)

- Modèle : FIR (peu de IIR)
- Algorithmes d'estimation paramétrique : gradient, moindres carrés
- Données séparées en deux groupes (apprentissage, test)
- La validation porte uniquement sur l'estimation de l'ordre du modèle
- *Critère* : minimisation de l'erreur quadratique sur les données de test

Pas de test statistiques!

C'est la technique utilisée dans le data mining

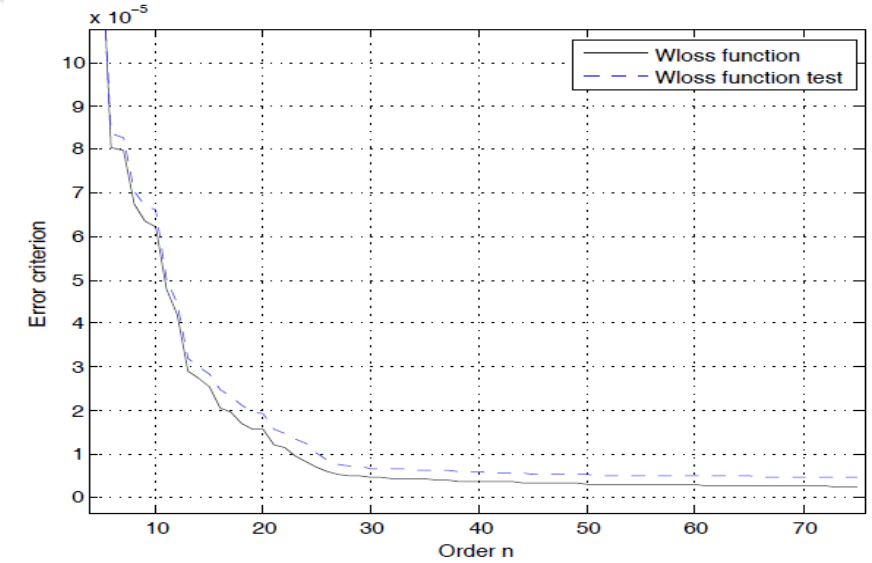
Pour cet exemple simulé avec bruit (n= max(na,nb+d)=3)
On estime n=5 avec la méthode « préhistorique »
Les méthodes actuelles donnent n=3 (na=2,nb=2,d=1)



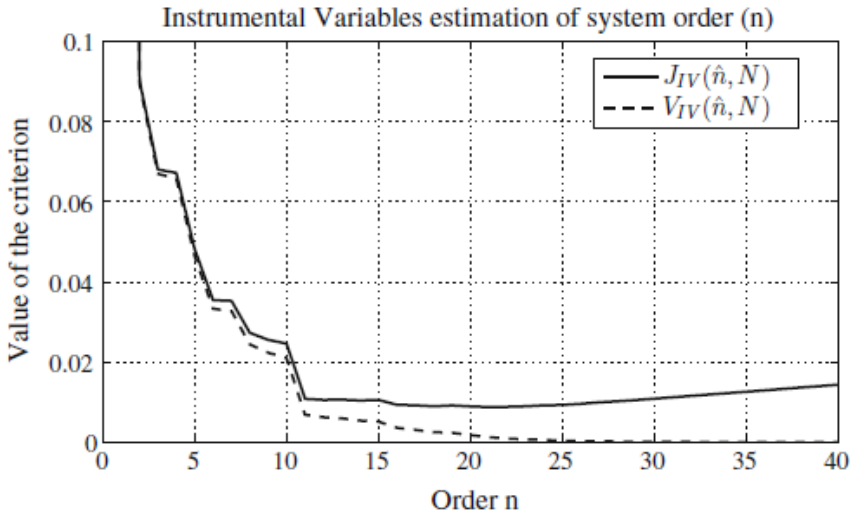
- ***Surestime le nb. de paramètres*** (10 param. au lieu de 4)
- ***L'estimation paramétrique est biaisée***
  (le modèle ne passe pas le test de blancheur)

« Identification préhistorique » appliquée au contrôle actif de vibrations (SAAI) (T. Airimitoaie)

Il n'y a pas de minimum sur les données de test !
(pas de surapprentissage)



L'ordre du modèle estimé par les méthodes « modernes » et qui passe les tests statistiques de validation est n=22



Data Mining

Acquisition des données sous un protocole

- Pas n’importe quelle type de données est utile pour estimer un modèle de comportement (prédictif)

- Il faut s’assurer qu’il y a une corrélation entre les données disponibles et l’objectif

- Les techniques d’analyse en composantes principales (ACP) sont une approche pour détecter les entrées (données) significatives

Acquisition des données sous un protocole

- Pas n'importe quelle type de données est utile pour estimer un modèle de comportement (prédictif)

- Il faut s'assurer qu'il y a une corrélation entre les données disponibles et l'objectif

- Les techniques d'analyse en composantes principales (ACP) sont une approche pour détecter les entrées (données) significatives

- *Pour les modèles exogènes il faut choisir des "excitations" (entrées) qui permettront d'estimer un modèle représentatif.*

- *Il faut que l'excitation soit "riche" (variée) pour que le modèle estimé soit indépendant de la forme particulière d'une entrée.*
  (condition pour une bonne prédiction en présence de nouvelles entrées)

Prétraitement des données

Le résultats de l'estimation d'un modèle de prédiction dépendent beaucoup du prétraitement des données

Techniques triviales:
-Mise à echelle
-Elimination des biais
(pour les mesures physiques mais aussi pour les résultats des enquétes)
-Synchronisation des données
-Elimination éventuelle des données abérantes

Techniques plus subtiles:
Agrégation de plusieurs informations et définition d'une nouvelle variable

Modèle de comportement (paramétrique)

La sortie du modèle de comportement est une fonction de plusieurs types de données (attributs) : $x_1, x_2, ...x_n$ et des paramètres $a_1, a_2, ...a_n$

La complexité du modèle va être donnée par le nombre des paramétres du modèle

Modèles → statiques

Modèles → dynamiques (on associe la notion de temps)

***Modèle statique*:** Poids du bar (y) en fonction de la taille (x)

$$y = ax + b = a_1 x_1 + a_2 x_2 = \mathrm{a}^T x = \theta^T \phi; \quad x_1 = x;\ x_2 = 1;\ a_1 = a;\ a_2 = b$$

Modèle de comportement (paramétrique)

***Modèle dynamique*:** Evolution de la position d'un bateau (y) en fonction de l'évolution du couple de commande (x)

$$y(t+1) = -a_1 y(t) - a_2 y(t-1) .... + b_1 x(t) + b_2 x(t-1) .....$$

L'évolution de la sortie y dépend des variables x1(t), x2(t)…xp(t) à l'instant t et de leur valeur à t-1, t-2… t-h (horizon).

Elle peut dépendre aussi des valeurs précédentes de y (mémoire dynamique)

Modèle de comportement (paramétrique)

***Modèle dynamique*:** Evolution de la position d'un bateau (y) en fonction de l'évolution du couple de commande (x)

$$y(t+1) = -a_1 y(t) - a_2 y(t-1) .... + b_1 x(t) + b_2 x(t-1) .....$$

L'évolution de la sortie y dépend des variables x1(t), x2(t)…xp(t) à l'instant t et de leur valeur à t-1, t-2… t-h (horizon).

Elle peut dépendre aussi des valeurs précédentes de y (mémoire dynamique)

Le chiffe d'affaire d'un site de vente en ligne dépend-il de l'évolution antérieure du chiffre d'affaire?

Remarque : on n'a pas trouvé des exemples de data mining prenant en considération une éventuelle mémoire dynamique du modèle

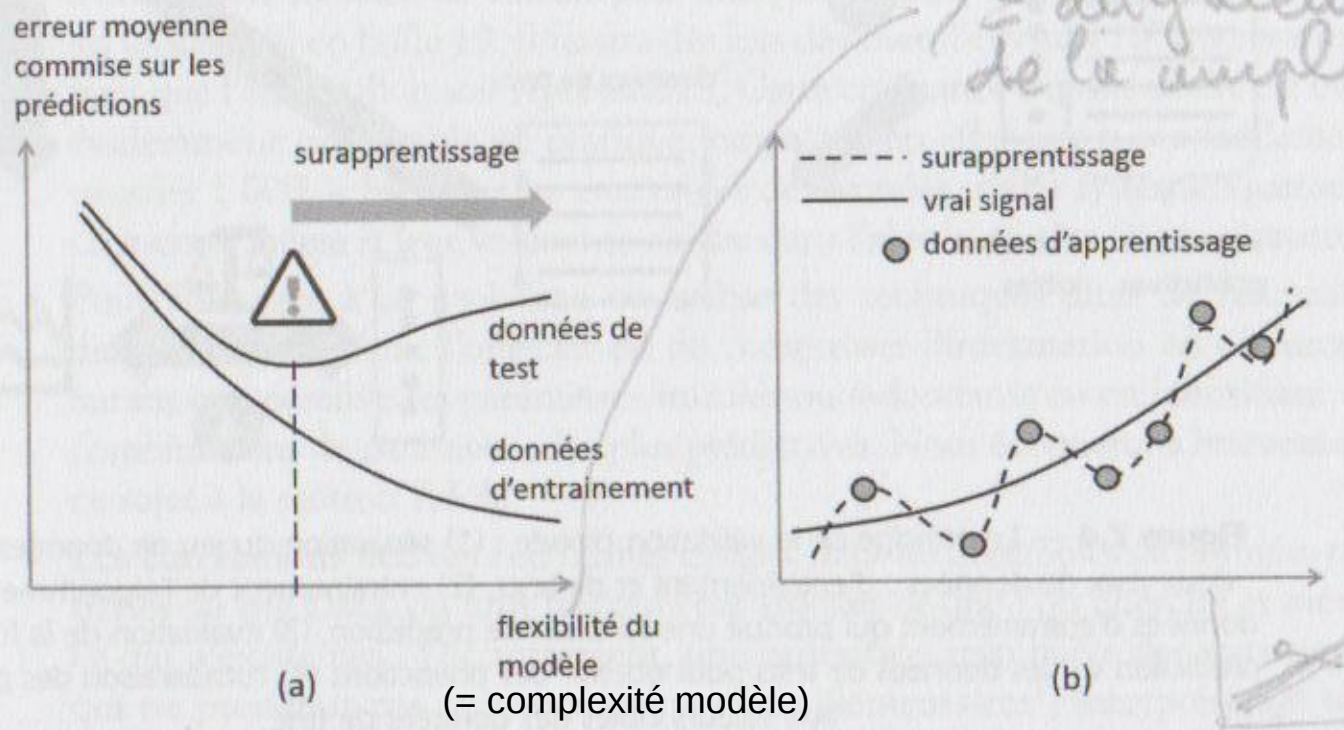
Remarque: on peut montrer qu'un site de vente en ligne a une mémoire dynamique

Estimation de la complexité du modèle

On cherche la structure de modèle de complexité *minimale* pouvant décrire les phénomènes qui nous intéressent.

On utilise le principe de *parcimonie (comme en identification de systèmes)*
-Un modèle à peu de paramètres est plus facile à estimer
-Un modèle trop compliqué est en général “non robuste” par rapport au changement d’environnement
(“risque de surapprentissage” – voir Big Data pg 113)

Estimation de la complexité du modèle dans le Data Mining



(= complexité modèle)

Figure 7.3 — Le surapprentissage : (a) montre comment évoluent les erreurs commises sur les données de tests et celles sur les données d'apprentissage à mesure que la flexibilité du modèle croît. Dans un premier temps, les deux erreurs diminuent de concert. Puis, lorsque s'amorce le surapprentissage, l'erreur sur les données d'apprentissage continue à diminuer alors que celle sur les données de test augmente. (b) illustre un exemple ou une courbe passe par tous les points d'apprentissage mais ne réussit pas à capter la tendance générale.

Big Data pg 113

Estimation des paramètres (1)

Objectif (sortie mesurée): y Sortie prédite ;$\hat{y}$

Erreur de prédiction : e (ou ε) = $y - \hat{y}$

Structure des modèles: linéaires, non linéaires, réseau de neurones, …..

Structure la plus utilisée*: la régression linéaire*

$$y = a_1 x_1 + a_2 x_2 + ... + a_n x_n = \mathrm{a}^T x = \theta^T \phi$$

$$a_1, a_2 .... a_n = paramètres; \quad x_1, x_2 ... x_n = variables\,(observations, mesures)$$

$$\mathrm{a}^T = \theta^T = [a_1, a_2 .... a_n]; \quad x^T = \phi^T = [x_1, x_2 ... x_n]$$

a = θ = vecteur des paramètres *x* = ϕ = vecteur des attributs (observations)

On mesure $y, x_1, x_2, ... x_n$ **. On ne connait pas** $a_1, a_2, ... a_n$

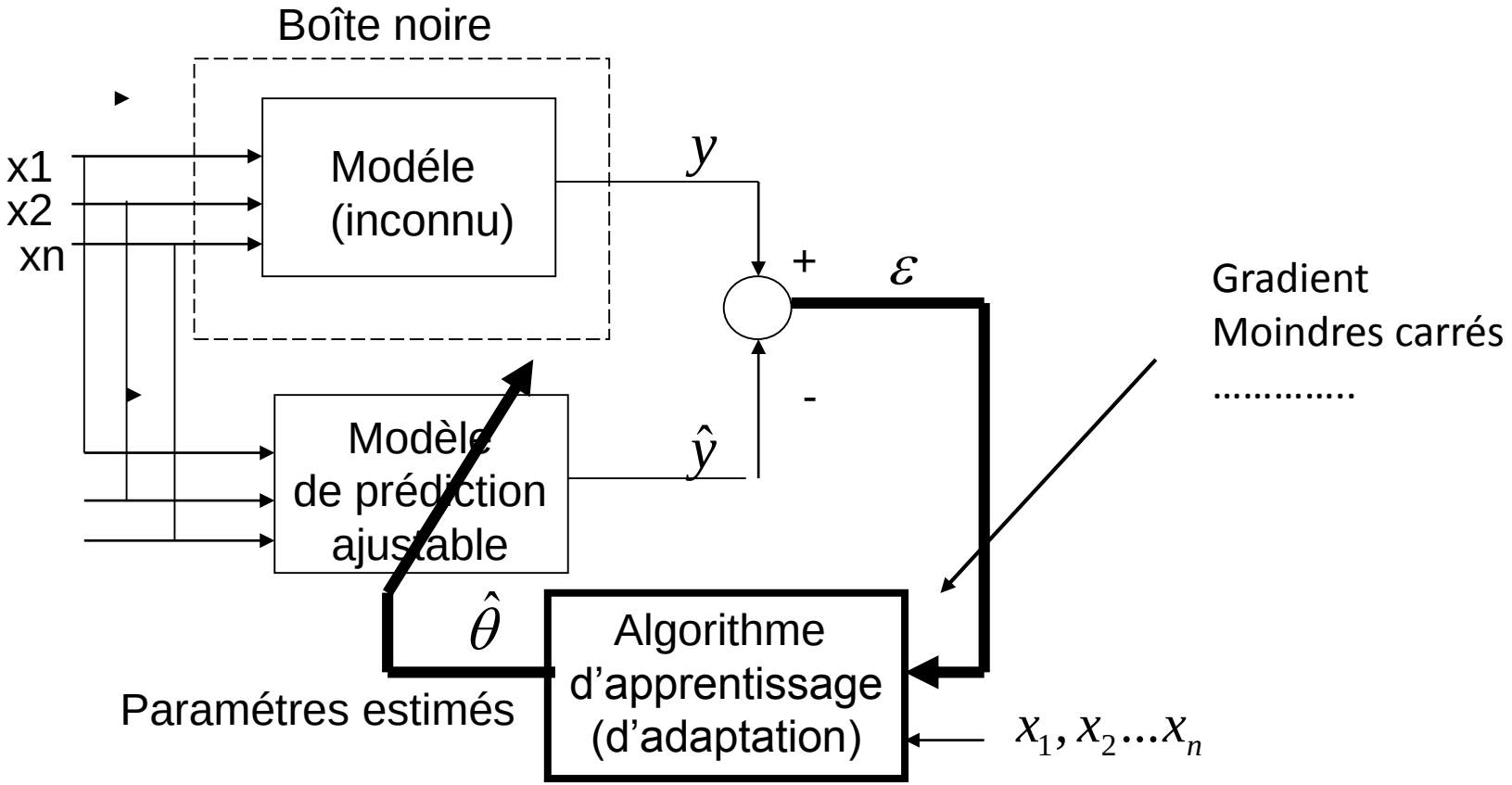
Modèle de prédiction (à trouver les bons paramètres)

$$\boxed{\hat{y} = \hat{a}_1 x_1 + \hat{a}_2 x_2 + ... + \hat{a}_n x_n = \hat{\mathrm{a}}^T x = \hat{\theta}^T \phi}$$

Estimation des paramètres (2)

Objectif de l'estimation des paramètres (apprentissage des paramètres):

A partir d'un jeux de données (horizon N)(= données d'entrainement)
trouver $\hat{a}_1, \hat{a}_2 .... \hat{a}_n$ qui minimisent un critère en termes d'erreurs de prédiction (en bref qui "minimise" l'erreur de prédiction)



Estimation des paramètres (3)

Préliminaire: on ne peut minimiser que des quantités positives

Types de critères (les plus utilisés)

$$\min \varepsilon^2(t+1)$$ Algorithmes du gradient (plusieurs versions)

$$\min \sum_{t=0}^{N} \varepsilon^2(t+1)$$ Critère et algorithme des “moindres carrés”

Estimation des paramètres (4)

Types d'algorithmes

Batch (off line); traitement en un coup d'un ensemble de N jeux de données
Recursive (on line): traitement itératif de chaque jeux de de données de 1 à N

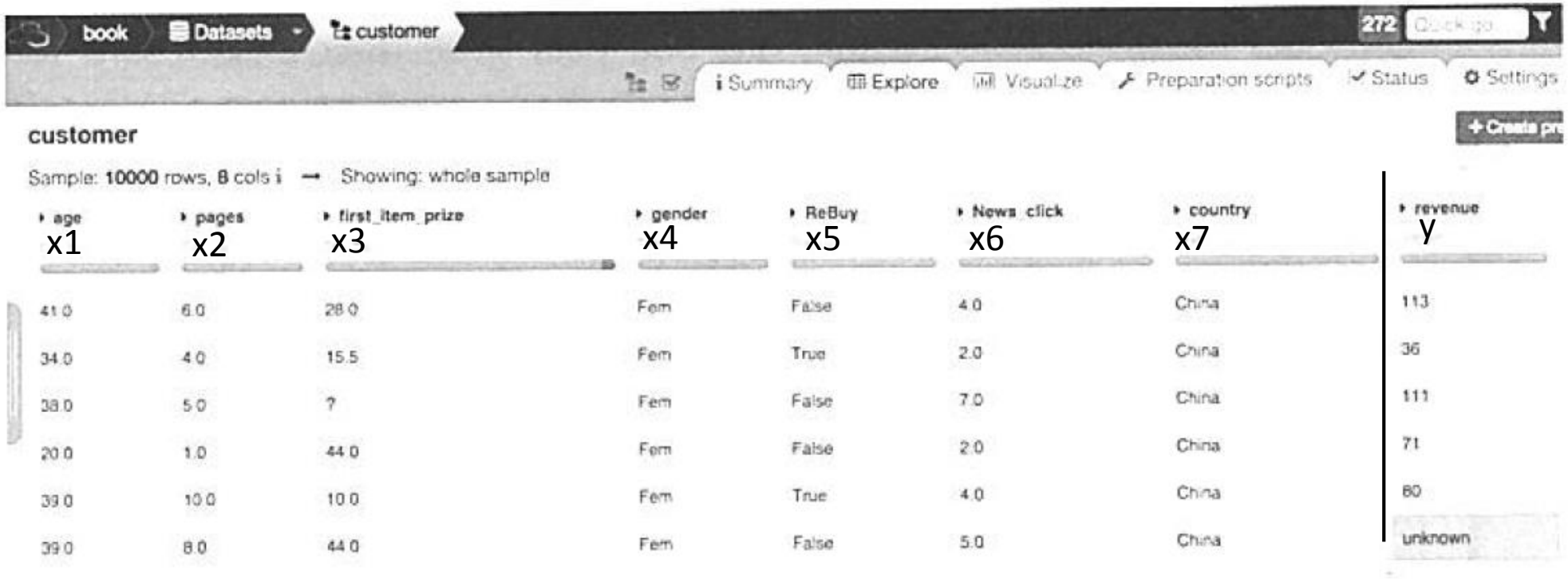
Seulement les algorithmes récursifs font de l'apprentissage (au sens strict)

Rem: les algos récursifs peuvent aussi bien être utilisés en *on line* qu'en *off line*

Rem: *les algorithmes fonctionnent correctement s'il le critère à minimiser a un seul minimum* (on dit que c'est un *critère convexe*)

Rem: la *convexité* dépend (mais pas seulement) de la structure des données

Exemple de modèle prédictif type régression linéaire



age (x1)	pages (x2)	first_item_prize (x3)	gender (x4)	ReBuy (x5)	News_click (x6)	country (x7)	revenue (y)
41.0	6.0	28.0	Fem	False	4.0	China	113
34.0	4.0	15.5	Fem	True	2.0	China	36
38.0	5.0	?	Fem	False	7.0	China	111
20.0	1.0	44.0	Fem	False	2.0	China	71
39.0	10.0	10.0	Fem	True	4.0	China	80
39.0	8.0	44.0	Fem	False	5.0	China	unknown

Figure 7.13 — Un fichier clients dont les variables prédictives sont `age` (l'âge), `pages` (le nombre de pages du site visité), `first_item_prize` (le prix du premier article acheté), `gender` (masculin ou féminin), `ReBuy` (le client a-t-il acheté le premier article plus d'une fois ?), `News_click` (nombre de fois où le client a cliqué sur une campagne de publicité du site), `country` (le pays dont provient l'adresse IP), `revenue` (le revenu généré par le client sur le site). Le nom du dataset est `customer`. © Dataiku.

La sortie (la variable cible)

Big Data pg 135

Exemple de modèle prédictif type régression linéaire

Venons-en maintenant à la régression linéaire. Celle-ci est effectuée après une normalisation initiale des données, ceci pour pouvoir légitimement comparer entre eux les coefficients de la régression.

Prédiction

Paramétres $\hat{a}_i$ de la regression linéaire

```
Revenue = constante + 0.002 * GDP_inhab
    + 1.1 * first_item_prize + 0.09 * age + 0.09 * pages
    + 13.6 * News_click - 0.00000002 * population
    - 18.4 * ReBuy_value_False - 8.9 * gender_value_Fem
```

On peut commencer par regarder les signes des différents coefficients. S'ils sont positifs cela traduit le fait qu'une augmentation de la variable associée se traduit par une augmentation du revenu et inversement. On voit par exemple qu'un client qui n'a pas racheté une deuxième fois son premier article (`ReBuy_value_False`) génèrera moins de revenu, tandis qu'un niveau de prix élevé pour ce premier article (`first_item_prize`) est un bon indicateur d'un client rémunérateur. Rien de surprenant à cela, mais il est toujours bon de vérifier la cohérence de l'algorithme.

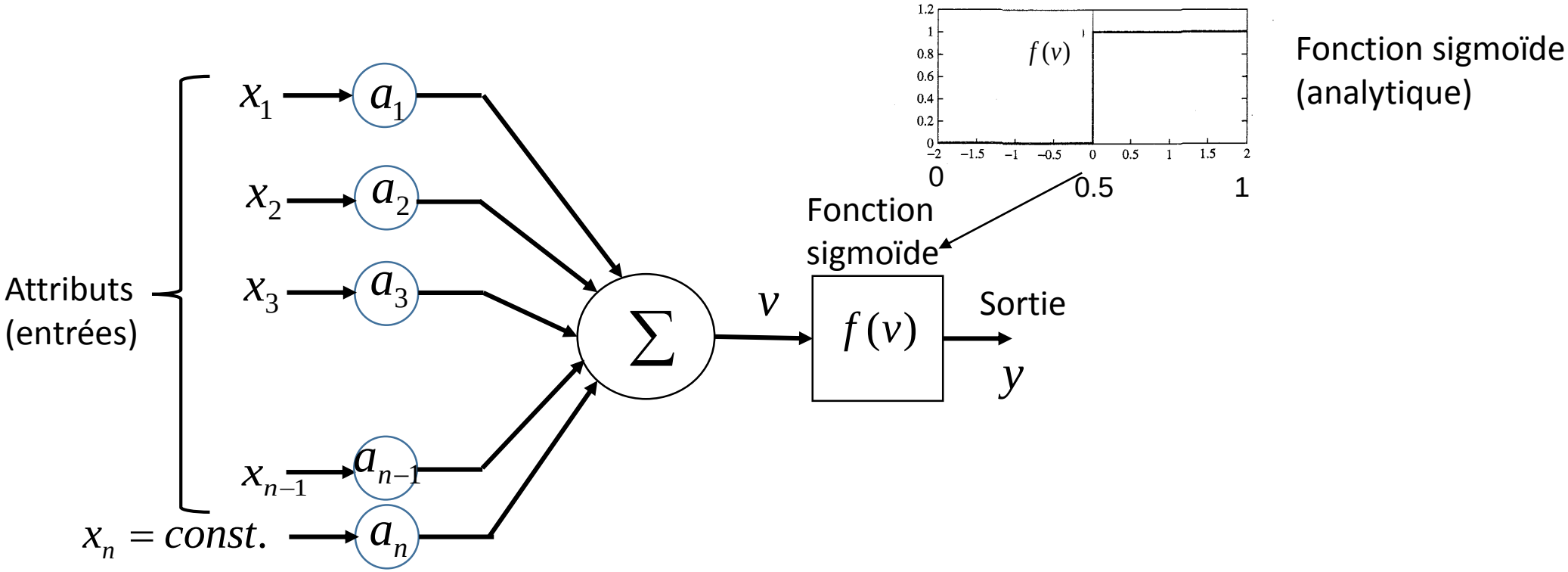
Big Data pg 144

Le cas des sorties binaires (0 ou 1)

Exemple: en fonction des variables prédictives sur un site de e-commerce on veut prédire s'il y aura achat (1) ou non (0)

Il faut définir des nouveaux types de modèles de prédiction dont les entrées (observations) sont continues (toutes valeurs possibles entre 0 et 10000) et la sortie est binaire (0 ou 1)

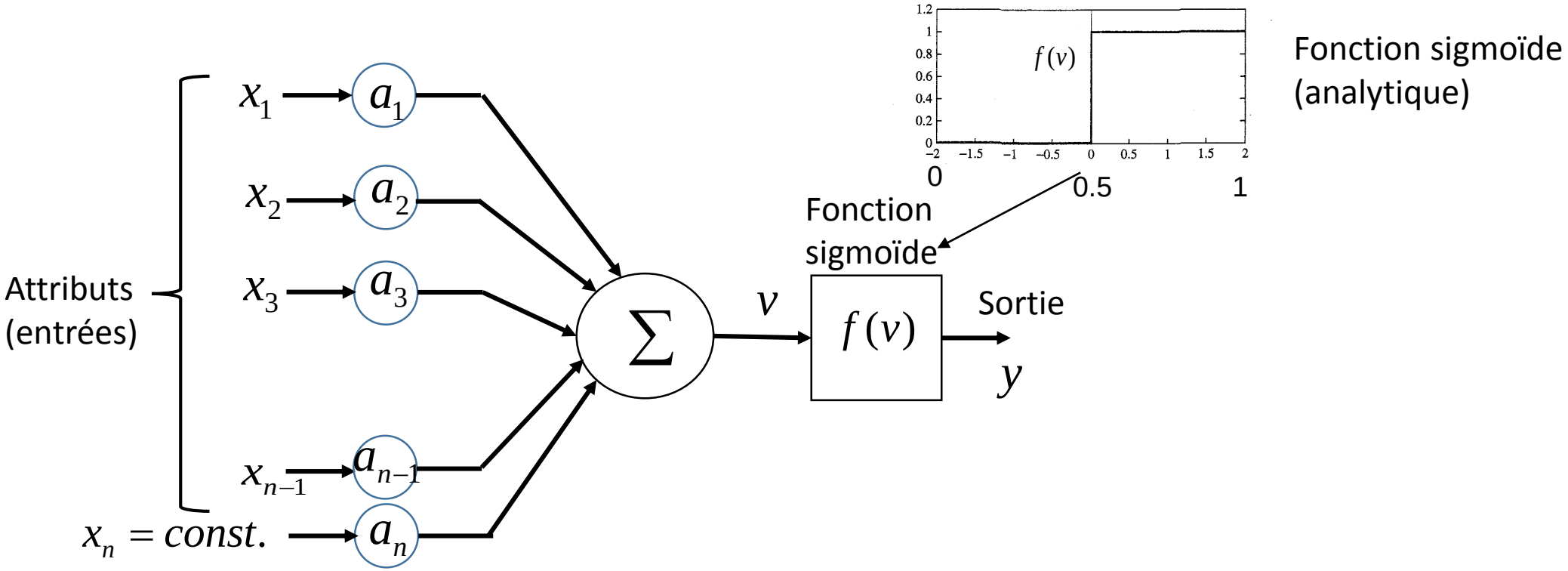
Connaissez vous cette structure ?



$$\theta^T = [a_1, a_2, a_3 .. a_n]; \phi^T = [x_1, x_2, x_3 .. x_n] \quad \Rightarrow \quad v = \theta^T \phi = \sum_1^n a_i x_i$$

C'est une régression linéaire dont la sortie (intermédiaire) v passe par une fonction nonlinéaire spéciale (sigmoïde)

Le module “Neurone” (perceptron)

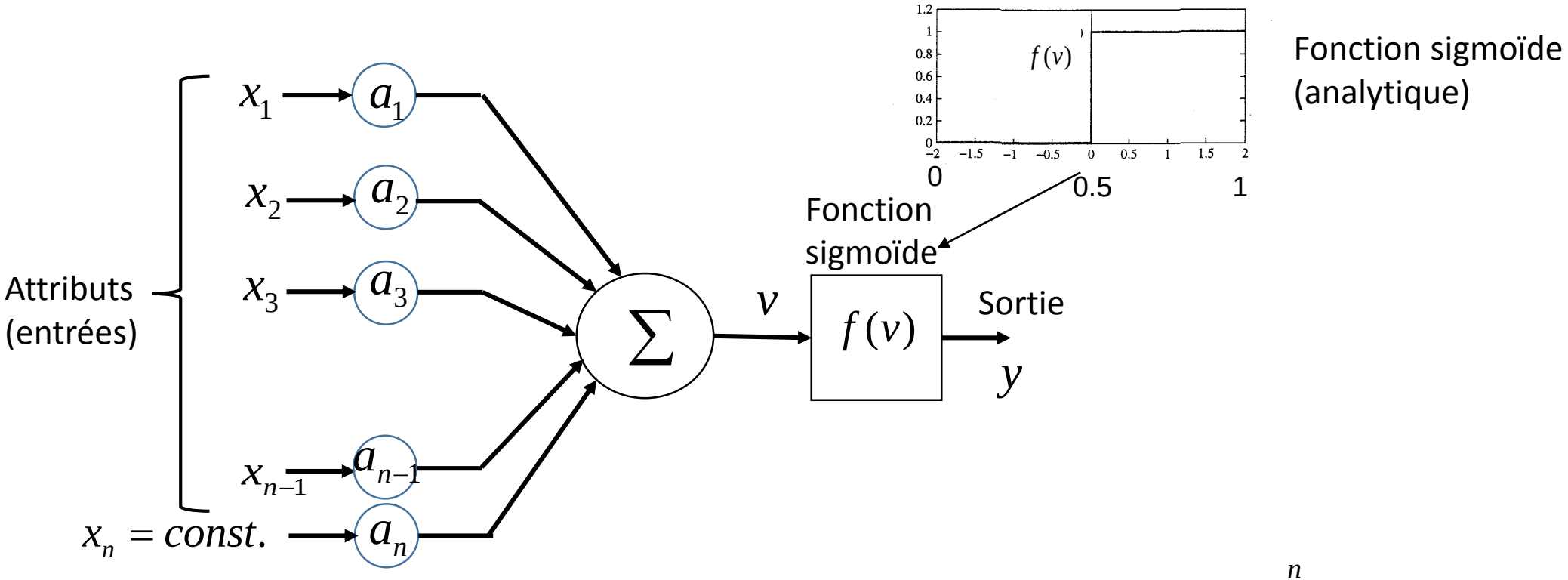


(Haykin,1998)

$$\theta^T = [a_1, a_2, a_3 .. a_n]; \phi^T = [x_1, x_2, x_3 .. x_n] \quad \Rightarrow \quad v = \theta^T \phi = \sum_1^n a_i x_i$$

C’est une régression linéaire dont la sortie (intermédiaire) v passe par une fonction nonlinéaire spéciale (sigmoïde)

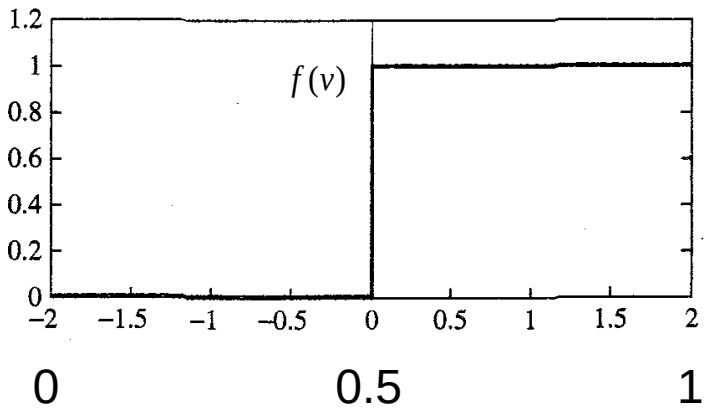
La Régression Logistique

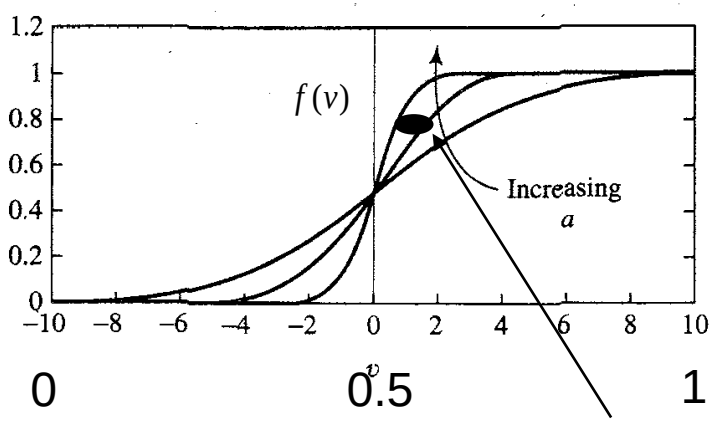


$$\theta^T = [a_1, a_2, a_3 .. a_n]; \phi^T = [x_1, x_2, x_3 .. x_n] \quad \Rightarrow \quad v = \theta^T \phi = \sum_1^n a_i x_i$$

C'est une régression linéaire dont la sortie (intermédiaire) v passe par une fonction nonlinéaire spéciale (sigmoïde)

Fonctions sigmoïdes





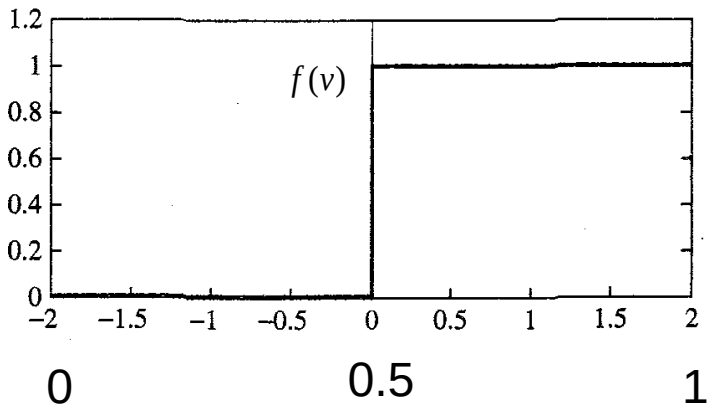
$$f_{sig}(x) = \frac{1}{1 + e^{-x}}$$

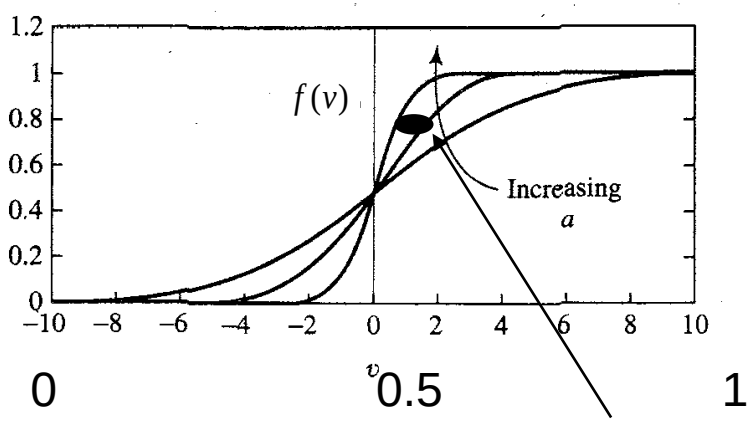
Il y a 80% de chances qu'il y ait un achat.

Permet d'associer un coefficient de confiance ou de probabilité.

Le résultat dépend du choix de la courbe sigmoïde (très pifométrique)

Fonctions sigmoïdes





$$f_{sig}(x) = \frac{1}{1 + e^{-x}}$$

Il y a 80% de chances qu'il y ait un achat.

Permet d'associer un coefficient de confiance ou de probabilité.

Le résultat dépend du choix de la courbe sigmoïde (très pifométrique)

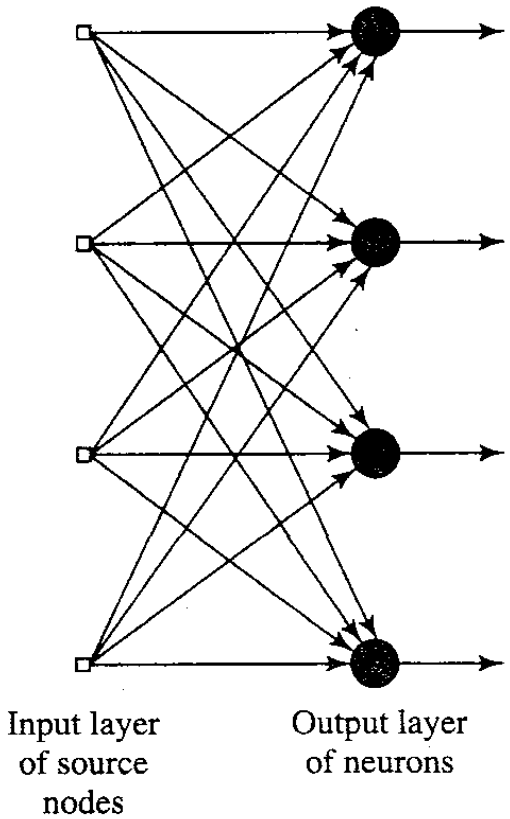
<u>Algorithmes d'apprentissage pour réseaux de neurones</u>

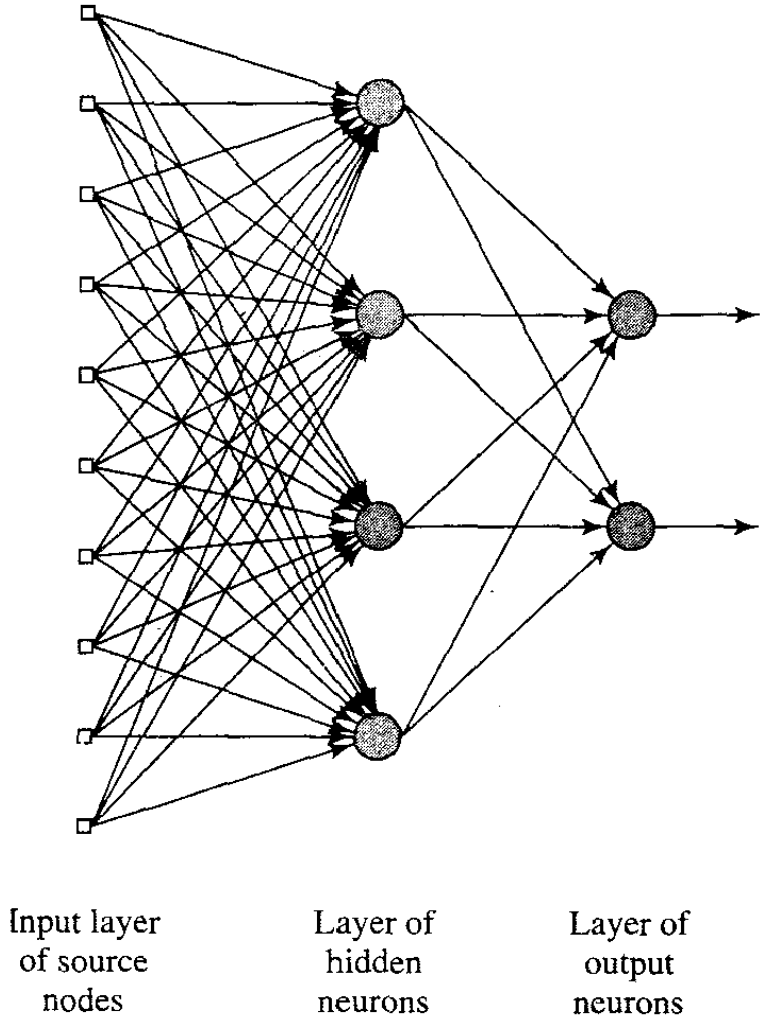
Ils sont similaires aux algorithmes pour les régressions linéaires
(c'est les composantes du vecteur des attributs qui changent)

Réseaux de Neurones

Une option pour la modélisation non-linéaire (non-structurée) et les arbres de régression

(Haykin ,1998)





Rem.: Les poids (paramètres) ne sont pas représentés

Apprentissage supervisé renforcé (avec excitations externes)

Dans certaines situations on dispose naturellement d'un *entrée* qu'on peut manipuler

Exemples:

Commande d'un moteur (tension)
Politique de promotion (rabais et leurs influence sur les ventes)

Apprentissage supervisé renforcé (avec excitations externes)

Dans certaines situations on dispose naturellement d'un *entrée* qu'on peut manipuler

Exemples:

Commande d'un moteur (tension)
Politique de promotion (rabais et leurs influence sur les ventes)

<u>Apprentissage en temps réel (adaptation des modèles)</u>

Les algorithmes d'apprentissage récursifs peuvent être utilisés en temps réel pour améliorer le ***modèle de prédiction*** en utilisant les nouvelles données disponibles

Modèle de prédiction type « arbre de décision »

Morgan, Sonquist (1963), Breiman (1984) : Références classiques pour les arbres de décision

- C'est sans aucun doute un objet de base pour les informaticiens qui l'ont décliné de multiple façons.
- Même si son utilisation en Data Mining est discutable, ça a donné lieu à un très grand nombre de travaux pour pouvoir faire du Data Mining avec des ***arbres de décision***.

Les arbres de décision sont traités par exemple dans:
Livre de Mitchell (Machine Learning, 1997) Sujet majeur
Livre de Poole, Mackworth (Artificial Intelligence, 2010) Sujet majeur
Livre de Haykin (Neural Network ,1999)

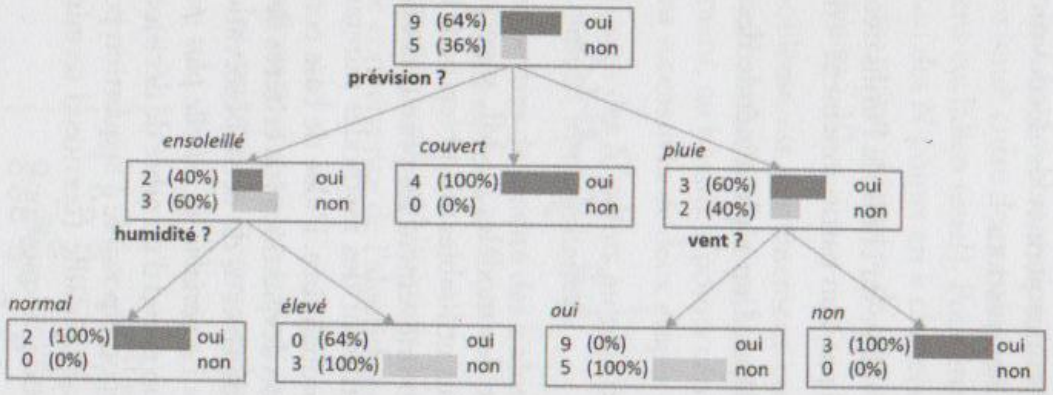
Modèle de prédiction type « arbre de décision »

Variables prédictives (observations)

Variable cible (dernière colonne *on joue* (oui-1) ou *on ne joue pas* (non-0)

	prévision	température	humidité	vent	oui
1	couvert	froid	normal	oui	oui
2	couvert	chaud	élevé	non	oui
3	couvert	chaud	normal	non	oui
4	couvert	moyen	élevé	oui	oui
5	pluvieux	froid	normal	oui	non
6	pluvieux	moyen	élevé	oui	non
7	pluvieux	froid	normal	non	oui
8	pluvieux	moyen	élevé	non	oui
9	pluvieux	moyen	normal	non	oui
10	ensoleillé	chaud	élevé	non	non
11	ensoleillé	chaud	élevé	oui	non
12	ensoleillé	moyen	élevé	non	non
13	ensoleillé	froid	normal	non	oui
14	ensoleillé	moyen	normal	oui	oui

(a)



(b)

Big Data pg128 (Quinlan ,1993)

Ce n'est pas une variable significative

Figure 7.11 — Construction d'un arbre de décision :
(a) l'ensemble d'apprentissage contient 14 observations décrites par quatre variables prédictives et une variable cible que l'on veut prédire.
(b) la succession de questions posées homogénéise progressivement la répartition des observations comme le montrent les histogrammes.

Modèle de prédiction type « arbre de décision »

Numéro	Ensoleillement	Température (°F)	Humidité (%)	Vent	Jouer
1	soleil	75	normale 70	oui	oui
2	soleil	80	élevé 90	oui	non
3	soleil	chaud 85	élevé 85	non	non
4	soleil	72	élevé 95	non	non
5	soleil	69	normale 70	non	oui
6	couvert	72	élevé 90	oui	oui
7	couvert	83	élevé 78	non	oui
8	couvert	64	normale 65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	froid 65	70	oui	non
12	pluie	moyen 75	élevé 80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui

Tableau 1 : Données "weather" (Quinlan, 1993)
(Rakotomalala)

- Pour appliquer la procédure, les variables prédictives doivent être « logiques » (à 2 ou plusieurs nivaux)
- On doit donc « fuzzifier » les données numériques
- La variable cible doit être logique
- Les méthodes de sélection des variables de segmentation par ordre d'importance sont assez claires
- Des nombreux logiciels sont disponibles
- Grand risque de « surapprentissage »
- La « profondeur » de l'arbre peut être associée à l'ordre du modèle (complexité)
- Il faut faire de *l'élagage*
- Difficile à interpréter quand la « profondeur » dépasse 5

Modèle de prédiction type « arbre de décision »

Numéro	Ensoleillement	Température (°F)	Humidité (%)	Vent	Jouer
1	soleil	75	normale 70	oui	oui
2	soleil	80	élevé 90	oui	non
3	soleil	chaud 85	élevé 85	non	non
4	soleil	72	élevé 95	non	non
5	soleil	69	normale 70	non	oui
6	couvert	72	élevé 90	oui	oui
7	couvert	83	élevé 78	non	oui
8	couvert	64	normale 65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	froid 65	70	oui	non
12	pluie	moyen 75	élevé 80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui

Tableau 1 : Données "weather" (Quinlan, 1993)
(Rakotomalala)

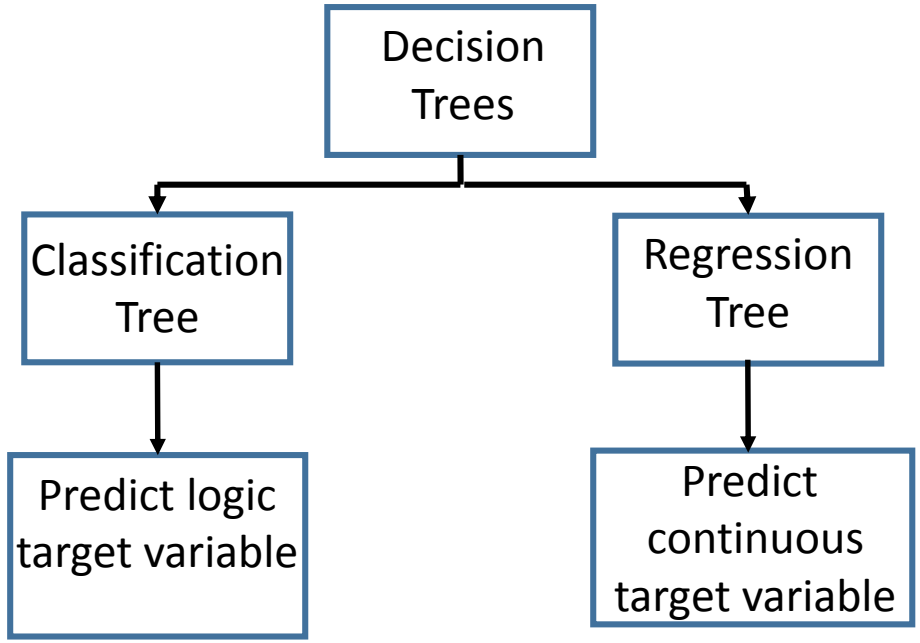
- Pour appliquer la procédure, les variables prédictives doivent être « logiques » (à 2 ou plusieurs nivaux)
- On doit donc « fuzzifier » les données numériques
- La variable cible doit être logique
- Les méthodes de sélection des variables de segmentation par ordre d'importance sont assez claires
- Des nombreux logiciels sont disponibles
- Grand risque de « surapprentissage »
- La « profondeur » de l'arbre peut être associée à l'ordre du modèle (complexité)
- Il faut faire de *l'élagage*
- Difficile à interpréter quand la « profondeur » dépasse 5

Ce n'est pas une variable logique

Comment on fait pour prédire le « chiffre d'affaire » d'un site de e-commerce?

On ne prend pas en compte la « mémoire dynamique »
(j'ai joué hier et avant-hier, vais-je jouer aujourd'hui si les conditions sont bonnes?)

CART – Classification and Regression Trees (Standard)



Arbre de régression :
Les feuilles sont des « régressions ». Les variables cibles des feuilles deviennent des variables prédictives, qui se combinent dans une nouvelle régression créant une nouvelle variable cible

- On peut avoir des combinaisons de régression linéaires et logistiques
- Les réseaux de neurones sont un bon exemple de réalisation de tels modèles

Comparaison régression/régression logistique - arbre de décision

- Malgré les nombreux benchmarks et les BD disponibles, on ne peut pas tirer actuellement une conclusion sur: quelle est la meilleure approche?
- En règle générale les résultats sont comparables (mais il faut beaucoup de doigté pour les arbres).
- En l'état actuel de la méthodologie, il semblerait qu'il y a aussi un aspect « problem dependent ».
- On recommande d'utiliser plusieurs approches pour le même problème (ça fait vendre des logiciels!).
- En l'état actuel de la méthodologie, le travail en amont pèse beaucoup sur les résultats.
- Les techniques utilisés en « régression linéaire » dans le Data Mining et Machine Learning sont d'un « autre âge » (*on doit pouvoir faire mieux!*).

Comparaison régression/régression logistique - arbre de décision

- Malgré les nombreux benchmarks et les BD disponibles, on ne peut pas tirer actuellement une conclusion sur: quelle est la meilleure approche?
- En régle générale les résultats sont comparables (mais il faut beaucoup de doigté pour les arbres).
- En l'état actuel de la méthodologie, il semblerait qu'il y a aussi un aspect « problem dependent ».
- On recommande d'utiliser plusieurs approches pour le même problème (ça fait vendre des logiciels!).
- En l'état actuel de la méthodologie, le travail en amont pèse beaucoup sur les résultats.
- Les techniques utilisés en « régression linéaire » dans le Data Mining et Machine Learning sont d'un « autre âge » (*on doit pouvoir faire mieux!).*

Il y a une grande marge de progression méthodologique pour les techniques de régression

Quelques pistes pour l'amélioration des techniques de régression en Data Mining

Prise en compte de l'expérience acquise en identification des systèmes dynamiques!

- Amélioration du traitement de la composante continue des modèles (biais sur la variable cible)
- Analyse de la « richesse » des données (concept « excitation persistante »)
- Analyse de la variabilité des paramètres identifiés (intervalle de confiance)
- Nouvelle techniques pour déterminer les attributs significatifs et leur nombre (estimation d'ordre)
- Techniques de validation sur la base de données d'apprentissage (tests statistiques)
- Utilisation d'autre algorithmes à la place de l'algorithme des moindres carrés.
- Prise en compte de la mémoire dynamique
- Adaptation des modèles de prédiction en temps réel après déploiement

Apprentissage supervisé actif (renforcé) en Data Mining – *un sujet de recherche*

- Les politiques de promotion sont des entrées exogènes (apprentissage actif)
- Les politiques de promotion ont un **coût**
- Les politiques de promotion ont comme objectif à terme: d'augmenter le CA, recruter des nouveaux clients, etc
- Elles doivent permettre de mettre en place des offres ciblées à partir des modèles de prédiction

Solution actuelle (Provost, Fawcett 2013)

- Promotions aléatoires (comme une SBPA)
- Entrée riche permettant une bonne identification des modèles de prédiction
- Il résulte un coût important à court terme et un temps de retour sur investissement long
- Il y a comme même une « success story »

Solution optimale (*à développer*)

- C'est un problème de « commande duale » (Feldbaum 1961)
- Il faut chercher une approximation de la commande duale permettant en temps réel d'ajuster la politique de promotion tout en minimisant le coût.

System Identification and Adaptation Techniques for Data Mining and Machine Learning in Big Data

Doctoral School: EEATS Grenoble
Host Research Laboratory: Control Dept, GIPSA- LAB, UMR CNRS 5216, Grenoble

PhD advisers: I.D. Landau Emeritus Research Director CNRS
Mails : [ioan-dore.landau]@gipsa-lab.grenoble-inp.fr, Phone: 0676462451 (IL)
External PHD Advisers :

Key words:
System identification, Adaptive techniques, Data Mining, Machine Learning, Big Data, Benchmarking

The Context

The tremendous development of large data bases has led to a new paradigm in data science: extraction of predictive models from the huge amount of data available. This is the current major challenge in data mining. Tools for manipulating big data have been developed (like MapReduce, Hadoop). Concepts have emerged which are not surprisingly very close to the concepts used in system identification from data (basically the same once the vocabulary is decoded). However in terms of algorithms there is a lot to be done from the perspective of system identification and adaptive control. The current techniques in data mining use learning techniques (developed under the term "machine learning") which largely ignores the research done in learning algorithms in the field of control since the end of the sixties (Saridis , Tsypkin ..).

An analysis of the current methodology in data mining for extracting prediction models show at the level of concepts an extremely strong similarity with the concepts and methodology developed for system identification from data, field which establishes its base back in the seventies. The current technology in data mining ignores most of the work done in the field of system identification. Just to mention the problems of dc components, of order (complexity) estimation (called "overfitting" in data mining), of the bias, of statistical validation etc. These problems are currently approached by ad-hoc solutions ignoring the fact that these problems have been deeply addressed in the system identification field. While the interest of adding an excitation (called "supervised learning") has been recognized, the solutions used in system identification are ignored.

A similar situation occurs in "deployment" (real-time use) of the prediction models which requires on line updating. The important development in "adaptive algorithms" for real time identification and adaptive control is ignored.

The use of the so called "logistic regression" based on the use of the sigmoid function to generate a logic variable from a continuous variable, brings immediately in view the "neural networks" which were deeply studied in the field of control (Haikin, Narendra, Soykete) .

Research subject and expected results

1) Reformulation of the data mining for extraction of prediction model in big data as an system identification problem
2) Development of a new procedure for data mining bringing in new algorithms for
   - Processing the off-set (steady state component)
   - Complexity estimation
   - Parameter estimation
   - Model validation
   - Active data mining (using external excitation)
3) Comparative evaluation using currently available data bases (benchmarking) like http://downloads.dataiku.com/training/livre-big-data/

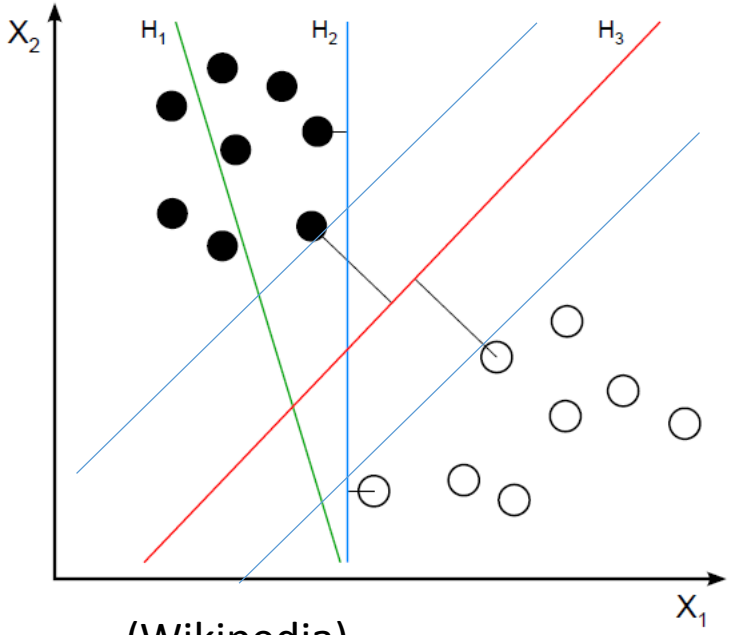
4) Improving current so called "machine learning" techniques by using the advanced parameter adaptation techniques for deployment of predictive models.

REFERENCES:
I.D. Landau "Digital Control Systems – design, identification and implementation", Springer , 2005, London
I.D. Landau et al. "Adaptive Control", Second edition, Springer 2011
T.E. Mitchell "Machine Learning" , Mc Graw Hill, 1997
S. Haikin "Neural Networks – a comprehensive foundation", Pearson Prentice Hall, 1999-2005,
F. Provost, T. Fawcett "Data Science for Business", O'Reylly, New York, 2013
P. Lemberger, M. Batty, M. Morel, J.L. Raffaelli "Big Data et Machine Learning", Dunod , Paris 2015
J.A.K. Soykete, J.P.I. Vandevalle, B. de Moor "Artificial Neural Networks for modelling and control of non-linear systems".
L. Ljung, H. Hjalmarsson, H Ohlsson "Four encounters with system identification" European J. of Control, Volume 17, Issues 5-6, September 2011, Pages 449-471
P. de Larminat « Changement Climatique », ISTE, 2014
J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle (Eds.) *Advances in Learning Theory: Methods, Models and Applications,* NATO Science Series III: Computer & Systems Sciences, Volume 190, IOS Press Amsterdam, 2003, 436pp. (ISBN: 1 58603 341 7).
K.S. Narendra, K. Parthasarati "Identification and control of dynamical systems using neural networks" IEEE Trans on Neural Networks, Vol 1, No. 1 pp 4-27
P. van Overschee, B. de Moor, "Subspace identification for linear systems". Kluwer Academic Publisher,
Y.Z. Tsypkin, "Adaptation and learning in Automatic Systems" Academic Press 1971

GRANTS: No specific grant attached to this subject. The candidate has to postulate for the grants offered in principle by GIPSA-LAB or by other organisms.

Expected background: the candidate should have a good background either in control engineering or in computer engineering.

Support Vector Machines (*Machines à Vecteur de Support*)



(Wikipedia)

Une autre « imposture »!

Algorithme de discrimination à marges maximales

Référence originale: Vapnik, Chervonenkis (1963)

Plusieurs algorithmes:

- Moindres carrés (Bart de Moor, 2003)
- Réseaux de neurones (Haykin, 1998)
- ……..

« Not only is the name opaque, but the method often is imbued with the sort of magic that derives from perceived effectiveness without understanding. » (Provost, Fawcett,2013)

En résumé

- L'Identification des Systèmes dynamiques c'est du Data Mining

- Les méthodologie utilisée dans le Data Mining (en général) correspond à la méthodologie utilisée en Identification des Systèmes il y a 40 ans!

- Critique des techniques actuelles utilisées en Data Mining:
  - On ne s'assure pas qu'on a extrait le « meilleur modèle » sur la base d'apprentissage avant de passer à la validation sur la base de test (absence de tests statistiques)
  - Les techniques estimation de la complexité sont dépassées
  - Absence de la prise en compte du « bruit » dans le problème d'estimation des paramètres
  - Utilisation exclusive de l'algorithme de moindres carrés qui donne des estimations biaisées en présence de bruit.

Où on va ?

Où on va ?

DDDM

Data-Driven Decision Making

Où on va ?

DDDM

Data-Driven Decision Making

En Automatique ont fait du DDDM: on identifie un modèle à partir des données, on calcule un régulateur à partir de ce modèle et on applique une commande fournie par le régulateur (qui dépend des données).

Où on va ?

DDDM

Data-Driven Decision Making

Evolution de la gouvernance des entreprises

Entrepreneurs → Gestionnaires → Commerciaux → Financiers → Informaticiens

REFERENCES:

I.D. Landau "Digital Control Systems – design, identification and implementation", Springer , 2005, London

I.D. Landau et al. "Adaptive Control", Second edition, Springer 2011

T.E. Mitchell "Machine Learning" , Mc Graw Hill, 1997

S. Haikin "Neural Networks – a comprehensive foundation", Pearson Prentice Hall, 1999-2005,

F. Provost, T. Fawcett "Data Science for Business", O'Reylly, New York, 2013

P. Lemberger, M. Batty, M. Morel, J.L. Raffaelli "Big Data et Machine Learning", Dunod , Paris 2015

J.A.K. Soykete, J.P.I. Vandevalle, B. de Moor "Artificial Neural Networks for modelling and control of non-linear systems".

L. Ljung, H. Hjalmarsson, H Ohlsson "Four encounters with system identification" European J. of Control, Volume 17, Issues 5-6, September 2011, Pages 449-471

P. de Larminat « Changement Climatique », ISTE, 2014

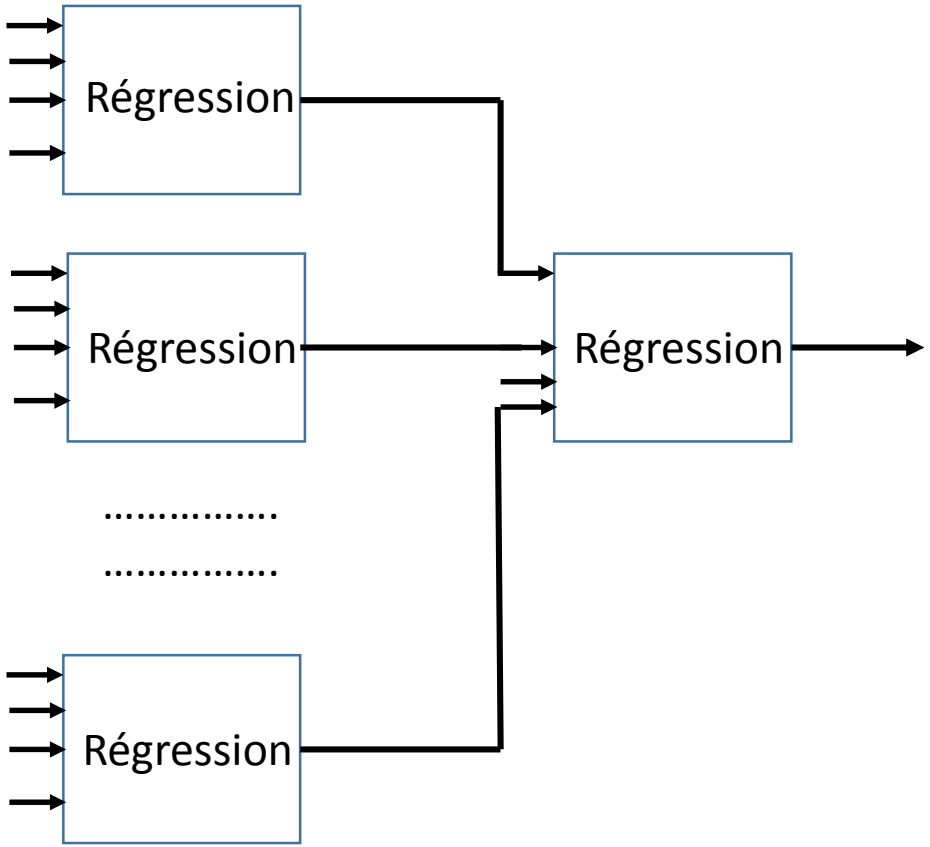
J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, J. Vandewalle (Eds.) *Advances in Learning Theory: Methods, Models and Applications,* NATO Science Series III: Computer & Systems Sciences, Volume 190, IOS Press Amsterdam, 2003, 436pp. (ISBN: 1 58603 341 7).

K.S. Narendra, K. Parthasarati "Identification and control of dynamical systems using neural networks" IEEE Trans on Neural Networks, Vol 1, No. 1 pp 4-27

P. van Overschee, B. de Moor, "Subspace identification for linear systems". Kluwer Academic Publisher,

Y.Z. Tsypkin, "Adaptation and learning in Automatic Systems" Academic Press 1971

Arbre de régression



Peut être inspiré par les « réseaux de neurones »?