Une histoire des idées en **Apprentissage Artificiel**

Antoine Cornuéjols

MMIP, AgroParisTech, Paris

EPAT - 3 mai 2010

Plan

1. Les pionniers
2. Les années 60-70
3. Avènement de l'analyse statistique de l'apprentissage
4. Et maintenant : les grandes questions ?
5. Conclusion

Plan

1. Les pionniers
   - De nouvelles bases conceptuelles
   - Des conférences marquantes
   - Caractéristiques de la première approche
2. Les années 60-70
3. Avènement de l'analyse statistique de l'apprentissage
4. Et maintenant : les grandes questions ?
5. Conclusion

Informatique, cybernétique, intelligence artificielle

Naissance de **quelque chose de neuf**

Machines et calcul

- Traitement de l'information
- Manipulations discrètes de symboles discrets
- programme $\equiv$ données ($\Rightarrow$ auto-modifiables)

Versus

- Gestält theory (analogique et global)
- Behaviorisme

Informatique, cybernétique, intelligence artificielle

Naissance de quelque chose de neuf

Très **interdisciplinaire** : *biologie, psychologie, ingénierie, mathématique, informatique*

Trois nouveautés conceptuelles importantes :

1. La **simulation** :
   - on simule pour comprendre, et on étudie des simulations
2. La **rétro-action** :
   - la physique est un idéal de science mais l'intelligence artificielle n'est pas de la physique
3. Le **codage de l'information** :
   - idéalisation et matérialisation : penser c'est manipuler des représentations

Informatique, cybernétique, intelligence artificielle

Naissance de quelque chose de neuf

Questions : Mémoire, adaptation, apprentissage, raisonnement, représentations

- Qu'est-ce que c'est ?
- Comment ça fonctionne ?

Idées nouvelles :

- Cybernétique, rétro-action, théorie du contrôle
- Rationalité et théorie des jeux
- (Shannon et Turing, 1943)
  Robustesse au bruit (colorblue cryptage ou transmission de l'information)
  $\Rightarrow$ réflexions sur la généralisation et l'induction

Informatique, cybernétique, intelligence artificielle

Naissance de quelque chose de neuf

Tout est TRÈS difficile à programmer : *l'apprentissage est essentiel*

- Premières approches fondées sur l'apprentissage par renforcement
- Développement de l'idée d'auto-organisation
- Question jugée centrale : l'apprentissage de **représentations internes**

- Assemblées de neurones et règle d'apprentissage [Hebb, 1949]

Extraordinaires conférences

Naissance de quelque chose de neuf

1. Les conférences Macy (1943 - 1953)
2. Hixon Symposium on Cerebral Mechanisms in Behavior (1948)
3. Session on Learning Machines (1955)
4. Dartmouth Summer School on Artificial Intelligence (1956)
5. Symposium on the "Mechanization of Thought Processes" (1958)

[Nilsson,10]] N. Nilsson. *The quest for artificial intelligence.* Cambridge University Press, 2010.

Session on Learning Machines (1955)

Quatre papiers importants

- Wesley Clark and Belmont Farley : « *Generalization of Pattern Recognition in a Self-Organizing System* » : Some pattern-recognition experiments on networks of neuron-like elements. Règle de Hebb. Allusion à capacité de généralisation.
- Gerald Dinneen (1924-) : « *Programming Pattern Recognition* ». Computational techniques for processing images. Suggère d'utiliser des filtres sur des images pixelisées en niveaux de gris.
- Oliver Selfridge (1926-2008) : « *Pattern Recognition and Modern Computers* ». Techniques for highlighting features in clean-up images" (coins, carrés, triangles)
- Allen Newell (1927-1992) : « *The Chess Machine: An Example of Dealing with a Complex Task by Adaptation* ». About programming a computer to play chess. Notions de buts, de recherche dans un espace de sous-buts, de recherche en meilleur d'abord, d'heuristique, de calcul des prédicats.

Dartmouth Summer School on Artificial Intelligence (1956)

- McCarthy (1927 -) : langage de la pensée (précurseur de Lisp).
- Allen Newell and Simon (1916-2001). Logic Theorist.
- Marvin Minsky (1927 -). Réseaux connexionnistes -> approche symbolique "consider a machine that would tend to build up within itself **an abstract model of the environment** in which it is placed. If it were given a problem it would **first explore solutions** within the internal abstract model of the environment and then **attempt external experiments**".

Symposium on the "Mechanization of Thought Processes" (1958)

- Minsky : méthodes pour la planification, l'apprentissage, et la reconnaissance des formes.
- McCarthy : logique des prédicats et Lisp.
- Oliver Selfridge : « *Pandemonium: A Paradigm for Learning* ».
  Une architecture hiérarchique de "démons" pour résoudre des problèmes + la suggestion d'un processus d'apprentissage.

Caractéristiques essentielles

- Exploration de règles d'adaptation *sans critère de succès explicite* (qui guide le processus d'apprentissage)
- Apprentissage **en-ligne**
- L'apprentissage dans CHECKER comme précurseur de l'*apprentissage par différences temporelles*.
- **Modèles locaux et combinaisons** (RN, Pandemonium)
- Importance de la **représentation** (attributs de description)

2 questions centrales :

1. le « *credit assignment problem* »
2. l'invention de nouveaux prédicats

L'exemple de CHECKER

Combinaison de descripteurs et attribution de mérite

[Arthur Samuel (1901 - 1990) ;
1952 (IBM-701), 1954 (IBM-704), avec apprentissage : 1956 ...]

Apprentissage de la fonction d'évaluation dans une approche MinMax.

$$\boxed{valeur(position) = \sum_{i=1}^{n} w_i \, f_i}$$

Deux problèmes :

- **Sélectionner** de bonnes fonctions de base f_i
- **Pondérer** l'importance de ces fonctions grâce aux w_i

L'exemple de CHECKER

Combinaison de descripteurs et attribution de mérite

Pondération des fonctions de bases :

Apprentissage de la fonction d'évaluation dans une approche MinMax.

Fonction linéaire de 38 attributs (n'utilisant que les 16 meilleurs).

Principe : modifier les poids à la racine pour que l'évaluation soit plus proche de celle ramenée par MinMax.

Précurseur de la méthode des différences temporelles [Sutton] en apprentissage par renforcement.

Et apprentissage par cœur de la valeur de certaines positions pour des parties jouées.

`http://www.fierz.ch/samuel.html`

L'exemple de CHECKER

Combinaison de descripteurs et attribution de mérite

Recherche de bonnes fonctions de bases :

Dans une collection de 32 fonctions, choix aléatoire de 16.

À chaque fois qu'une fonction de base a eu la moins bonne pondération, son score est augmenté de 1.

Quand ce score dépasse 32, cette fonction est éliminée et remplacée par une autre du pool.

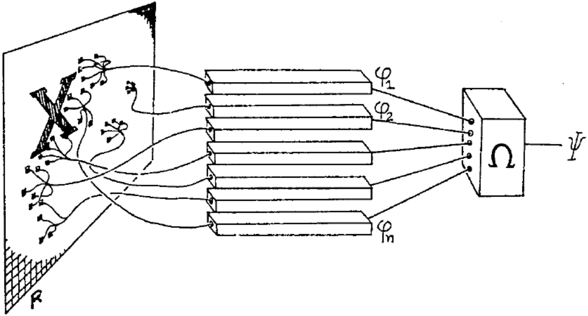
Jugé pas très satisfaisant par Samuel qui voudrait pouvoir **inventer** de nouvelles fonctions de base

Plan

1. Les pionniers
2. Les années 60-70
    - Le premier connexionnisme
    - Reconnaissance des Formes
    - Systèmes experts et apprentissage
3. Avènement de l'analyse statistique de l'apprentissage
4. Et maintenant : les grandes questions ?
5. Conclusion

Premier connexionnisme
Le perceptron de Rosenblatt

Frank Rosenblatt (1928 - 1969)



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

Premier connexionnisme
L'algorithme du perceptron

Apprentissage des poids w_i

Principe (**règle de Hebb**) : en cas de succès, ajouter à chaque connexion quelque chose de proportionnel à l'entrée et à la sortie.

Règle du perceptron : apprendre seulement en cas d'échec.

Algorithme 1 : Algorithme d'apprentissage du perceptron

tant que *non convergence* **faire**
  si *la forme d'entrée est correctement classée* **alors**
    ne rien faire
  sinon

$$\boxed{\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\ \mathbf{x}_i\ y_i}$$

  fin
  Passer à la forme d'apprentissage suivante
fin

Premier connexionnisme

L'algorithme du perceptron

Deux théorèmes fondamentaux :

Le perceptron **peut apprendre tout ce qu'il peut représenter** !!

[David Block, 1962], [Albert Novikoff, 1963]

L'apprentissage s'effectue **en un nombre fini d'étapes**

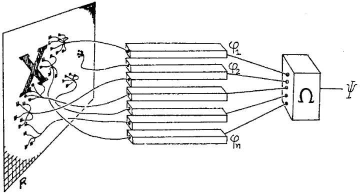
(si une séparatrice linéaire existe)

$T \leq \frac{2R^2}{M^2}$ où $M = \min_{1 \leq i \leq m}\{y_i \; d(\mathbf{x}_i, H^\star)\}$ pour un certain séparateur $H^\star$.

Premier connexionnisme

L'algorithme du perceptron : **caractéristiques**

1. Algorithme en-ligne
2. Ne pouvait pas tout apprendre !?
   - Car ne peut pas tout représenter
   - $\Rightarrow$ Avoir de bonnes fonctions de base (détecteurs locaux)
   - $\Rightarrow$ Savoir les combiner de manière précise



Blocage

Premier connexionnisme

Le problème de l'apprentissage

Encore une démarche exploratoire.

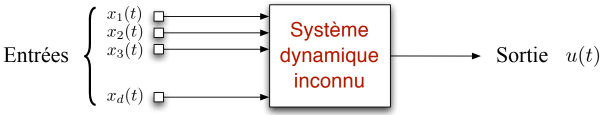
Pas de principe normatif et générique sous-jacent

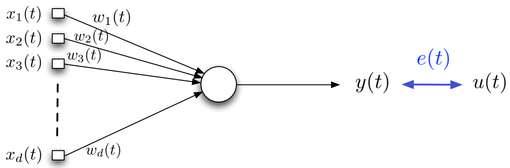
Mais des problèmes qui commencent à se préciser.

Premier connexionnisme
Règle de Widrow-Hoff

Conçue dans le cadre du **filtrage adaptatif**.

Chercher un **modèle linéaire d'un signal temporel** : $y(t) = \sum_{k=1}^{M} w_k(t) x_k(t)$





Premier connexionnisme
Règle de Widrow-Hoff

$$\boxed{\ell(\mathbf{w}) = \frac{1}{2} e^2(t)}$$

Méthode de gradient :

$$\frac{\partial \ell(\mathbf{w})}{\partial \mathbf{w}} = e(t)\, \frac{\partial e(t)}{\partial \mathbf{w}}$$

$$e(t) = u(t) - \mathbf{x}^\top(t)\, \mathbf{w}(t) \quad \text{d'où :} \quad \frac{\partial e(t)}{\partial \mathbf{w}(t)} = -\mathbf{x}(t)$$

$$\frac{\partial \ell(\mathbf{w})}{\partial \mathbf{w}(t)} = -\mathbf{x}(t)\, e(t)$$

$$\boxed{\mathbf{w}(t+1) \;=\; \mathbf{w}(t) + \eta\, \mathbf{x}(t)\, e(t)}$$

[Widrow-Hoff:60] B. Widrow and M. Hoff. *Adaptive Switching Circuits.* IRE WESCON Conv. Rec. Pt.4, pp.96-104.

Premier connexionnisme
Questions

Comment **comparer** des apprentissages ?

- Pas de méthodologie

Apprentissage des **descripteurs**

- Un problème essentiel

Apprentissage **hiérarchique**

- Comment combiner et apprendre un modèle hiérarchique ?

Approche bayésienne de la reconnaissance des formes
La tâche

Associer une **forme d'entrée** à une **décision**

« *Étant donnés des exemples de signaux complexes et les décisions correctes associées, trouver automatiquement les bonnes décisions pour une séquence d'exemples futurs.* »[Ripley, 96]

- Apprentissage (supervisé)
- Échantillon d'apprentissage
- Échantillon de test
- Bonne décision

[Ripley:96] Bernard Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

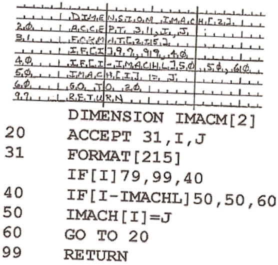
[Duda et al.:01] R. Duda, P. Hart and D. Stork. *Pattern classification.* Wiley Interscience (2nd Ed., 2001).

Approche bayésienne de la reconnaissance des formes

Exemple : la reconnaissance de caractères manuscrits

Applications

- Reconnaissance de ***caractères***
- Reconnaissance de la ***parole***
- Reconnaissance de ***gestes*** (lecture sur les lèvres)
- Reconnaissance de ***particules*** (trajectoires dans les chambres à bulles)
- ...



Approche bayésienne de la reconnaissance des formes
Questions

Comment décrire les formes ?

- Notion d'attributs

Comment représenter la connaissance ?

- Distributions de probabilité
  - sur les décisions possibles
  - sur les exemples

Qu'est-ce qu'une décision ?

- Modèle génératif
- Modèle discriminants
- Modèle par fonction de décision

Approche bayésienne de la reconnaissance des formes
Modèles génératifs vs. Modèles discriminants

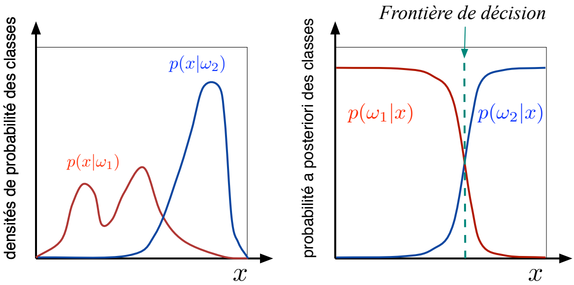


Figure: *Exemple d'une tâche de classification à deux classes. À gauche, selon l'**approche générative**, on a estimé les probabilités conditionnelles $\mathbf{p}_{\mathcal{X}|\mathcal{Y}}(\mathbf{x}|u_k)$. À droite, selon l'**approche discriminative**, seules les probabilités a posteriori sont estimées. La connaissance des détails des distributions $\mathbf{p}_{\mathcal{X}|\mathcal{Y}}(\mathbf{x}|u_k)$ n'a pas d'effet sur la détermination du seuil de décision.*

Approche bayésienne de la reconnaissance des formes

Théorie de la décision optimale

Quantifie le compromis entre les décisions en fonction de leur probabilité et des coûts de mauvaise décision.

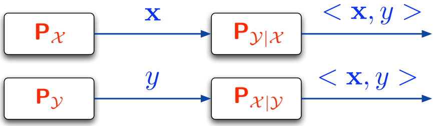


Figure: *Scenarii de génération d'exemples.*

Coût d'une décision $h(\mathbf{x})$:

$$L(h) \;=\; \begin{cases} \mathbf{P}_{\mathcal{X}\mathcal{Y}}\{h(\boldsymbol{x}) \neq y\} & \text{(si même coût de mauvaise décision)} \\ \mathbb{E}[\ell(h(\boldsymbol{x}), y)] \;=\; \displaystyle\int_{\boldsymbol{x}\in\mathcal{X}, y\in\mathcal{Y}} \ell(h(\boldsymbol{x}), y)\; \mathbf{P}_{\mathcal{X}\mathcal{Y}}\; d(\boldsymbol{x}, y) & \end{cases}$$

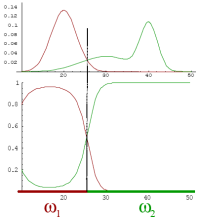
Approche bayésienne de la reconnaissance des formes
Théorie de la décision optimale

Règle de Bayes

$$P(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_i)\,P(\mathcal{C}_i)}{p(\boldsymbol{x})}$$

Décision bayésienne

$$\mathcal{C}^\star = \begin{cases} \underset{\mathcal{C}_k \in \mathcal{H}}{\text{ArgMax}}\, P(\mathcal{C}_k|\mathbf{x}) \\ \underset{\mathcal{C}_k \in \mathcal{H}}{\text{ArgMin}} \left\{ \sum_{k=1}^{C} \ell(\mathcal{C}^\star, \mathcal{C}_k) \cdot \mathbf{P}(\mathcal{C}_k|\mathbf{x}) \right\} \end{cases}$$



Il faut connaître $P(\mathcal{C}_k)$ et les lois de probabilité $p(\boldsymbol{x}|\mathcal{C}_i)$!

Approche bayésienne de la reconnaissance des formes

Apprentissage...

... de $P(\mathcal{C}_k)$ et les lois de probabilité $p(\boldsymbol{x}|\mathcal{C}_i)$

En général, on se donne UNE distribution de probabilité paramétrée pour représenter l'ensemble des distributions $\mathbf{P}(\boldsymbol{x}|\mathcal{C}_i)$: $\mathbf{p}(\mathbf{x}|\theta)$.

E.g. un mélange de C gaussiennes $\Rightarrow C \cdot (d + d^2)$ paramètres.

Approche bayésienne de la reconnaissance des formes

Principes inductifs

Soit $\mathcal{S} = \langle(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\rangle$ l'échantillon d'apprentissage

Règle du Maximum A Posteriori (MAP)

$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMax}} \; \mathbf{P}(h \mid \mathcal{S}) = \underset{h \in \mathcal{H}}{\text{ArgMax}} \; \frac{\mathbf{P}(\mathcal{S} \mid h)\, \mathbf{P}(h)}{\mathbf{P}(\mathcal{S})} = \underset{h \in \mathcal{H}}{\text{ArgMax}} \; \mathbf{P}(\mathcal{S} \mid h)\, \mathbf{P}(h)$$

Cas de données i.i.d.

$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMax}} \left[\mathbf{P}(h) \prod_{i=1}^{m} \mathbf{P}(\mathbf{z}_i \mid h) \right] = \underset{h \in \mathcal{H}}{\text{ArgMax}} \left[\log(\mathbf{P}(h)) + \sum_{i=1}^{m} \log(\mathbf{P}(\mathbf{z}_i \mid h)) \right]$$

Règle du Maximum de Vraisemblance (MLE)

Hypothèses équiprobables

$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMax}} \prod_{i=1}^{m} \mathbf{P}(\mathbf{z}_i \mid h) = \underset{h \in \mathcal{H}}{\text{ArgMax}} \sum_{i=1}^{m} \log(\mathbf{P}(\mathbf{z}_i \mid h))$$

Approche bayésienne de la reconnaissance des formes

Grandes familles de méthodes

- Les méthodes *paramétriques*, où l'on suppose que les $\mathbf{p}(\mathbf{x} \mid \omega_i)$ possèdent une certaine forme analytique ; en général, on fait l'hypothèse qu'elles sont des distributions gaussiennes.
- Les méthodes *non paramétriques*, pour lesquelles on estime localement les densités $\mathbf{p}(\mathbf{x} \mid \omega_i)$ au point $\mathbf{x}$ en observant l'ensemble d'apprentissage autour de ce point. Ces méthodes sont implémentées par la technique des *fenêtres de Parzen* ou l'algorithme des *k-plus proches voisins*.
- Les méthodes semi-paramétriques, pour lesquelles nous ne connaissons pas non plus la forme analytique des distributions de probabilités. Nous supposons cependant que ces distributions appartiennent à des familles et que les « hyper-paramètres » qui les caractérisent à l'intérieur de cette famille peuvent être déterminés.

Approche bayésienne de la reconnaissance des formes
Grandes familles de méthodes

- Les méthodes *paramétriques*, où l'on suppose que les $\mathbf{p}(\mathbf{x} \mid \omega_i)$ possèdent une certaine forme analytique ; en général, on fait l'hypothèse qu'elles sont des distributions gaussiennes.
- Les méthodes *non paramétriques*, pour lesquelles on estime localement les densités $\mathbf{p}(\mathbf{x} \mid \omega_i)$ au point $\mathbf{x}$ en observant l'ensemble d'apprentissage autour de ce point. Ces méthodes sont implémentées par la technique des *fenêtres de Parzen* ou l'algorithme des *k-plus proches voisins*.
- Les méthodes semi-paramétriques, pour lesquelles nous ne connaissons pas non plus la forme analytique des distributions de probabilités. Nous supposons cependant que ces distributions appartiennent à des familles et que les « hyper-paramètres » qui les caractérisent à l'intérieur de cette famille peuvent être déterminés.

Approche bayésienne de la reconnaissance des formes
Avantages et limites

- Utilise naturellement la connaissance préalable et les nouvelles données
- Régularise naturellement (par les $\mathbf{P}(\mathcal{C})$)

mais ...

- Fléau de la dimensionnalité
- Problème si H $\neq$ F
- Question de la sélection de modèle

Conclusion provisoire

On a précisé la tâche

- E.g. Apprentissage supervisé
- Généralisation (*échantillon d'apprentissage ; échantillon de test*)

On a des méthodes

- Critère de décision optimale
- Critères inductifs (e.g. MAP, MLE, ...)

Nouveaux présupposés

- Distributions de probabilité sous-jacentes ($\mathbf{P}(\mathcal{C}_k)$, $\mathbf{p}(\mathbf{x} \mid \mathcal{C}_k)$)
- Données i.i.d. (indépendamment et identiquement distribuées)

Intelligence artificielle symbolique

Connaissances et raisonnement

Les techniques de la reconnaissance des formes **ne permettent pas** de :

- **apprendre des descriptions** (vs. des règles de discrimination) [McCarthy, Stanford, 1971, pour la robotique]
- **apprendre les règles** d'un système expert
- **apprendre des descriptions structurées**

Intelligence artificielle symbolique

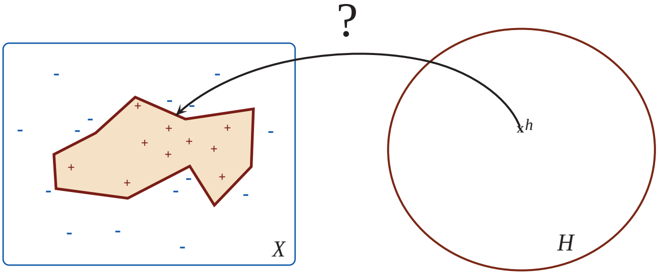
Connaissances et raisonnement

- Représentations logiques
- Règles d'inférence
- Systèmes Experts et bases de règles

On ne parlera pas d'Arch, ni de SOAR, ni d'ACT*

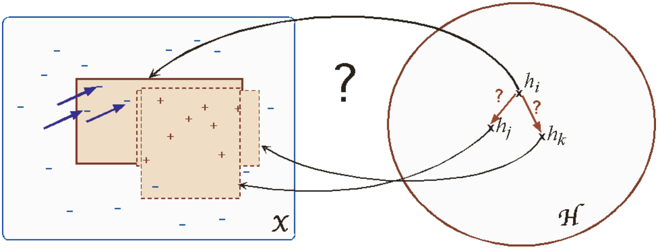
Intelligence artificielle symbolique

Apprentissage de règles



Intelligence artificielle symbolique

Apprentissage de règles



Intelligence artificielle symbolique

Apprentissage de règles

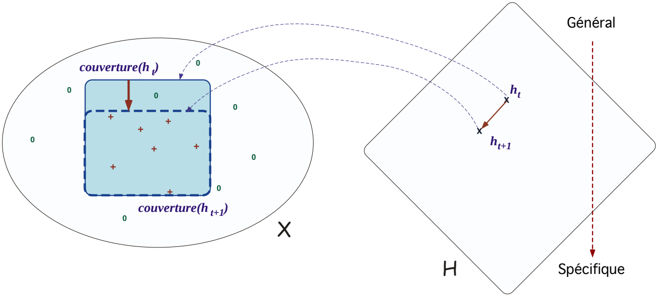


Figure: *La relation d'inclusion dans $\mathcal{X}$ induit la relation de généralisation dans $\mathcal{H}$. Ici, $h_{t+1} \preceq h_t$.*

Intelligence artificielle symbolique

Apprentissage de règles

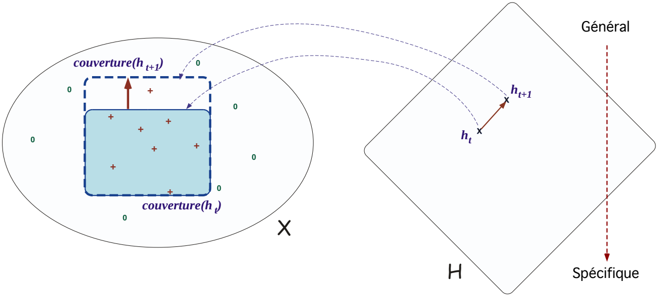


Figure: *La relation d'inclusion dans $\mathcal{X}$ induit la relation de généralisation dans $\mathcal{H}$. Ici, $h_{t+1} \succeq h_t$.*

Intelligence artificielle symbolique
Apprentissage de règles

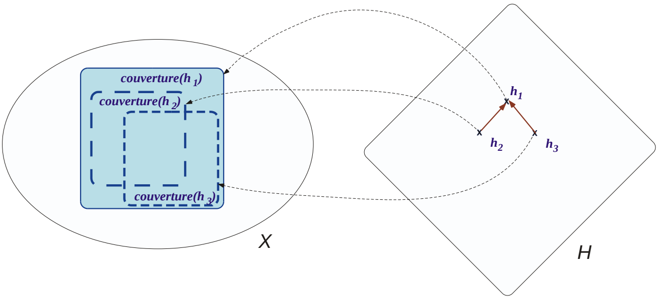


Figure: *La relation d'inclusion dans $\mathcal{X}$ induit la relation de généralisation dans $\mathcal{H}$. Il s'agit d'une relation d'ordre partielle : ici, les hypothèses h_2 et h_3 sont incomparables entre elles, mais elles sont toutes les deux plus spécifiques que h_1.*

Intelligence artificielle symbolique
Apprentissage de règles

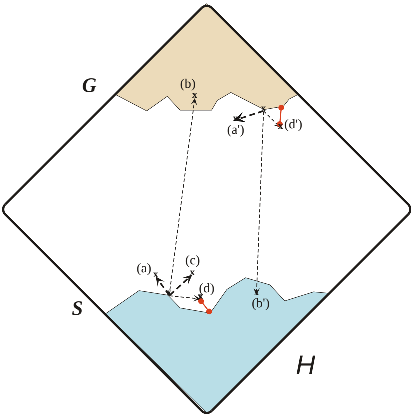


Figure: *Cette figure schématise les différents cas possibles lors de la mise à jour des ensembles S et G par l'algorithme d'élimination des candidats.*

Intelligence artificielle symbolique

Apprentissage de règles et espace des versions

Algorithme 2 : Algorithme d'apprentissage du perceptron

tant que *non convergence* **faire**
  si *la forme d'entrée est correctement classée* **alors**
    ne rien faire
  sinon

$$\boxed{\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\ \mathbf{x}_i\ y_i}$$

  fin
  Passer à la forme d'apprentissage suivante
fin

Intelligence artificielle symbolique

Apprentissage de règles et espace des versions

Algorithme 3 : Algorithme d'élimination des candidats.

Résultat : Initialiser G comme l'hypothèse la plus générale de $\mathcal{H}$
Initialiser S comme l'hypothèse la moins générale de $\mathcal{H}$
pour chaque *exemple* $\mathbf{x}$ **faire**
 si $\mathbf{x}$ *est un exemple positif* **alors**
  Enlever de G toutes les hypothèses qui ne couvrent pas $\mathbf{x}$
  pour chaque *hypothèse s de S qui ne couvre pas* $\mathbf{x}$ **faire**
   Enlever s de S
   `Généraliser`(s,$\mathbf{x}$,S)
   c'est-à-dire : ajouter à S toutes les généralisations minimales h de s telles que :
   - h couvre $\mathbf{x}$ et
   - il existe dans G un élément plus général que h

   Enlever de S toute hypothèse plus générale qu'une autre hypothèse de S
  fin
 sinon
  Enlever de S toutes les hypothèses qui couvrent $\mathbf{x}$
  pour chaque *hypothèse g de G qui couvre* $\mathbf{x}$ **faire**
   Enlever g de G
   `Spécialiser`(g,$\mathbf{x}$,G)
   c'est-à-dire : ajouter à G toutes les spécialisations maximales h de g telles que :
   - h ne couvre pas $\mathbf{x}$ et
   - il existe dans S un élément plus spécifique que h

   Enlever de G toute hypothèse plus spécifique qu'une autre hypothèse de G
  fin
 fin
fin

Intelligence artificielle symbolique

Apprentissage de règles et espace des versions

Applications

- Learning apprentices
- Inférence grammaticale
- Explanation-Based Learning : PRODIGY, ...
- SOAR, ACT*, ...

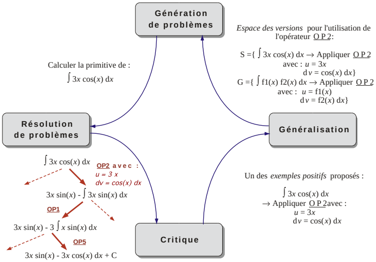


Figure: *Un exemple de cycle d'apprentissage dans le système* LEX.

Intelligence artificielle symbolique

Apprentissage de règles et espace des versions

Leçons

1. Apprentissage = recherche dans un espace d'hypothèses
2. Intérêt de disposer d'une structure dans $\mathcal{H}$ reflétant la relation d'inclusion dans $\mathcal{X}$ (*apprentissage de concept*)
3. Très efficace : très petits échantillons

Critère inductif :

- Compréhensibilité
- Risque empirique nul

Intelligence artificielle symbolique

Apprentissage de règles et espace des versions

Limites

- $\mathcal{H} = \mathcal{F}$ (pas de bruit dans les données)
- Difficulté de trouver des opérateurs de généralisation / spécialisation (avoir un treillis de généralisation)
- Complexité des opérations de calcul de subsomption (plus phénomène de transition de phase révélé plus tard (1999))

Apprentissage artificiel

Importance de la structures sur $\mathcal{H}$ pour l'exploration

Pas de $\mathcal{H}$

Plus-proches-voisins

$\mathcal{H}$ muni d'une distance

Réseaux de neurones ; régression logistique ; modèles bayésiens ; HMM ; ...

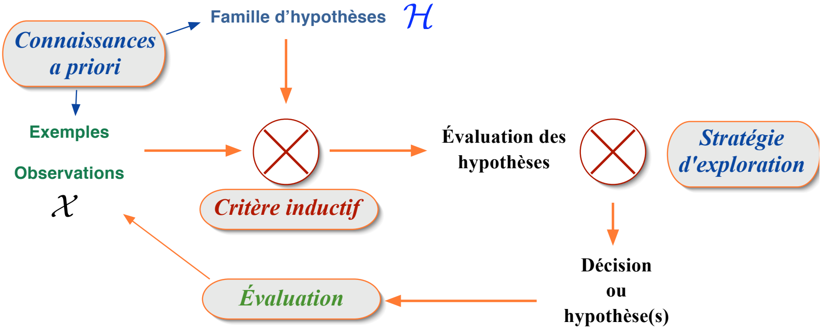
- Optimisation directe (e.g. pseudo-inverse)
- Adaptation itérative = descente de gradient

$\mathcal{H}$ muni d'une relation de généralité

Inférence grammaticale ; Induction de règles ; Apprentissage relationnel

- Apprentissage symbolique
- Bruit

L'apprentissage
Ingrédients



Apprentissage de règles et questionnement PAC

La fin d'un monde, le début d'un autre

Quelles **classes de concepts** peuvent être appris ?
(concepts logiques : k-CNF, DNF, μ-expressions)

Définition d'un **protocole d'apprentissage** : appels possibles à :

- EXAMPLES : fournit un exemple '$+$' tiré i.i.d.
- ORACLE : dit si l'exemple choisi par l'apprenant est '$+$' ou '$-$'

Apprenabilité

- Complexité polynomiale en taille f (programme à apprendre) et nombre de variables.
- $\forall f \in \mathcal{H}$ et $\forall$ $\mathbf{p}_{\mathcal{X}\mathcal{Y}}$, $\quad P^m\left(R_{\text{Réel}}(h^\star_{\mathcal{S}})\right) \geq \varepsilon) < \delta$

Intention

Montrer que la classe des concepts apprenables est non vide mais limitée, d'où nécessité d'un apprentissage cumulatif, hiérarchique et guidé.

[Valiant:84] Leslie Valiant. *A theory of the learnable.* Communications of the ACM 27:11 (1984) pp.1134-1142.

Apprentissage de règles et questionnement PAC

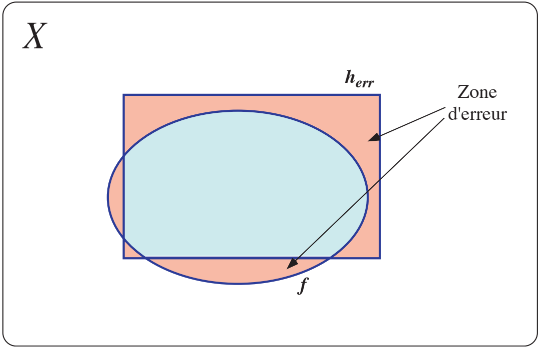


FIG.: *La zone d'erreur dans le cas de l'apprentissage d'un concept ou fonction binaire définie sur* $\mathcal{X}$.

Apprentissage de règles et questionnement PAC

Quelle est la probabilité qu'une hypothèse h_{err} soit de $R_{\text{Emp}}(h_{err}) = 0$ et de $R_{\text{Réel}}(h_{err}) = P_{\mathcal{X}}(h_{err} \Delta f) \geq \varepsilon$?

Probabilité de « survie » de h_{err} après une observation : $1 - \varepsilon$

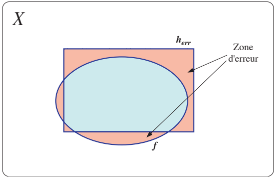
Après m exemples (i.i.d.) : $(1 - \varepsilon)^m$

Probabilité qu'une hypothèse survive : $|\mathcal{H}|(1 - \varepsilon)^m$
(union d'évènements disjoints)

On cherche les conditions pour que :
$P^m(|R_{\text{Réel}}(h^\star_{\mathcal{S}}) - R_{\text{Réel}}(h^\star)| \geq \varepsilon) < \delta$

En posant $\delta = |\mathcal{H}|(1 - \varepsilon)^m$, on obtient : $m \geq \frac{1}{\varepsilon} \ln \frac{|\mathcal{H}|}{\delta}$

Ou encore, ε varie en : $\varepsilon = \dfrac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{m}$
(Convergence rapide (quand $\mathcal{H} = \mathcal{F}$))



- La garantie sur l'apprentissage dépend de m et $|\mathcal{H}|$
- Analyse en pire cas : pour toute hypothèse (convergence uniforme sur $\mathcal{H}$) et pour toute distribution $\mathbf{p}_{\mathcal{X}\mathcal{Y}}$

Apprentissage de règles et questionnement PAC

Un tournant

Le côté « **tradition** » :

- Apprentissage de concept
- Question : quels concepts (logiques) sont apprenables ?
- $\mathcal{H} = \mathcal{F}$ (apprentissage humain)

Le côté **révolutionnaire** :

- Très forte abstraction de l'algorithme d'apprentissage : suit le principe MRE
- On suppose « seulement » qu'il est capable de trouver une hypothèse de risque empirique minimal (nul).

Apprentissage de règles et questionnement PAC

Deux papiers précurseurs injustement moins connus :

[Stone, 77]

Un algorithme d'apprentissage $\mathcal{A}$ a la propriété de *consistance universelle* si :

$$\forall\ \mathbf{p}_{\mathcal{X}\mathcal{Y}} : \qquad R_{\textit{Réel}}(\mathcal{A}(\mathcal{S}_m)) \;=\; R_{\textit{Réel}}(h^{\star}_{\mathcal{S}_m}) \underset{m \to \infty}{\longrightarrow} R_{\textit{Réel}}(h^{\star})$$

Stone montre par une preuve élégante qu'un système d'apprentissage particulier, la classification par les k-plus-proches-voisins, est universellement consistant.

[Pearl, 78]

Pose la question de la préférence pour des hypothèses simples.

Montre qu'elles sont naturellement tirées d'espaces $\mathcal{H}$ de capacité faible et donc dont le risque réel est proche du risque empirique.

[Stone:77] Charles Stone. *Consistent nonparametric regression.* Ann. Statist., 5(4) :595–620, 1977.

[Pearl:78] Judea Pearl. *On the connection between the complexity and credibility of inferred models.* Int. J. Gen. Syst., 4 :255–264, 1978.

Apprentissage

Retour vers le futur : les k-ppv

Questions :

- Comment **représenter les données** pour pouvoir définir une distance ?
- Quelle mesure de **distance** ?
  - Euclidienne ?
- **Normalisation** ?
- Quelle **valeur de** k ?
  - Fenêtres de Parzen (et fonctions noyau)
- **Effets des grandes dimensions** : espace vide et homogène
- Asymptotiquement consistant et quasi optimal
  - Mais **avec des petits échantillons** ?
- **Efficacité calculatoire** ?

Plan

1. Les pionniers
2. Les années 60-70
3. Avènement de l'analyse statistique de l'apprentissage
   - Analyse statistique de l'apprentissage
   - Vers de nouveaux principes inductifs
   - SVM et méthodes à noyaux
   - Les réseaux connexionnistes : old and new again
   - Le « *no-free lunch theorem* »
   - Le reste
4. Et maintenant : les grandes questions ?
5. Conclusion

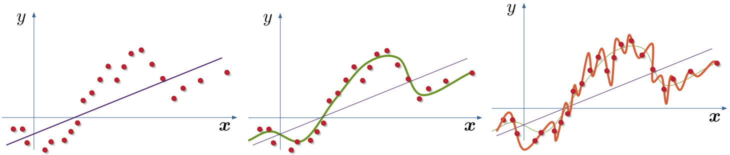
Analyse statistique de l'apprentissage

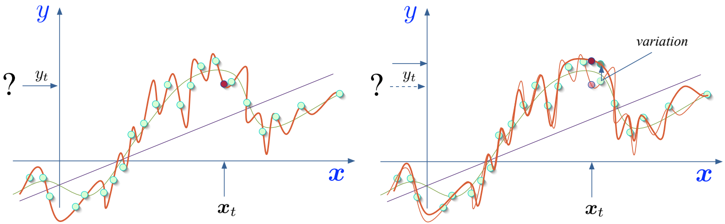
L'induction est un problème mal-posé

1. Conditions nécessaires et suffisantes pour la validité du principe MRE
2. Bornes et vitesses de convergence
3. Nouveaux principes inductifs ?
4. Nouveaux algorithmes ?

[Vapnik:95] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995

Analyse statistique de l'apprentissage





Analyse statistique de l'apprentissage

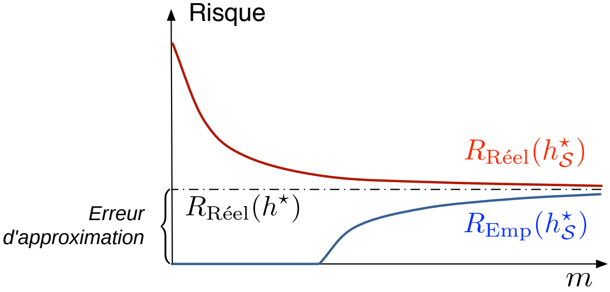
Attention !

[Vapnik et Chervonenkis, 68-91]

1. L'induction est un **problème mal-posé**
   - $f \xrightarrow{\text{tirage i.i.d.}} \mathcal{S}$
   - Induction $\equiv$ trouver f à partir de $\mathcal{S}$
   - mais $\mathcal{S} \underset{\delta}{\longrightarrow} \mathcal{S}_\delta$ peut conduire à f très différent de f_δ
2. Le principe de **Minimisation du Risque Empirique** n'est **pas évident**
3. Gros problème quand échantillon « petit »

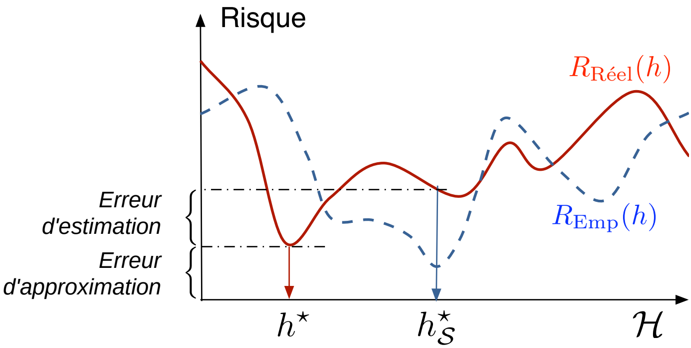
Analyse statistique de l'apprentissage

Pertinence du principe MRE



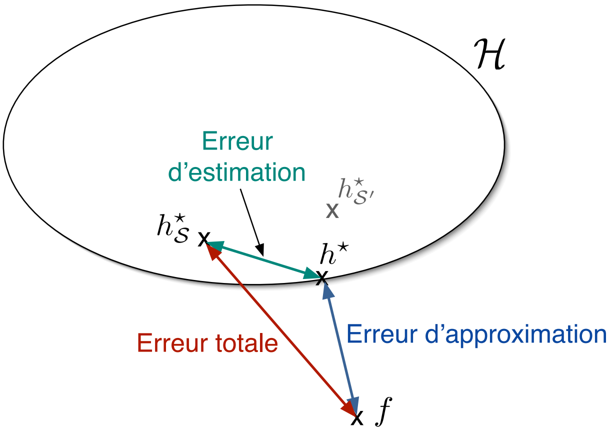
Analyse statistique de l'apprentissage

Pertinence du principe MRE



Analyse statistique de l'apprentissage

Le compromis biais-variance



Analyse statistique de l'apprentissage

Conditions de pertinence du MRE

Pour $|\mathcal{H}|$ fini

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \; \leq \; R_{\text{Emp}}(h) \; + \; \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2\,m}} \right] \; > \; 1 - \delta$$

Rq. : Lorsque $\mathcal{H} = \mathcal{F}$, **la convergence est beaucoup plus rapide** puisqu'elle se fait en $\mathcal{O}(1/m)$ au lieu de $\mathcal{O}(\sqrt{1/m})$.

Pour $|\mathcal{H}|$ infini

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R_{\text{Réel}}(h) \; \leq \; R_{\text{Emp}}(h) \; + \; \underbrace{g(\log|\mathcal{H}|, \delta, m)}_{\text{Terme exprimant la richesse de } \mathcal{H}} \right] \; > \; 1 - \delta$$

Analyse statistique de l'apprentissage

Mesures de richesse de $\mathcal{H}$ et régularisation

Mesures de richesse de $\mathcal{H}$:

- Modèles paramétrés et hypothèses de distribution normale : AIC ou BIC
- Dimension de Vapnik-Chervonenkis
- Complexité de Rademacher
- Nombres de couverture
- Marge
- ...

Nouveaux critères inductifs régularisés

Analyse statistique de l'apprentissage

Critères inductifs régularisés

Contrôler $d_{\mathcal{H}}$

1. "Sélection de modèle"
2. Puis choix de $h \in \mathcal{H}$

$$\hat{h} \;=\; \text{ArgMin}_{h \in \mathcal{H}} \left[R_{Emp}(h) + \text{Capacité}(\mathcal{H}) \right]$$

Régularisation

- Contrôler directement la complexité de h

$$\hat{h} \;=\; \text{ArgMin}_{h \in \mathcal{H}} \left[R_{Emp}(h) + \lambda \, \text{Reg}(h) \right]$$

Analyse statistique de l'apprentissage

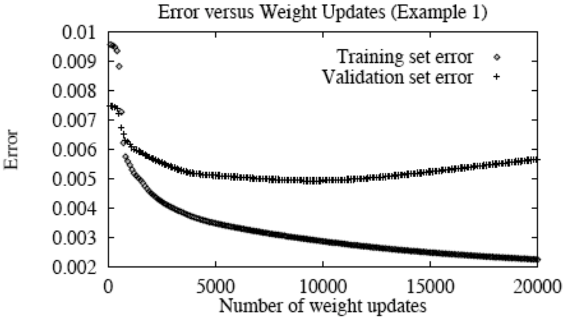


Figure: Évolution des courbes d'apprentissage et de test en fonction du nombre de présentations (weight updates) de la base d'exemples d'apprentissage.

Vers de nouveaux principes inductifs

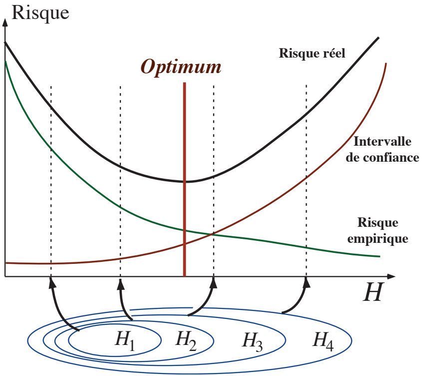


Figure: *Le principe SRM.*

SVM et méthodes à noyaux

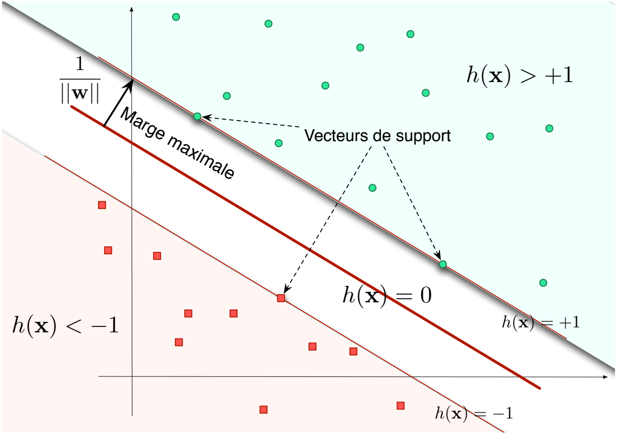
Et si ...

...

- On s'arrangeait pour **avoir** $R_{\text{Emp}}(h^{\star}_{\mathcal{S}}) = 0$
  - Convergence rapide (en $\mathcal{O}(1/m)$)
  - On joue sur la richesse de $\mathcal{H}$
- On cherchait une **séparatrice linéaire**
  - Problème d'optimisation convexe
- ... de **marge maximale**
  - Contrôle la richesse de $\mathcal{H}$

$\Rightarrow$ on aurait les SVM !

SVM et méthodes à noyaux



SVM et méthodes à noyaux

Réalisation technique → astuce des noyaux

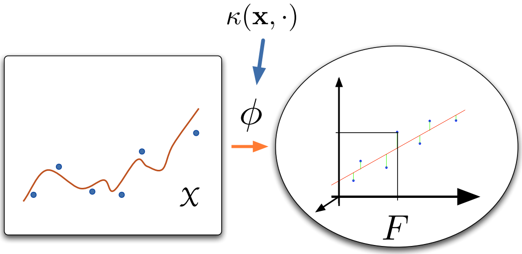
$$\boxed{h^*(\mathbf{x}) \;=\; (\mathbf{w}^\star \;\; \mathbf{x}) \;+\; w_0^\star \;=\; \sum_{i=1}^{m} \alpha_i^\star \, u_i \,.\, \langle \mathbf{x}_i \,,\, \mathbf{x} \rangle \;+\; w_0^\star}$$

- Représentation ***non paramétrique*** : fonction des exemples d'apprentissage
  - La complexité de l'hypothèse dépend de $|\mathcal{S}|$
- **Parcimonieux**
  - Seulement les $m' \leq m$ exemples support
  - Solution la plus robuste (large marge) $\Rightarrow$ donc nécessite moins de précision (le moins de bits)
- Ne dépend pas de d, mais de m'

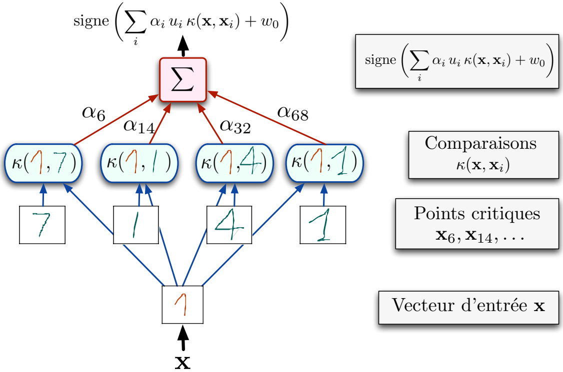
SVM et méthodes à noyaux

Réalisation technique → astuce des noyaux

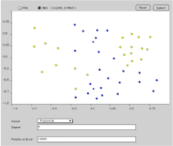
$$\boxed{h(\mathbf{x}) \;=\; \text{signe}\left\{\sum_{i=1}^{m} \alpha_i^\star \, u_i \, \kappa(\mathbf{x}, \mathbf{x}_i) \;+\; w_0^\star\right\}}$$

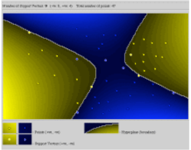


SVM et méthodes à noyaux

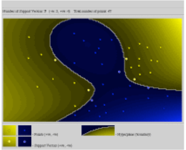


SVM et méthodes à noyaux

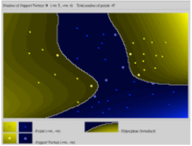




(5-, 4+)



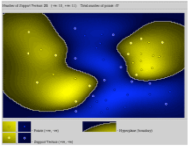
(3-, 4+)



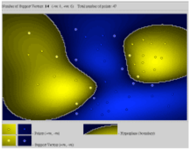
(5-, 4+)

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

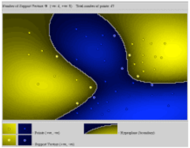
Ici *fonction polynomiale* **de** *degré 2, 5, 8* **et** *C = 10000*



(10-, 11+)



(8-, 6+)



(4-, 5+)

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

SVM et méthodes à noyaux

Théorème de représentation

Toute fonction $\hat{h}$ minimisant un **risque empirique régularisé** admet une **représentation de la forme** :

$$\boxed{\hat{h}(\cdot) \;=\; \sum_{i=1}^{m} \alpha_i \, \kappa(\mathbf{x}_i, \cdot)}$$

Theorem (*representer theorem*)

Soit un noyau reproduisant $\kappa \;:\; \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, *un échantillon d'apprentissage* $S_m \in (\mathcal{X} \times \mathcal{Y})^m$, *et un risque empirique quelconque* R_{Emp} *(muni d'une fonction de perte* ℓ*). Soit* $\text{Reg} \;:\; \mathbb{R} \rightarrow [0, \infty[$ *une fonction croissante strictement monotone. Soit* $\mathcal{H}_\kappa$ *l'espace hilbertien induit par le noyau reproduisant* κ*. Alors, toute fonction* $\hat{h} \in \mathcal{H}_\kappa$ *minimisant le risque régularisé :*

$$\hat{h}(\mathbf{x}) \;=\; \underset{h \in \mathcal{H}}{\text{ArgMin}} \left\{ R_{\text{Emp}}(h, \mathcal{S}) + \lambda \, \text{Reg}(h) \right\}$$

admet une représentation de la forme : $\hat{h}(\cdot) \;=\; \sum_{i=1}^{m} \alpha_i \, \kappa(\mathbf{x}_i, \cdot) \quad \alpha_i \in \mathbb{R}, \forall i.$

SVM et méthodes à noyaux

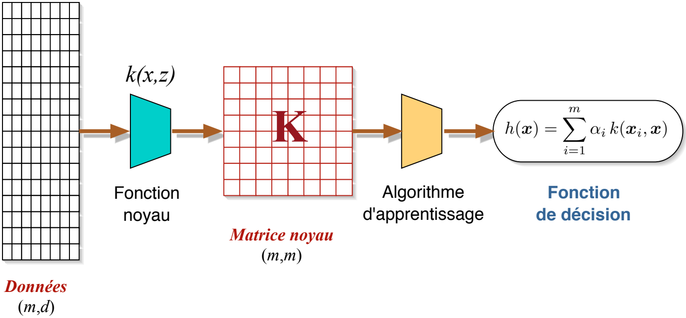


Figure: *Chaîne de traitements générique des méthodes à noyaux.*

SVM et méthodes à noyaux

Bilan conceptuel

1. Retour aux **approches non paramétriques**
2. Retour aux **modèles additifs** (combinaisons linéaires)
3. Nouvelle perspective sur la **découverte de re-descripteurs**
    - Sélection dans un « réservoir »
4. Importance de la **notion de marge**

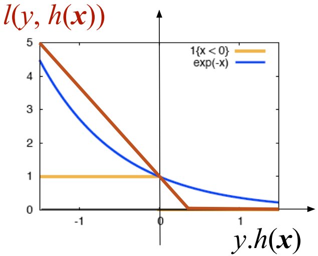
SVM et méthodes à noyaux

Bilan conceptuel et technique

Optimisation du critère inductif régularisé

Minimisation du ϕ-risque empirique

- "hinge loss"
- Fonction de perte exponentielle
- ...



Deux rôles :

- Régulariser
- Faciliter l'optimisation
  - Différentiabilité
  - Convexité

SVM et méthodes à noyaux
Développements

Nécessité de justifier les SVM (stratification *a posteriori*)

Mesure d'adéquation *a posteriori*

- Le cadre de la *félicité* (*luckiness framework*)
- Capacité à « comprimer » l'échantillon d'apprentissage
- Analyse par stabilité du risque empirique
- Prise en compte de l'erreur induite par les imperfections de l'algorithme de recherche dans $\mathcal{H}$
- Transition de phase dans la mesure de couverture des hypothèses

$\Rightarrow$ Retour à des **caractéristiques de l'algorithme** d'apprentissage

Les réseaux connexionnistes : le retour (2)

Les réseaux à architecture profonde

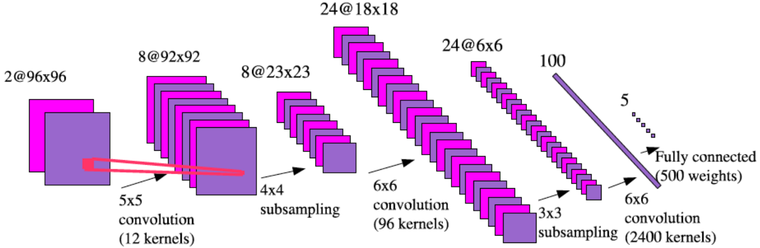


Figure: *Architecture d'un réseau connexionniste utilisé pour traiter des images de la base NORB. L'entrée consiste en une paire d'images, dont le système extraie 8 descripteurs de taille* 92×92 *calculés sur des imagettes* 5×5. *Les sorties de ces descripteurs sont reprises par 8 descripteurs* 23×23, *24 descripteurs* 18×18, *24 descripteurs* 6×6 *et une couche de 100 neurones complètement connectés aux 5 neurones de sortie qui donnent la distance avec les vecteurs cibles. (Repris de [**?**].)*

Les réseaux connexionnistes : le retour (2')

Le *reservoir computing*

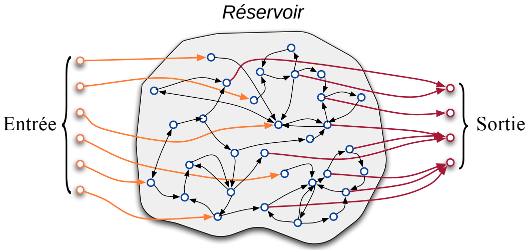


Figure: *Un réseau connexionniste à « réservoir ». L'apprentissage ne se fait que sur la couche de sortie.*

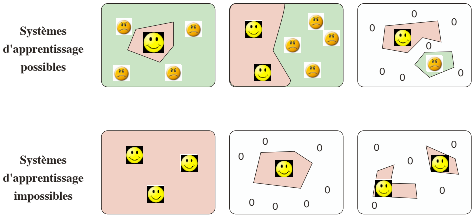
Le « *no-free lunch theorem* »

Toute corrélation entre le passé et le futur est possible

Theorem (No-free-lunch theorem (Wolpert, 1992))

Pour tout couple d'algorithmes d'apprentissage $\mathcal{A}_1$ et $\mathcal{A}_2$, caractérisés par leur distribution de probabilité a posteriori $\mathbf{p}_1(h|\mathcal{S})$ *et* $\mathbf{p}_2(h|\mathcal{S})$, *et pour toute distribution $d_{\mathcal{X}}$ des formes d'entrées $\mathbf{x}$ et tout nombre m d'exemples d'apprentissage :*

En moyenne uniforme sur toutes les fonctions cible f dans $\mathcal{F}$:
$\mathbb{E}_1[R_{\text{Réel}}|f,m] \; - \; \mathbb{E}_2[R_{\text{Réel}}|f,m] \; = \; 0.$



De quoi n'avons-nous pas parlé

1. Le **boosting** (issu d'une question en informatique)
2. **Autres méthodes d'apprentissage**
   - Arbres de décision
   - HMM / modèles graphiques
3. **Autres types d'apprentissage**
   - Apprentissage non supervisé / Fouille de données
   - Apprentissage par renforcement
4. **Autres paradigmes**
   - Induction de Solomonoff (et MDLP)
   - Identification à la limite
   - Prédiction de séquences individuelles
   - Approche par la physique statistique (transition de phase)

Plan

1. Les pionniers
2. Les années 60-70
3. Avènement de l'analyse statistique de l'apprentissage
4. Et maintenant : les grandes questions ?
5. Conclusion

Les défis (*challenges*)

Challenges Pascal

- RTE-5: Recognizing Textual Entailment Challenge
- Active Learning Challenge
- Large Scale Hierarchical Text Classification
- Visual Object Classes Challenge 2009
- Morpho Challenge 2009
- GRavitational lEnsing Accuracy Testing 2008 (GREAT08)
- KDD cup 2009: Fast Scoring on a Large Database Challenge
- Causality challenge
- Visual Object Classes Challenge 2008
- Large Scale Learning Challenge
- Human-machine comparisons of consonant recognition in noise Challenge

Grandes directions

Nouvelles tâches

- Ranking
- Apprentissage et réseaux (graphes)
- Apprentissage dans espaces de très grande dimension : puces ADN ; Vision ; ...
- Apprentissage de sorties structurées
- Recherche de "*deep models*" (modèles causaux hiérarchiques)

Questions théoriques

- Nouveaux critères de succès (e.g. AUC, ...)
- Nouveaux opérateurs de voisinage (graph Laplacian)
- Nouvelles méthodes de réduction de dimensionnalité
- Nouvelles mesures de proximité dans $\mathcal{X} \times \mathcal{Y}$
- Recherche de causalité et apprentissage hiérarchique

Grandes directions

Apprentissage de très gros volumes de données

Very large scale learning

Motivation pratique

- Très gros volumes de données
- e.g. GREAT08, EGEE, Wall-mart, flux de télévision, ...

Types d'apprentissages et directions de recherches

- Optimisation : gradient stochastique
- Optimisation (e.g. nouvelles fonctions de perte de substitution (*surrogates*))
- Apprentissage incrémental

Grandes directions

Combinaison d'apprentissages locaux

Motivation pratique

- Intelligence ambiante. Apprenants en réseaux
- Apprentissage en-ligne

Types d'apprentissages

Apprentissage multi-tâche ; apprentissage semi-supervisé ; co-apprentissage (et méthodes d'ensemble) ; apprentissage incrémental ; apprentissage hiérarchique (par décomposition de tâches)

Questions théoriques

- Notion de « relatedness »
- Transfert

Grandes directions

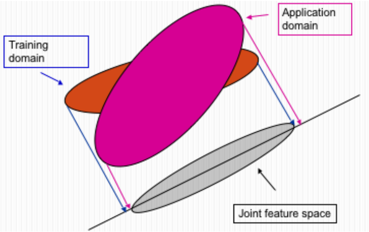
Combinaison d'apprentissages locaux

Le problème du transfert [Shai Ben-David, 2009]

We propose to embed the original attribute space(s) into some feature space in which

- The two tasks look similar.
- The source task can still be well classified.

Then, treat the images of points from both distributions as if they are coming from a unique distribution.



For a class $\mathcal{H}$ of predictors : $d_{\mathcal{H}}(\mathcal{S}_S, \mathcal{S}_T) = 1 - 2\,\mathrm{Min}_{h \in \mathcal{H}}\, \mathrm{err}(h)$

Use classifiers in $\mathcal{H}$ to try to separate the two distributions. Failure to separate means that the distributions are close.

Grandes directions
Combinaison d'apprentissages locaux

Le problème du transfert [Shai Ben-David, 2009]

Theorem Assume we have m be a random labeled sample of the source domain, S, and let $\tilde{\mathcal{U}}_S$ and $\tilde{\mathcal{U}}_T$ be random unlabeled samples of size m' from $\tilde{\mathcal{D}}_S$ and $\tilde{\mathcal{D}}_T$ respectively. Then with probability $1-\delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \boxed{\sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)}}$$

$$+\lambda + d_{\Delta\mathcal{H}}(\tilde{\mathcal{U}}_S, \tilde{\mathcal{U}}_T) + \boxed{\sqrt{\frac{16}{m'}\left(d \log(2m') + \log \frac{4}{\delta}\right)}}$$

More simply : $\boxed{\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\mathcal{H}}(\mathcal{S}_S, \mathcal{S}_T) \inf_{h \in \mathcal{H}} (\mathrm{Er}_T(h) + \mathrm{Er}_S(h))}$

where d is the VC-dim of $\mathcal{H}$ and : $\lambda = \inf_{h \in \mathcal{H}} (\mathrm{Er}_T(h) + \mathrm{Er}_S(h))$

Note that this is a measure of the relatedness of labelling functions of the two tasks.

Grandes directions

Apprentissage **à partir de flux de données**

Motivation pratique

- Omniprésence de flux de données
- Pas (peu) de stockage possible : “*one-pass learning*”
- e.g. consommation électrique personnalisée, vidéos, ...

Types d’apprentissages et directions de recherches

- Apprentissage en-ligne
- Changement co-varié (covariate shift)
- Dérive de concept
- Dilemme oubli-mémoire

Grandes directions

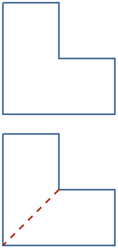
Apprentissage **à partir de flux de données**

Apprentissage en-ligne et transfert

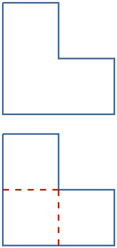
Peut-on prédire qu'il est plus facile d'apprendre dans l'ordre : $XOR(x_1, x_2) \vee x_1$ puis $XOR(x_1, x_2)$ que dans l'ordre $XOR(x_1, x_2)$ puis $XOR(x_1, x_2) \vee x_1$?

Pourquoi l'ordre suivant est-il catastrophique ?
Consigne : découper la figure suivante en n parties superposables.

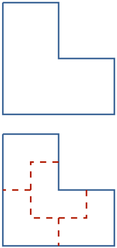
En **2** :



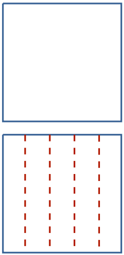
En **3** :



En **4** :



En **5** :



Grandes directions

Apprentissage à partir de flux de données

Apprentissage en-ligne et transfert

Quels outils conceptuels ?

- **Méthodes bayésiennes ou statistiques** ?
- Caractérisation de la **représentation des concepts** ?
  - Complexité algorithmique ?
- Théorie des **espaces de Riemman** (*géométrie de l'information*)
  - Transport parallèle et dérivées co-variantes ?
- ...

Plan

1 Les pionniers

2 Les années 60-70

3 Avènement de l'analyse statistique de l'apprentissage

4 Et maintenant : les grandes questions ?

5 Conclusion

Conclusion

Résumé

1. **50s** : Ivresse et exploration
   - Apprentissage continu par renforcement
2. **60s-70s** : Apprentissage supervisé et méthodes
   Imitation de l'apprentissage humain
   - neurologique / symbolique
   - Induction $\equiv$ généralisation
   - Importance des représentations
3. **80s-90s** : Cadre statistique de l'apprenabilité
   - Abstraction des algorithmes (MRE)
   - ... et des représentations (notion de capacité)
   - Mais présupposé i.i.d. et représentations pauvres
4. **2005 - ...** : Ouverture
   - Tâches plus riches : plus seulement i.i.d., supervisées et représentations pauvres
   - Prise en compte d'autres caractéristiques des algorithmes
   - Nouveaux cadres conceptuels, au-delà du cadre statistique ?

Conclusion

The idea of a learning machine may appear paradoxical to some readers.
Alan Turing (1912 - 1954), 1950.

La prédiction est difficile, surtout lorsqu'il s'agit de l'avenir
Niels Bohr (1885 - 1962).

Conclusion


Apprentissage
artificiel
2e édition
Antoine Cornuéjols est professeur à AgroParisTech. Il enseigne l'apprentissage artificiel dans plusieurs grandes écoles et en Master. Ses recherches portent notamment sur l'apprentissage en ligne, l'apprentissage à partir de flux de données ainsi que sur des applications en bioinformatique et sciences du vivant.
Laurent Miclet est professeur à l'ENSSAT (www.enssat.fr) de Lannion, université de Rennes-I, et responsable du projet CORDIAL de l'IRISA (www.irisa.fr). Il a enseigné l'apprentissage artificiel et la reconnaissance des formes dans plusieurs grandes écoles et en Master. Ses recherches portent en particulier sur l'apprentissage pour le dialogue homme-machine et les technologies vocales.
Les programmes d'intelligence artificielle sont aujourd'hui capables de reconnaître des commandes vocales, d'analyser automatiquement des photos satellites, d'assister des experts pour prendre des décisions dans des environnements complexes et évolutifs (analyse de marchés financiers, diagnostics médicaux...), de fouiller d'immenses bases de données hétérogènes, telles les innombrables pages du Web...
Pour réaliser ces tâches, ils sont dotés de modules d'apprentissage leur permettant d'adapter leur comportement à des situations jamais rencontrées, ou d'extraire des lois à partir de bases de données d'exemples.
Ce livre présente les concepts qui sous-tendent l'apprentissage artificiel, les algorithmes qui en découlent et certaines de leurs applications. Son objectif est de décrire un ensemble d'algorithmes utiles en tentant d'établir un cadre théorique pour l'ensemble des techniques regroupées sous ce terme « d'apprentissage artificiel ».
À qui s'adresse ce livre ?
• Ce livre s'adresse tant aux décideurs et aux ingénieurs qui souhaitent mettre au point des applications qu'aux étudiants de niveau Master 1 et 2 maîtrise en école d'ingénieurs, qui souhaitent un ouvrage de référence sur ce domaine clé de l'intelligence artificielle.
Sommaire
I. Les fondements de l'apprentissage · Première approche de l'induction · Environnement méthodologique · II. Apprentissage par exploration · Induction et relation d'ordre · Programmation logique inductive · Transfert de connaissance · Inférence grammaticale · Apprentissage par évolution · III. Apprentissage par optimisation · Modèles linéaires · Réseaux connexionnistes · Réseaux bayésiens · HMM (modèles de Markov cachés) · Inférence d'arbres · IV. Apprentissage par approximation et interpolation · Méthodes à noyaux · Apprentissage bayésien · Apprentissage par renforcement · V. Au-delà de l'apprentissage supervisé · Combinaisons d'experts · Classification non supervisée et fouille de données. Apprentissage semi-supervisé · Nouvelles tâches et nouvelles questions · Annexes et bibliographie.
Code éditeur : G12471 • ISBN : 978-2-212-12471-2
55 €
EYROLLES
2e édition
Apprentissage artificiel
Algorithmes
A. Cornuéjols
L. Miclet
EYROLLES
2e édition
Antoine Cornuéjols - Laurent Miclet
A
Apprentissage
artificiel
Concepts et algorithmes
Préface de Jean-Paul Haton
EYROLLES