Apprentissage et Décision

Apprentissage et Décision

Objectifs

1. Classification automatique
2. Programmation Dynamique
3. Discrimination Fonctionnelle
4. Connexionisme
5. Statistiques bayésiennes
6. K-ppv
7. Méthodes stochastiques
8. Approches syntaxiques et structurelles

Apprentissage (automatique ou artificiel) :
Machine Learning

Cette notion englobe toute méthode permettant de construire un modèle de la réalité à partir de données.

Il existe deux tendances principales :

- *Celle issue de l'IA qualifiée de symbolique*
- *Celle issue des statistiques qualifiée de numérique*

Fouille de données (Extraction de connaissances à partir des données) :
Data Mining (Knowledge discovery in data)

La fouille de données prend en charge le processus complet d'extraction de connaissances : stockage dans une BD, sélection des données à étudier, nettoyage de ces données, puis utilisation des apprentissages symboliques et numériques afin de proposer des modèles à l'utilisateur, enfin validation des modèles proposés.

Précision vs. Généralisation

Le grand dilemme de l'apprentissage.

La précision est définie par un écart entre une valeur mesurée ou prédite et une valeur réelle. Apprendre avec trop de précision conduit à un "sur-apprentissage", comme l'apprentissage par coeur, pour lequel des détails insignifiants (dûs au bruit) sont appris.

Apprendre avec trop peu de précision, conduit à une surgénéralisation telle que le modèle s'applique même quand l'utilisateur ne le désire pas.

Il existe des mesures de généralisation, et l'utilisateur peut fixer le seuil de généralisation qu'IL juge optimal.

Classification :

La classification (voire Analyse de données) consiste à regrouper des ensembles d'exemples non supervisés en classes.

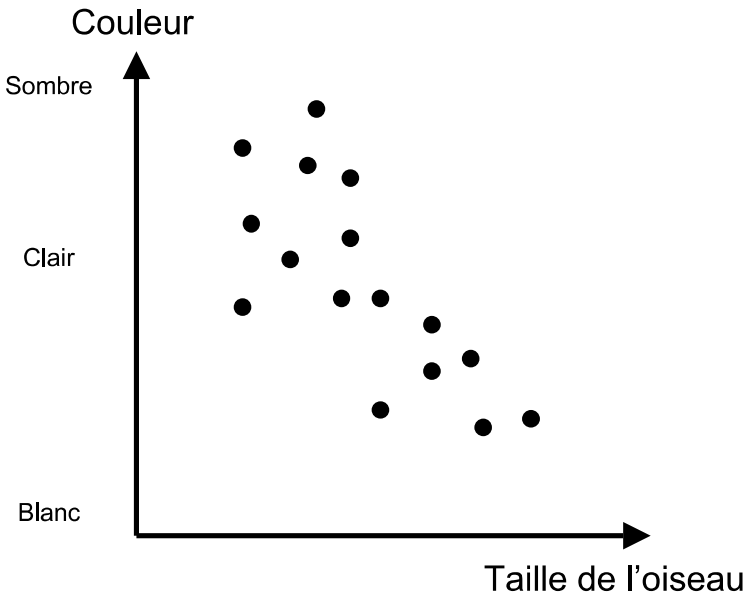
Ces classes sont souvent organisées en une structure (clustering). Si cette structure est un arbre, alors on parle de taxonomie ou de taxinomie (taxonomy).

Il s'agit par exemple de prévoir l'appartenance d'un oiseau observé à la classe "canard" ou "oie".

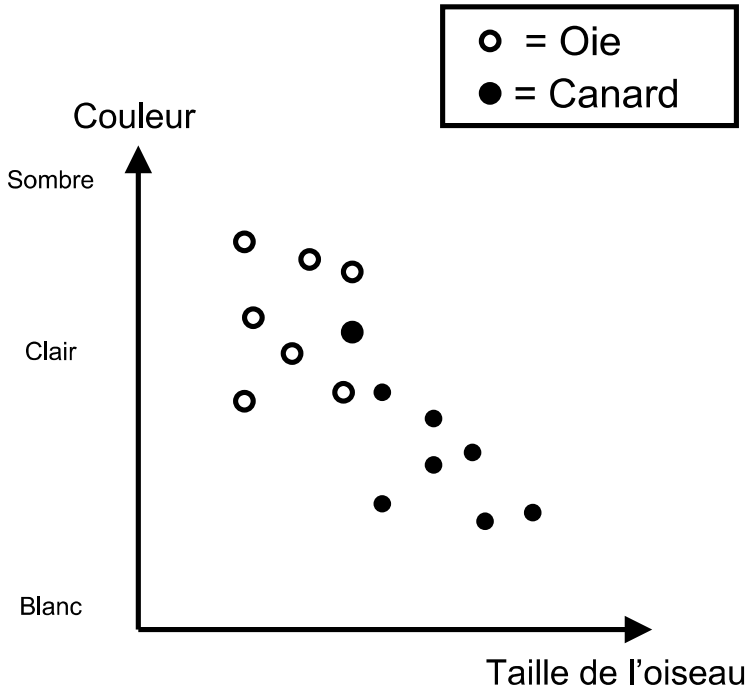
Deux champs industriels de l'Apprentissage :

1. La Reconnaissance des Formes (image, parole, signaux bio-médicaux)
2. La Fouille de Données

Exemples d'apprentissage :

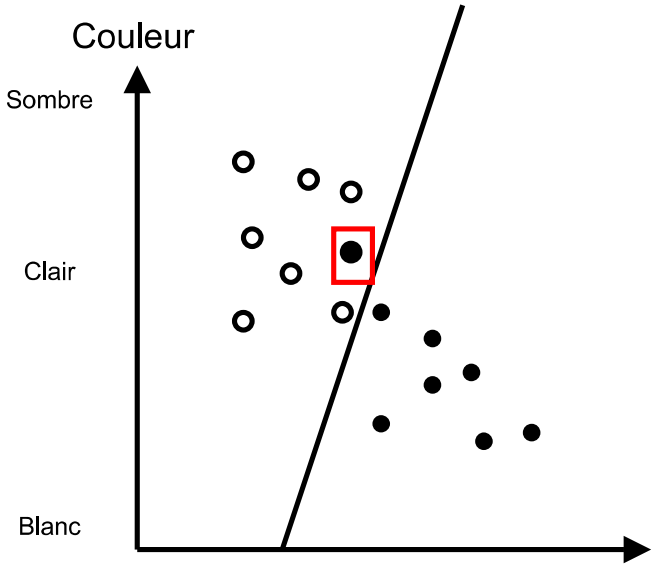


Oiseaux Observés par un débutant

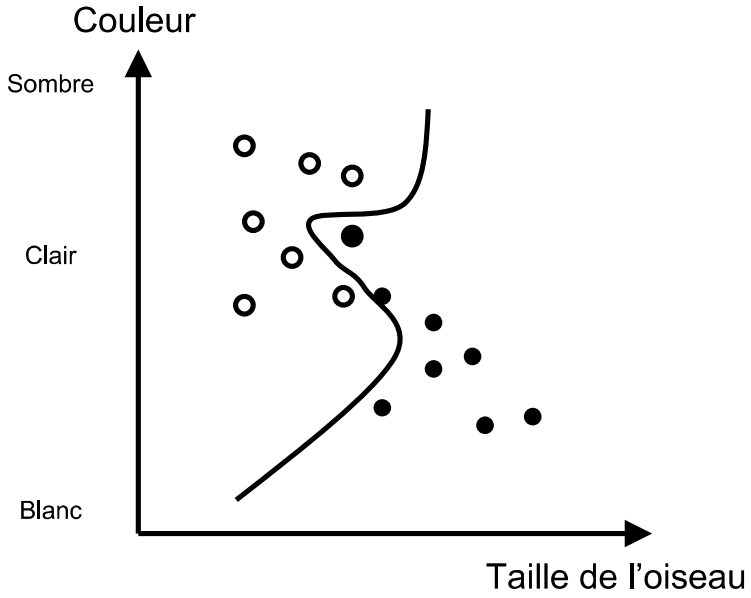


Données Etiquetées par un expert (superviseur)

Exemples d'apprentissage :



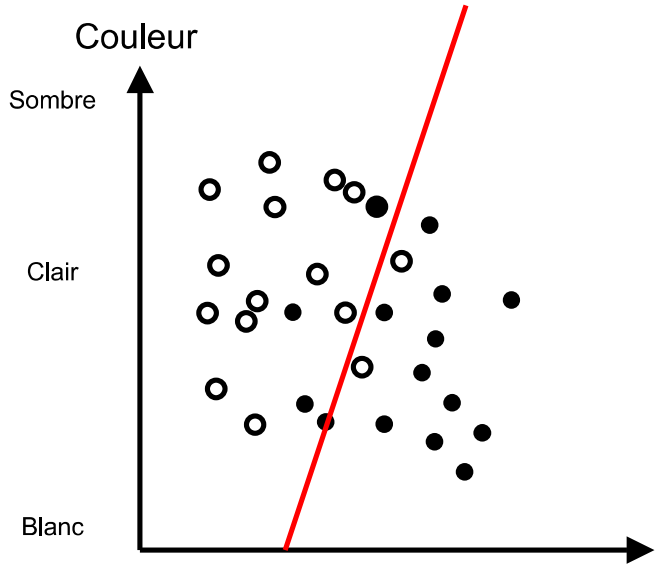
Apprentissage d'une règle de décision simple



Apprentissage d'une règle de décision complexe

Exemples d'apprentissage :

○ = Oie
● = Canard



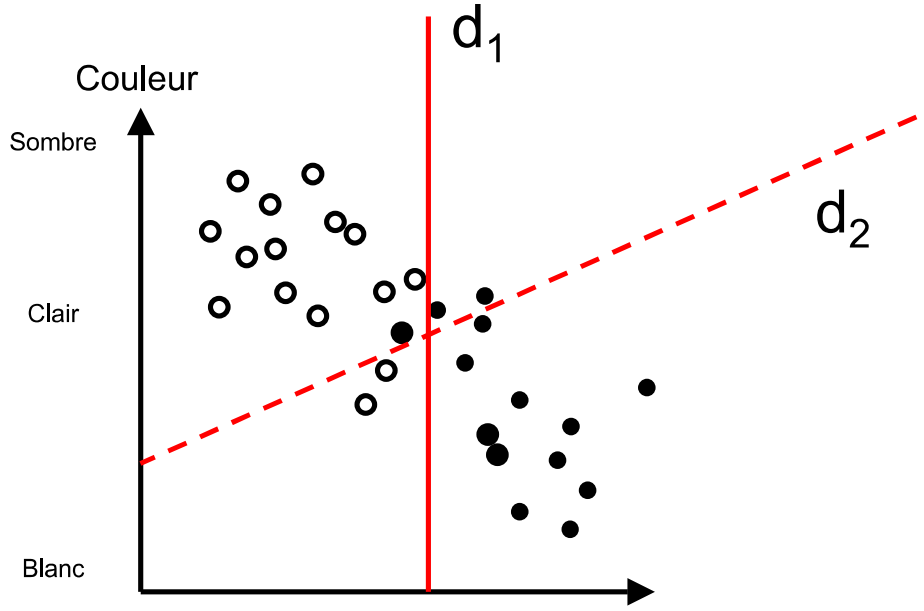
Test en Décision sur d'autres oiseaux à partir de la règle de décision simple

Principe :

définir les **fonctions de discrimination *f*** permettant de séparer **partiellement ou totalement** les classes représentées par les vecteurs paramètres *x* de leurs échantillons. L'ensemble des x exemples représente **l'ensemble d'apprentissage S**.

Objectifs :

- Il s'agit d'apprendre un concept sous la forme géométrique la plus simple : celle d'un hyperplan.
- Apprentissage de surfaces séparatrices linéaires dans un espace de représentation nécessairement numérique.
- Lien avec les Réseaux de Neurones et les SVM



- (d_1) minimise le nombre d'erreurs dans l'ensemble d'apprentissage
- (d_2) minimise le risque bayésien d'erreurs par rapport à l'ensemble d'apprentissage

Hypothèses :

Les classes sont linéairement séparables.

Remarque : hypothèse pas plus injustifiée que la classique hypothèse statistique *a priori* gaussienne

Dans $\Re^n$, une surface linéaire est un hyperplan f :

$$f(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \ldots + \omega_n x_n = W^t . X = 0$$

$$avec \quad X = (1, x_1, x_2, \ldots, x_n)^t = (1, x)^t$$

$$et \quad W = (w_0, w_1, w_2, \ldots, w_n)^t = (w_0, w)^t$$

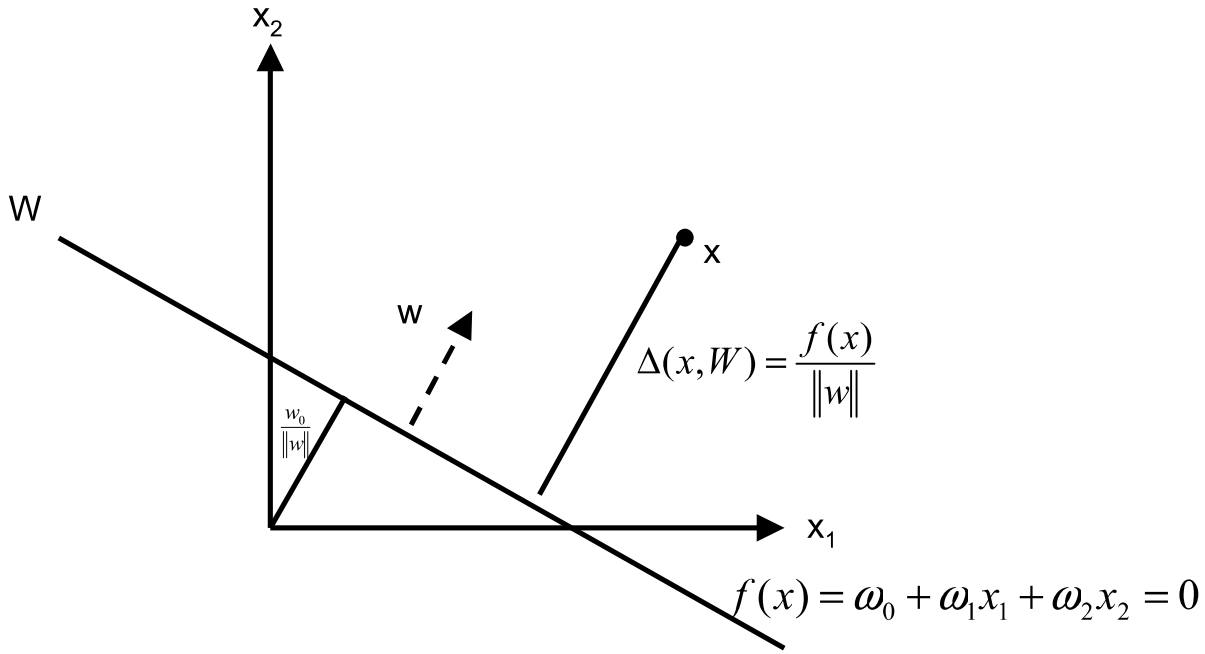
Dans $\Re^2$, une droite est définie par :

$$f(x) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 = 0$$

que l'on va écrire $\quad avec \quad x = (x_1, x_2)^t$

$$f(x) = \omega_0 + w^t x = 0 \qquad et \quad w = (w_1, w_2)^t$$

Remarque : Δ est une distance signée



Apprentissage :

- Déterminer les coefficients W de ces fonctions discriminantes à partir de formes connues

- Algorithme du **Perceptron** : le plus ancien et le plus simple

Cas de 2 classes :

$$f(X) = W^t.X \begin{cases} > 0 & si \quad X \in C_1 \\ < 0 & si \quad X \in C_2 \end{cases}$$

On pose

$$\begin{cases} \forall X \in C_1, & Y = X \\ \forall X \in C_2, & Y = -X \end{cases}$$

Alors

$$W^t.Y > 0 \Leftrightarrow \begin{cases} W^t.X > 0 & si \quad X \in C_1 \\ W^t.X < 0 & si \quad X \in C_2 \end{cases}$$

Cas de 2 classes :

Si on range maintenant les *m* vecteurs *X* comme colonnes d'une matrice *M,* le problème de séparation linéaire revient à la recherche d'un vecteur W dans $\Re^{n+1}$, tel que : $W^t M = B^t$

où B est un vecteur positif inconnu de $\Re^{m+1}$

Comme en général M n'est pas inversible et que les données ne sont pas réellement complètement linéairement séparables, il faut trouver l'hyperplan W le meilleur possible selon par exemple le critère :

$$J(W,B) = \frac{1}{2}\left\|W^t M - B^t\right\|^2$$

Cas de 2 classes :

Or, si W et B minimisent bien J(W,B) : $J(W,B) = \sum_{x \in S \quad mal\ classés\ par\ W} \Delta(x,W)^2$

Ce qui revient à se placer dans le cadre du critère aux moindres carrés.

Donc dans l'hypothèse que les distances entre W et les points mal classés par W sont répartis selon une distribution gaussienne, et que ceux-ci sont les seuls à compter dans le positionnement de W, la minimisation de J(W,B) est la recherche du meilleur hyperplan *a posteriori* au sens bayésien, cad l eplus probable connaissant les données.

Cas de 2 classes :

$$J(W,B)=\frac{1}{2}\left\|W^tM-B^t\right\|^2 \rightarrow \nabla_A J(W,B)=(W^tM-B^t)M^t$$

qui atteint son minimum pour $(W^tM-B^t)=0$

soit $W^t = B^tM^+$

où M^+ est la pseudo-inverse de $M = M^T(MM^T)^{-1}$ (user de la SVD si nécessaire)

Comme on ne connait pas B, l'algorithme doit réaliser une minimisation de J(W,B) sous la contrainte B positif ou nul

Cas de 2 classes :

Méthode globale basée sur une descente de gradient :

Algorithme de Ho et Kashyap

INPUT : B_0 et α positifs quelconques

- $t \leftarrow 0$
- Tant que *critère d'arrêt non satisfait* faire
  - $W_{(t)}^T \leftarrow B_{(t)}^T M^+$
  - $B_{(t+1)}^T \leftarrow B_{(t)}^T + \alpha \lfloor W_{(t)}^T M - B_{(t)}^T \rfloor$
  - $t \leftarrow t+1$
- Fin tant que

OUTPUT : l'hyperplan optimal W

$W^T_{(t)} = B^T_{(t)} M^+$ puis trouver $B_{(t+1)}$ tel que $J(W(t),B(t+1)) \leq J(W(t),B(t))$

or $\nabla_{B_{(t)}} J(W_{(t)}, B_{(t)}) = -2(W_{(t)}^T M - B_{(t)}^T)$ d'où $B_{(t+1)}^T = B_{(t)}^T + \alpha(W_{(t)}^T M - B_{(t)}^T)$

or B>=0, d'où $\lfloor W_{(t)}^T M - B_{(t)}^T \rfloor = 0$ si $W_{(t)}^T M - B_{(t)}^T < 0$. De plus on peut avoir $\alpha_{(t)}$