Apprentissage Artificiel

Fondements et méthodologie de l'apprentissage

Ewa Kijak

ESIR/Univ. Rennes

Sommaire

Sommaire

Premier contact

Qu'est ce que la tâche d'apprentissage **supervisé** ?

- On dispose d'un ensemble d'exemples **étiquetés**.
- A partir de ces exemples, on doit attribuer une étiquette à un nouvel exemple.

example	label
train	
ant	−
bat	+
dolphin	−
leopard	+
sea lion	−
zebra	+
shark	−
mouse	+
chicken	−
test	
tiger	
tuna	
platypus	

Premier contact

Autre exemple jouet

- un enfant, de retour chez lui après l'école, veut savoir s'il peut aller jouer avec ses voisins
- il a une certaine expérience, basée sur les décisions de sa mère des 8 jours précédents
- formellement, ce problème d'apprentissage consiste à trouver une règle de décision binaire à partir de 8 exemples.
- les 8 jours précédents (i.e. les exemples) sont décrits dans la table suivante avec 4 attributs

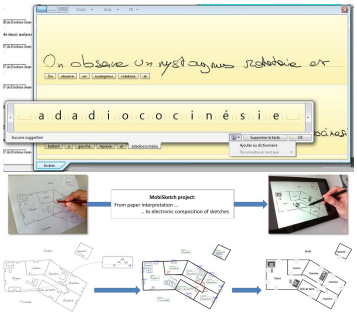
Premier contact

	mes devoirs sont faits ou pas	maman est de bonne humeur	il fait beau	mon goûter est pris	Décision
1	faits	faux	vrai	faux	oui
2	pas faits	vrai	faux	vrai	oui
3	faits	vrai	vrai	faux	oui
4	faits	faux	vrai	vrai	oui
5	pas faits	vrai	vrai	vrai	non
6	pas faits	vrai	faux	faux	non
7	faits	faux	faux	vrai	non
8	faits	vrai	faux	faux	non
today	faits	vrai	faux	faux	?

Premier contact

Exemples d'applications

- Apprentissage pour la navigation
  - Apprentissage de trajets (robots)
- Discrimination
  - Reconnaissance de l'écriture manuscrite, de la parole
  - Identification de locuteur / de signature
  - Reconnaissance de codes postaux, de plaques d'immatriculations



Premier contact

- Apprentissage à mieux jouer
  - S'adapter à l'adversaire
  - Ne pas répéter ses fautes
  - Apprendre à jouer en équipe (équipe de robots)
- Catégorisation/classification
  - Reconnaissance de concepts
  - Classification d'images
- ...

Premier contact

Qui a étiqueté les exemples donné ?
→un "oracle", la "Nature"

Cadre

- tout apprentissage peut être considéré comme l'apprentissage de la représentation d'une fonction
- un exemple est un couple $< \boldsymbol{x}, f(x) >$ où $\boldsymbol{x}$ est la valeur d'entrée et $f(x)$ la valeur de sortie
- étiquette $= f(x)$, où f est une fonction cible inconnue donnée par un oracle
- un exemple $\boldsymbol{x}$ est décrit par un ensemble d'attributs

Premier contact

Objectif

- déterminer cette fonction cible f à partir d'un ensemble d'exemple
- en fait, on ne pourra pas trouver f mais une hypothèse h parmi un ensemble (espace) $\mathcal{H}$ d'hypothèses qui se rapproche le plus de f

⇒apprentissage = problème de recherche dans un espace

Différents types d'inférence

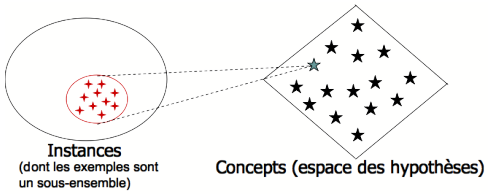
Syllogisme

a) Tout homme est mortel
b) Or Socrate est un homme
c) Donc Socrate est mortel

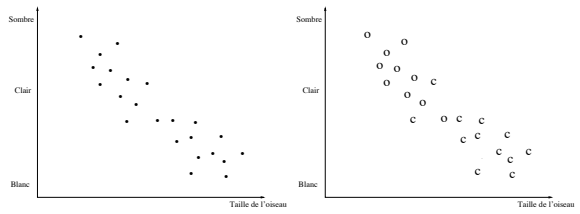
- Inférence déductive : avec a et b on trouve c →la plus simple
- Inférence abductive : avec a et c on trouve b →le diagnostic
- Inférence inductive : avec b et c on trouve a →généralisation

Classification inductive supervisée

Le système d'apprentissage (apprenant) cherche à trouver une description du concept (classifieur) qui « explique » les instances données en exemples.

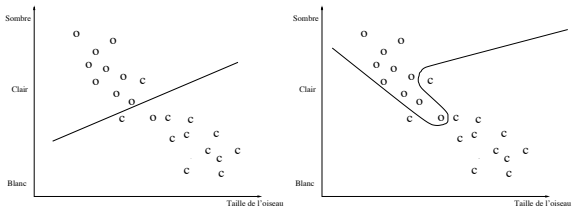


Un exemple ornithologique



O = oie ; C = cygne
(à gauche) dans l'espace de représentation.
(à droite) étiqueté par l'expert.

Un exemple ornithologique



(à gauche) Une règle de décision simple,
(à droite) une règle de décision complexe.

Différents types d'apprentissage

- Supervisé / non-supervisé
  - supervisé : l'apprenant reçoit des exemples comprenant la valeur d'entrée et de sortie
  - notion d'expert pour donner la valeur de sortie des exemples
- Par paquets (batch) / incrémental
  - batch : tous les exemples sont pris en compte dès le début de l'apprentissage
  - incrémental : exemples pris un par un, amélioration de l'hypothèse courante

Problématique de l'apprentissage inductif

4 questions à se poser en apprentissage à partir d'exemples :

1. choix de la description des exemples
2. choix de l'espace d'hypothèse
3. algorithme d'apprentissage (parcours de l'espace d'hypothèses)
4. évaluation de l'apprentissage

Sommaire

La représentation des exemples

- **exemple** = couple $(\boldsymbol{x}, c)$, où :
  - $\boldsymbol{x} \in \mathcal{X}$ est la description de l'objet à classer,
  - $c \in C$ représente la classe de $\boldsymbol{x}$.
- L'espace $\mathcal{X}$ est appelé l'**espace de représentation** ou l'**espace des instances**.

La représentation des exemples

Les éléments de $\mathcal{X}$ peuvent être détaillés comme un ensemble d'**attributs**, ce qui se note : $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_d)$.

- Un attribut $\boldsymbol{x}_i$ qui prend ses valeurs dans $\mathbb{R}$ est dit **numérique**.
- Un attribut $\boldsymbol{x}_i$ qui prend ses valeurs dans $\mathbb{B}$ est **binaire**.
- Un attribut $\boldsymbol{x}_i$ qui est une suite d'éléments d'un alphabet Σ est une **séquence**

On parle :

- d'**apprentissage numérique** quand $\mathcal{X} = \mathbb{R}^d$
- d'**apprentissage symbolique** quand $\mathcal{X} = \mathbb{B}^d$ ou $\mathcal{X} = \Sigma^*$.

Description des exemples

Choisir les bons attributs

- discriminants
- permettant un apprentissage efficace

Différents types d'attributs

- quantitatifs (discrets ou continus)
- qualitatifs (booléens, symboliques)

→Influence le choix de la technique d'apprentissage

L'exemple du cygne

Description d'un oiseau par l'ensemble des valeurs suivantes :

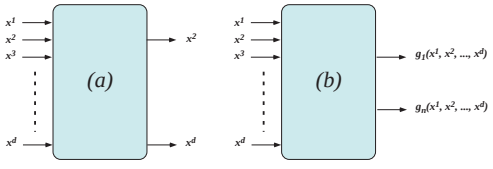
- sa *taille* (attribut numérique) ;
- son *sexe* (attribut binaire) ;
- la *couleur* de son bec (attribut nominal : une couleur-type parmi une dizaine sans relation d'ordre entre elles) ;
- son *genre* (dans la hiérarchie des naturalistes, cette variable est au-dessus de l'*espèce* et au-dessous de la *famille*).

Exemple : Cygne Chanteur (*Cygnus Cygnus L.*) caractérisé par l'observation des attributs :
$\mathcal{X} = (taille = 152cm, sexe = male, couleurdubec = jaune, genre = Anatidae)$

Description binaire, exemple

	vole	*a des poils*	*pond des œufs*	
cisticole	1	0	1	oiseau
ornithorynque	0	1	1	mammifère
rhinolophe	1	1	0	mammifère
percnoptère	1	0	1	oiseau
apteryx	0	0	1	oiseau
céphalorhynque	0	0	0	mammifère

Sélection et extraction d'attributs



(a) Sélection d'attributs
retient les attributs les plus pertinents parmi les d attributs de l'espace d'entrées.

(b) Extraction d'attributs
transforme les attributs de l'espace d'entrée, ici par une fonction de combinaison g, pour en construire de nouveaux en nombre restreint.

Classe, concept

- **classe** de l'exemple = un entier c dans :

$$\mathcal{C} = \{\omega_1, \omega_2, ..., \omega_C\}$$

où C désigne le nombre de classes possibles.

- $C = 2$ ⇒on utilise le mot : **concept**
  - partage de l'espace de représentation en deux parties : l'une où il est vérifié, l'autre où il est invalidé.
  - On note : $\mathcal{C} = \{VRAI, FAUX\}$ (ou parfois $\mathcal{C} = \{+, -\}$)
  - On appelle **contre-exemples** les données classées *FAUX* (on garde le mot d'**exemples** pour les autres).

Sommaire

La simplicité : un exemple instructif

Problème Quel est le chiffre a qui prolonge la séquence :

$$1\,2\,3\,5\ \ldots\ a$$

La simplicité : un exemple instructif

Solution(s) Quelques réponses valides :

La simplicité : un exemple instructif

Généralisation Il est facile de démontrer ainsi que n'importe quel nombre est une prolongation correcte de n'importe quelle suite de nombres.

Le rasoir d'Occam

La solution la meilleure fait intervenir le moins de concepts.
Dans l'exemple mathématique, la solution $a = 8$ est préférable : elle ne nécessite que le concept d'addition.
D'une manière générale, le principe du *rasoir d'Occam* (c. 1425) conduit à choisir, pour une valeur explicative égale, la solution la plus simple. On doit ignorer les entités non informatives.

"Pluralitas non est ponenda sine necessitate"

"Non sunt multiplicanda entia praeter necessitate."

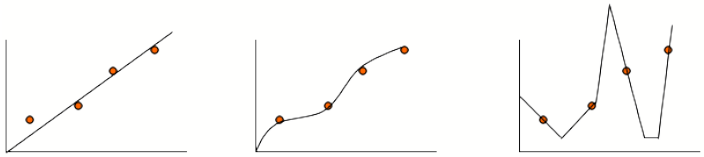
La nécéssité d'un biais

- On peut toujours expliquer n'importe quelle solution si se place dans un cadre assez complexe.
- On doit donc se fixer une famille de concepts à l'intérieur de laquelle on cherchera la meilleure explication des données.
- C'est ce qu'on appelle se donner un *biais d'apprentissage*.
- Le biais dépend de la représentation des données.
- Le compromis simplicité / efficacité devra guider le choix du biais.

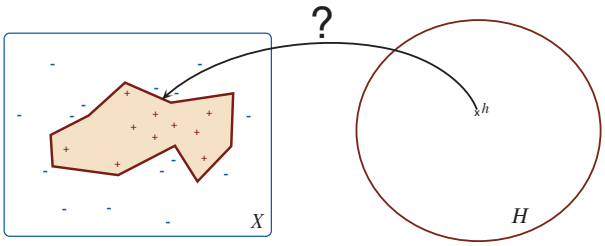
Biais inductif

Utilité
un biais permet de ne pas apprendre "n'importe quoi", surtout en non-supervisé

Un biais courant : le MDL →rasoir d'Occam
Minimum Description Length : préférer la solution la plus simple (description la plus compacte)



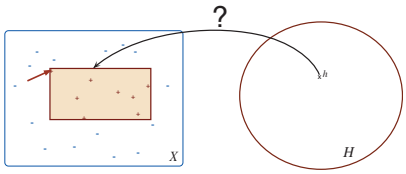
L'espace des hypothèses



Introduction d'un espace d'hypothèses $\mathcal{H}$:

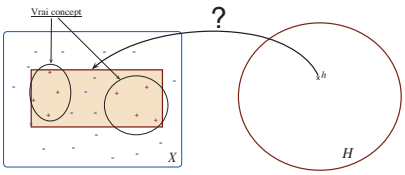
- 1 point de $\mathcal{H}$ ⇔ une hypothèse
- 1 hypothèse ⇔ une partition de l'espace des entrées $\mathcal{X}$.

Langage des hypothèses et généralisation



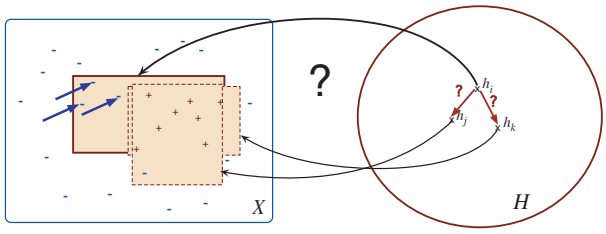
Le langage de représentation des hypothèses $\mathcal{L}_{\mathcal{H}}$ correspond aux parties de $\mathcal{X}$ qui sont des rectangles. Dans ce cas, la donnée du point '+' flèché implique que tous les points inscrits dans le rectangle dont il délimite un angle sont de classe '+'.

Langage des hypothèses et généralisation



Le langage de représentation des hypothèses $\mathcal{L}_{\mathcal{H}}$ correspond aux rectangles et la vraie partition vraie de la Nature, correspondant aux exemples positifs, est représentée par les deux patatoïdes. Dans ce cas, il est impossible d'approximer correctement le concept cible à l'aide d'une hypothèse de $\mathcal{H}$.

L'exploration de l'espace des hypothèses

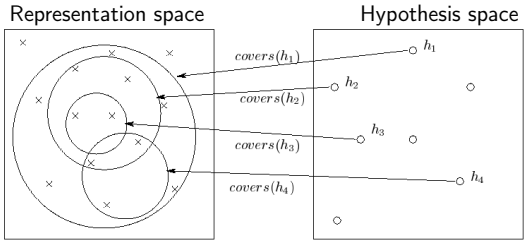


Si l'hypothèse courante h_t est insatisfaisante, il faut que l'apprenant cherche une nouvelle hypothèse dans $\mathcal{H}$: où doit-il chercher ?

Couverture

Définition

- On appelle couverture d'une hypothèse h l'ensemble des exemples "expliqués" par h
- La couverture est donc une relation de $\mathcal{X}$ dans $\mathcal{H}$

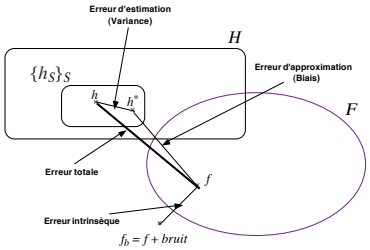


Les types d'erreurs en apprentissage

Les sources d'erreur en apprentissage par généralisation sont de trois types :

- Les données peuvent être bruitées, fausses, mal étiquetées
  →*erreur intrinsèque*
- L'espace $\mathcal{H}$ où l'on cherche une hypothèse est trop restreint
  →*erreur d'approximation, biais inductif*
- L'algorithme de recherche dans $\mathcal{H}$ ne fonctionne pas bien
  →*erreur d'estimation, variance*

Les différentes erreurs d'apprentissage



- *f est la fonction cible. f_b est celle qui a produit $\mathcal{S}$.*
- *h^* est la meilleure pour $\mathcal{S}$ et $\mathcal{H}$.*
- *h est la fonction trouvée par l'algorithme d'apprentissage sur les données $\mathcal{S}$.*

Le compromis "biais / variance"

Quand $\mathcal{H}$ est restreint :

- La meilleure solution dans $\mathcal{H}$ est facile à trouver (variance ↘)
- Mais elle peut être éloignée de la vraie solution (biais ↗)

Quand $\mathcal{H}$ est large :

- La meilleure solution dans $\mathcal{H}$ est difficile à trouver (variance ↗)
- C'est dommage, car elle est sans doute plus proche de la vraie solution (biais ↘)

La tâche d'apprentissage

- L'apprenant cherche dans l'espace $\mathcal{H}$ une fonction h qui approxime au mieux la réponse désirée de l'oracle.
- La qualité de son travail est définie par l'*espérance de perte* sur les situations possibles dans $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$.
- Pour chaque entrée $\boldsymbol{x}_i$ et réponse de l'oracle $\boldsymbol{u}_i$, on mesure une *perte* ou *coût* $l(\boldsymbol{u}_i, h(\boldsymbol{x}_i))$, coût d'avoir pris la décision $\boldsymbol{y}_i = h(\boldsymbol{x}_i)$ quand la réponse désirée était $\boldsymbol{u}_i$.
- L'espérance de coût, ou *risque réel* est alors :

$$R_{Réel}(h) = \int_{\mathcal{Z}=\mathcal{X}\times\mathcal{U}} l(\boldsymbol{u}, h(\boldsymbol{x}))dF(\boldsymbol{x}, \boldsymbol{u})$$

- Le problème de l'induction est donc de chercher à minimiser le risque réel inconnu à partir du seul échantillon d'apprentissage $\mathcal{S}$.

Quelques fonctions de risque

- **Classification :** $\boldsymbol{u}_i$ et $h(\boldsymbol{x}_i)$ sont des numéros de classe

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

- **Régression :** $\boldsymbol{u}_i$ et $h(\boldsymbol{x}_i)$ sont des réels

$$l(\boldsymbol{u}_i, h(\boldsymbol{x}_i)) = (\boldsymbol{u}_i - h(\boldsymbol{x}_i))^2$$

Le principe inductif

- Le *principe inductif* prescrit ce que doit vérifier la fonction h recherchée, en fonction à la fois de la notion de risque et de l'échantillon d'apprentissage observé $\mathcal{S} = \{(\boldsymbol{x}_1, \boldsymbol{u}_1), (\boldsymbol{x}_2, \boldsymbol{u}_2), ..., (\boldsymbol{x}_m, \boldsymbol{u}_m)\}$, dans le but de minimiser le risque réel.
- Il faut le distinguer de la *méthode d'apprentissage* (ou algorithme) qui décrit sa réalisation effective.
- Pour un principe inductif donné, il y a de nombreuses méthodes d'apprentissage qui résultent de choix différents pour régler les problèmes computationnels.

Le principe inductif ERM

On choisit l'hypothèse minimisant le risque empirique (*Empirical Risk Minimization*)

- Le risque empirique est la perte moyenne mesurée *sur l'échantillon d'apprentissage* $\mathcal{S}$:

$$R_{emp}(h) = \frac{1}{m}\sum_{i=1}^{m} l(\boldsymbol{u}_i, h(\boldsymbol{x}_i))$$

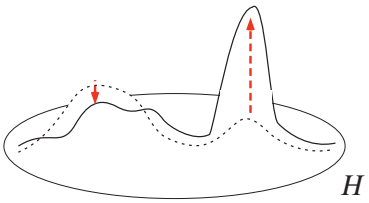
- L'idée est que l'hypothèse qui s'accorde le mieux aux données, si elles sont représentatives, décrit correctement le monde en général.

Le principe inductif bayésien

Choisir l'hypothèse la plus probable étant donné $\mathcal{S}$.

- On définit une distribution de probabilité sur l'espace des fonctions hypothèse.
- La connaissance du domaine préalable à l'apprentissage s'exprime sous la forme d'une distribution de probabilité *a priori* sur les hypothèses.
- L'échantillon d'apprentissage est alors considéré comme une information modifiant la distribution de probabilité sur $\mathcal{H}$.
- On peut choisir l'hypothèse la plus probable *a posteriori Maximum A Posteriori* (*MAP*) ou adopter une hypothèse composite résultant de la moyenne des hypothèses pondérée par leur probabilité *a posteriori* (*"vraie" approche bayésienne*).

Le principe inductif bayésien



L'espace des hypothèses $\mathcal{H}$ est supposé muni d'une densité de probabilités *a priori*. L'apprentissage consiste à modifier cette densité en fonction des exemples d'apprentissage.

Sommaire

Evaluation de l'apprentissage

Problèmes
- comment savoir que les hypothèses retenues se comporteront bien avec de nouvelles données ?
- comment éviter l'over-fitting et l'apprentissage par coeur ?

Solutions
- séparer les données en jeu d'entraînement et jeu de test (cross-validation, leave-one out)

Echantillon

Un **échantillon** est un ensemble fini d'exemples.
On fait l'hypothèse indispensable que les exemples sont tirés de manière aléatoire et indépendante selon les C distributions de probabilités $P(\mathcal{X}, \mathcal{C})$.

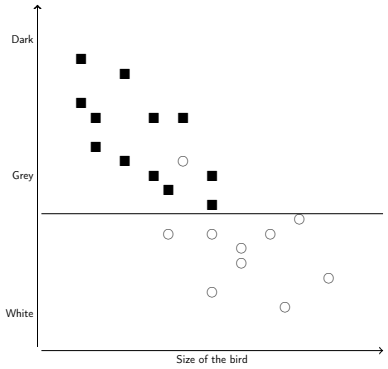
On distinguera dans la suite trois sortes d'échantillons :

- d'**apprentissage**
- de **validation**
- et de **test**.

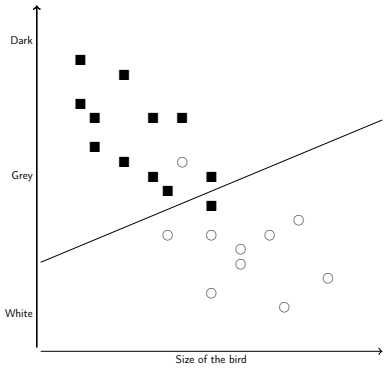
Apprentissage d'une règle de classification

- Une **règle de classification** ou de **décision** h est une application définie sur $\mathcal{X}$ à valeurs dans $\mathcal{C}$.
- **Apprentissage** d'une règle de classification :
  1. choix d'un ensemble $\mathcal{H}$ de règles possibles ;
  2. trouver une règle h dans $\mathcal{H}$, à partir de l'examen d'un échantillon d'**apprentissage**.
- Echantillon de **validation** (optionnel) : utilisé comme "contrôleur" dans l'algorithme d'apprentissage.
- Echantillon de **test** (indispensable) : utilisé pour vérifier la qualité de l'apprentissage réalisé.

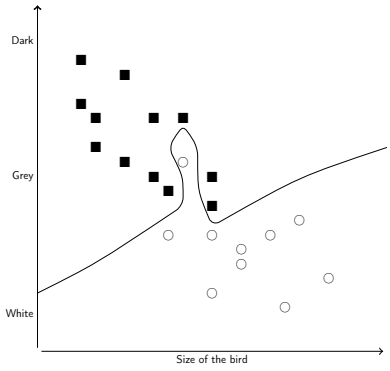
Apprentissage d'une règle de classification simple



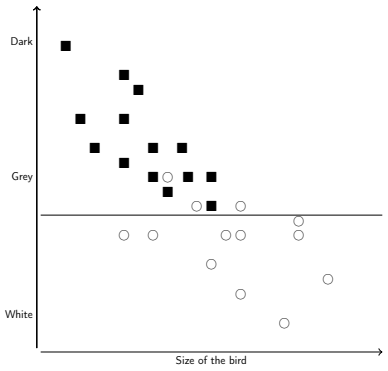
Apprentissage d'une autre règle de classification simple



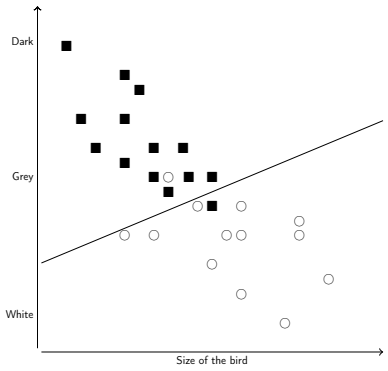
Apprentissage d'une règle de classification complexe



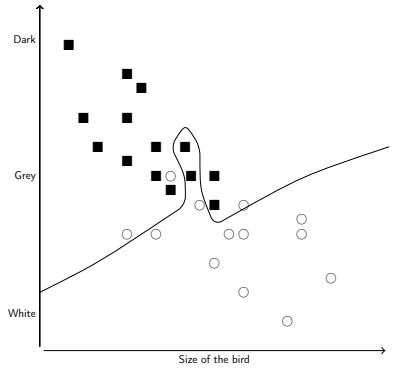
Test de la première règle de classification simple



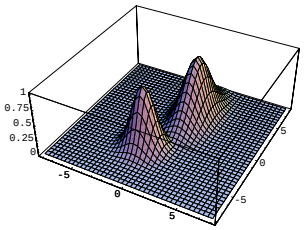
Test de la deuxième règle de classification simple



Test de la troisième règle de classification

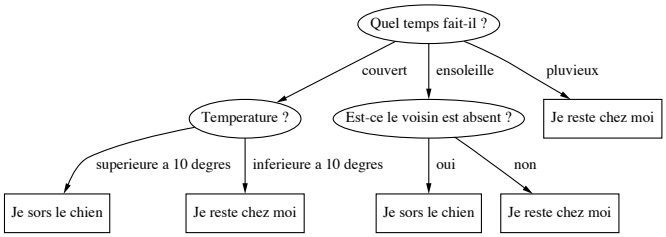


Apprentissage de distribution de probabilités



Deux distributions de probabilité correspondant à deux classes d'objets bi-dimensionnels.

Apprentissage d'arbres de décision



Taux d'erreur apparent et réel

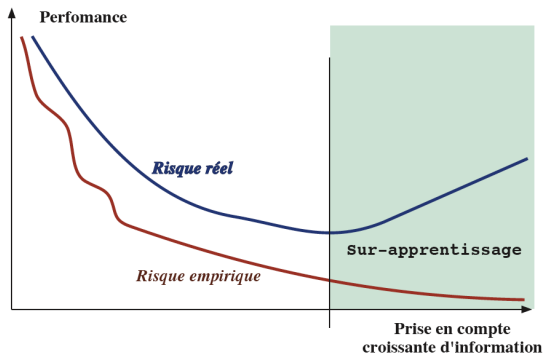
Soit m la taille de l'ensemble d'apprentissage
Soit m_{err} le nombre d'exemples de cet ensemble qui sont mal classés par une certaine règle h choisie dans $\mathcal{H}$.

Le **taux d'erreur apparent** ou **risque empirique** de h est :

$$f_{err}(h) = m_{err}/m$$

La **probabilité d'erreur** ou **taux d'erreur réel** ou **risque réel** de h, que l'on note $P_{err}(h)$, est la probabilité que h classe mal un exemple tiré selon $P(\mathcal{X}, \mathcal{C})$.

Sur-apprentissage



Estimation du taux d'erreur apparent et du taux d'erreur réel

Division de l'ensemble des exemples en deux parties :

- le premier est utilisé pour l'apprentissage de la règle h (**ensemble d'apprentissage**)
  →calcul du risque empirique
- le second sert à sa validation a posteriori (**ensemble (d'exemples) de test**)
  →estimation du risque réel

Mesure plus fine du taux d'erreur apparent : **matrice de confusion**.

Matrice de confusion

Définition

La **matrice de confusion** $M_h(i,j)$ d'une règle de classification est une matrice $C \times C$ dont l'élément générique donne le nombre d'exemples de test de la classe i qui ont été classés dans la classe j.

Estimation du risque réel

L'estimation $\widehat{P_{err}}$ de la probabilité d'erreur exacte P_{err} du classificateur appris
= somme des termes non diagonaux de M, divisée par la taille de l'ensemble de test, si toutes les erreurs sont considérées comme également graves.

Matrice de confusion

Un peu de vocabulaire

Pour un concept A et la matrice de confusion :

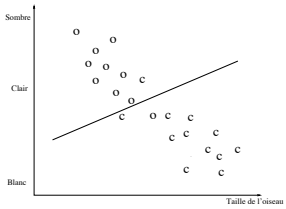
Classés \ Exacts	A	$\overline{A}$
A	a	b
$\overline{A}$	c	d

Le concept A est :

- *VRAI* exactement $(a+c)$ fois dans les données
- et *FAUX* $(b+d)$ fois.

Le classificateur trouve $(a+b)$ fois A comme *VRAI* et $(c+d)$ fois A comme *FAUX*.

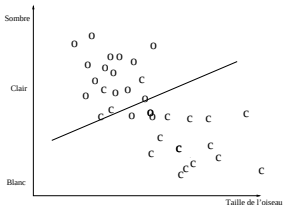
a = VP	nombre de *vrais positifs* ou de *bonnes acceptations* ou de *détections* ou de *reconnaissances*
b = FP	nombre de *faux positifs* ou de *fausses alarmes*
c = FN	nombre de *faux négatifs* ou de *faux rejets* ou de *non-détections* (*miss*)
d	nombre de *vrais négatifs* ou de *rejets corrects*
$\frac{b+c}{a+b+c+d}$	taux d'*erreur*
$\frac{a+d}{a+b+c+d}$	taux de *reconnaissance* ou *puissance*
$\frac{a}{a+b}$	*précision*
$\frac{a}{a+c}$	*rappel*, ou *sensibilité*
$\frac{d}{b+d}$	*spécificité*



Une règle de décision simple pour séparer les oies des cygnes.

Matrice de confusion d'apprentissage. Erreur empirique : $\frac{1+1}{9+1+1+11} \simeq 9\%$

	O	C
O	9	1
C	1	11



Le test de la règle simple sur d'autres oiseaux.

Matrice de confusion de test. $\widehat{P_{err}} = \frac{3+4}{14+3+4+13} \simeq 24\%$.

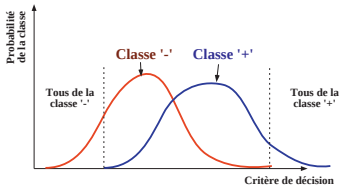
	O	C
O	14	4
C	3	13

Courbes ROC

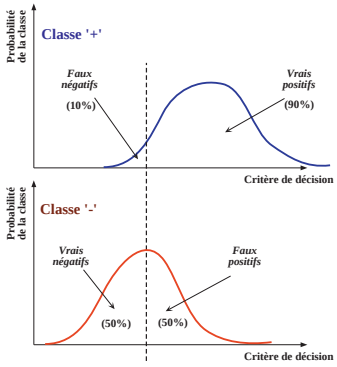
2 types d'erreurs :

- *faux positifs* : probabilité d'accepter l'hypothèse alors qu'elle est fausse
- *faux négatifs* : probabilité de rejeter l'hypothèse alors qu'elle est vraie

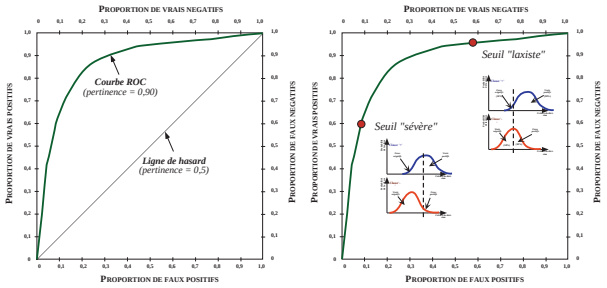
Comment arbitrer entre ces types d'erreurs ?



Courbes ROC



Courbes ROC

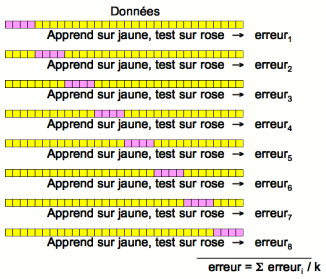


La validation croisée et le *leave-one-out*

Cas d'application :

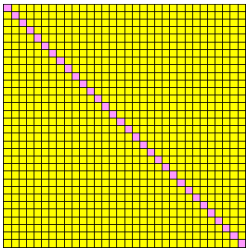
- peu de données
- meilleure estimation de l'erreur réelle

⇒**validation croisée** (*cross-validation*)



La validation croisée et le *leave-one-out*

L'estimateur *leave-one-out* est la limite : on fait m apprentissages sur $m-1$ données et on teste sur la dernière.



La classification obtenue estime avec précision la probabilité d'erreur de la méthode d'apprentissage.