

## Préambule

# Apprentissage Automatique

MASTER MICR  
Céline Rouveirol

1

- **Enseignant** : Céline Rouveirol
  - Email : [celine.rouveirol@lipn.univ-paris13.fr](mailto:celine.rouveirol@lipn.univ-paris13.fr)
  - Téléphone : 01 49 40 35 78
- **Organisation**
  - Cours: le lundi après midi de 13h45 à 15h15
  - TD/TP : le lundi après-midi 15h30-17h avec H. Soldano
- **Evaluations**
  - Contrôle continu: étudier un article, présenter un résumé et faire un exposé (10 minutes)
  - Examen final décembre 2006.

2

## Plan

- Définition et motivation de l'apprentissage automatique
- Quelques disciplines autour de l'apprentissage automatique
- Apprentissage supervisé
- Exemple de représentation symbolique
- Quelques applications de l'apprentissage symbolique
- Quelques outils (weka pour les TPs)
- Quelques références
- Les arbres de décision et l'algorithme ID3.

3

## Sources de ce cours

- Transparents de Aomar Osmani, cours Apprentissage Symbolique, 2006-2007
- Transparents sur l'apprentissage inductif (chapitre 18) de Artificial Intelligence: A Modern Approach (Second Edition), de Stuart Russell and Peter Norvig
- Notes de cours d'Apprentissage à partir d'exemples, de François Denis et Rémi Gilleron, Université de Lille 3.

4

## Apprentissage Naturel

Peut être défini comme une modification du comportement résultant d'une interaction répétée avec l'environnement.

Quelques notions de bases en psychologie expérimentale :

- Apprentissage par cœur et conditionnement
- Apprentissage par renforcement (punition-récompense)
- Apprentissage de concepts
- Résolution de problème
- Généralisation/spécialisation
- La nécessité d'un biais
- Le biais de simplicité (ou rasoir d'Occam)

5

## Apprentissage artificiel, définitions

- Wikipedia: « As a broad subfield of **artificial intelligence**, **machine learning** is concerned with the design and development of **algorithms** and techniques that allow computers to "learn".
- Wikipedia: « L'apprentissage automatique fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine (au sens large) d'**évoluer** grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques ».
- At a general level, there are two types of learning: **inductive**, and **deductive**. Inductive machine learning methods extract rules and patterns out of massive data sets.

6

## Apprentissage artificiel, définitions

- «L'apprentissage dénote des changements dans un système qui ... lui permet de faire la même tâche plus efficacement la prochaine fois» Herbert Simon
- «L'apprentissage permet des modifications utiles dans notre cerveau» Marvin Minsky

7

## Apprentissage artificiel, définitions

Du point de vue des machines, on dira qu'une machine apprend dès lors qu'elle change sa structure, son programme ou ses données en fonction de données en entrée ou de réponses à son environnement de sorte à ce que ses performances futures deviennent meilleures.

**L'objectif de l'apprentissage artificiel est de concevoir des programmes pouvant s'améliorer automatiquement avec l'expérience.**

8

## Pourquoi l'apprentissage artificiel?

- Certaines tâches ne sont bien définies que via un ensemble d'exemples : on est capable de spécifier des couples d'entrées/sorties mais pas de relations explicites et concises entre les entrées et les sorties.
- Il est possible que les quantités importantes de données renferment des corrélations et des relations importantes que les méthodes d'AA permettent de découvrir (fouille de données)
- Les machines produites par l'homme peuvent ne pas fonctionner sur tous les environnements car certains aspects de ces environnements peuvent être inconnus au moment de la conception des dites machines.
- La quantité de connaissances disponibles a propos de certaines situations sont telles que le cerveau humain ne puisse les expliciter. L'AA peut permettre de mieux exploiter ces connaissances.
- L'environnement change constamment. L'AA permet aux machines de s'adapter aux changements sans refaire systématiquement la conception après chaque changement de l'environnement.

9

## Pourquoi l'apprentissage artificiel?

Un des dix plus grands enjeux du XXI ème siècle (MIT Technology review):

- Compréhension et amélioration de l'apprentissage humain (ex: instruction assistée par ordinateur)
  - Découverte de nouvelles connaissances ou structures (ex: fouille de données)
  - Paramétrage automatique de systèmes complexes et/ou dynamiques
- Applications de l'apprentissage:
    - Traitement du langage naturel (fouille de textes), reconnaissances des formes, moteurs de recherche, diagnostic médical, bioinformatique, biochimie, finance (détection de fraude (à la carte bancaire), ...) , analyse des marchés boursiers, jeux, robotique, ...

10

## Quelques questions auxquelles l'AA tente de répondre

- Quels sont les algorithmes qui approximent au mieux les fonctions à apprendre ?
- Comment le système d'apprentissage est conçu pour s'approcher de la solution au problème ?
  - En jouant sur le nombre d'exemples d'apprentissage
  - En complexifiant les hypothèses de représentation
- Comment la modélisation du problème influence la précision de l'apprentissage
  - Données bruitées
  - Plusieurs sources de données
- Quelles sont les limites théoriques de l'apprenabilité ?
- Comment les connaissances a priori peuvent aider l'apprentissage ?

11

## Quelques disciplines autour de l'AA

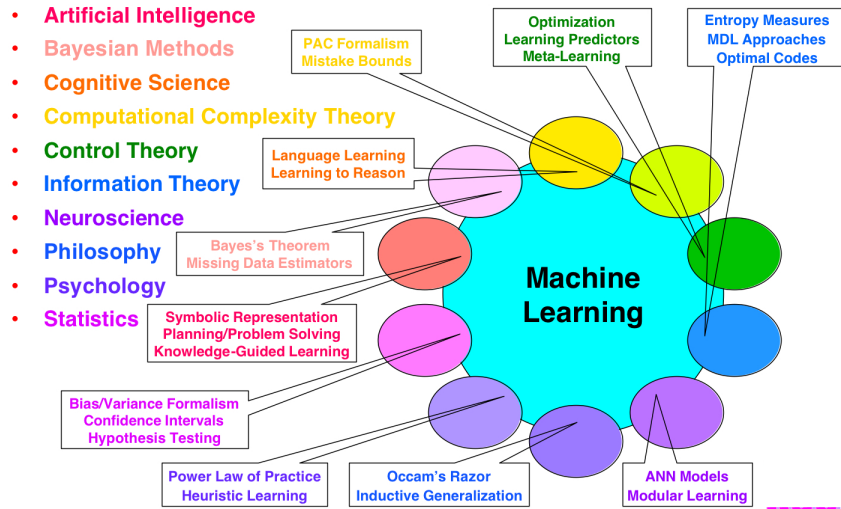
L'apprentissage artificiel se situe au carrefour de plusieurs disciplines dont

- les statistiques : Test d'hypothèses, ...
- l'intelligence artificielle : représentation symbolique, ...
- la philosophie : généralisation induction, ...
- la théorie de l'information : entropie
- la biologie : réseaux de régulation, ...
- les sciences cognitives : apprentissage du langage, ...
- la théorie de la complexité : PAC, ...
- etc.

D'après Tom Mitchell : le meilleur moyen d'approcher l'apprentissage artificiel est de le voir à travers ces différentes perspectives.

12

## Quelques disciplines autour de l'AA



13

## Principaux paradigmes d'apprentissage

- Apprentissage par coeur (mémoire associative)
- Analogie (correspondances entre deux représentations différentes)
- Apprentissage par renforcement (récompense/punition à la fin d'une séquence d'actions)

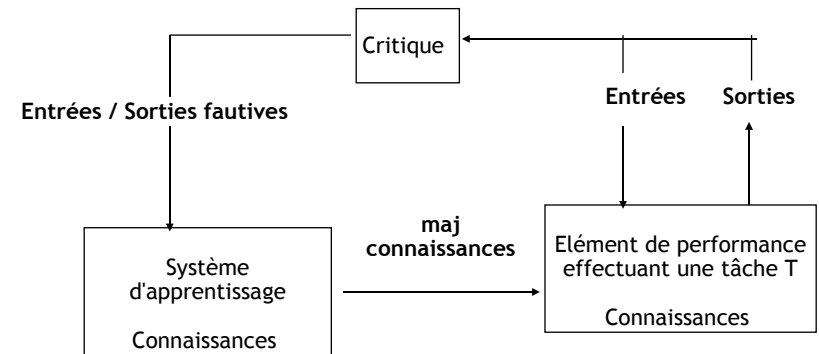
14

## Principaux paradigmes d'apprentissage

- Classification/Catégorisation (ACP, nuées dynamiques, cartes de Kohonen...)
- Découverte (règles d'association...)
- Apprentissage supervisé de fonctions (analyse statistique des données, séparation linéaire/non-linéaire (SVM, réseaux de neurones), k-NN...)

15

## Apprentissage supervisé: une modélisation



On dit qu'un système d'apprentissage **améliore** un élément de performance effectuant une tâche T, si la tâche T, d'après la mesure d'évaluation du critique s'effectue mieux avant qu'après apprentissage

16

## Apprentissage supervisé: exemple

- Entrée du système d'apprentissage : observations + classe (attribut booléen ou discret)
- Sorties du système d'apprentissage: un ensemble de règles permettant de discriminer les classes entre elles
- But : classifier de nouvelles observations.

### Application : apprentissage de règles de diagnostic

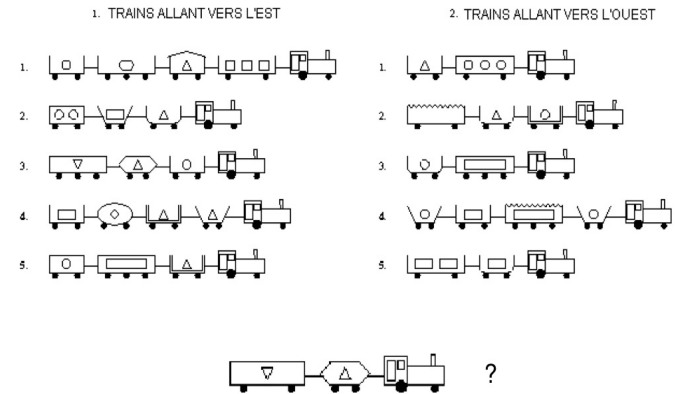
Élément de performance : outil automatique de détection de pannes (un systèmes à base de connaissances).

Entrées du système d'apprentissage : les situations non reconnues, associées à la panne correcte

Mesure d'évaluation : pourcentages de pannes reconnues

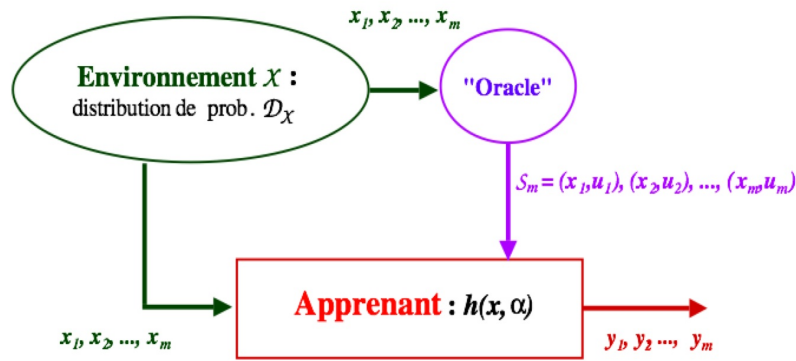
17

## Apprentissage supervisé: problème jouet



18

## Apprentissage supervisé, plus formellement



19

## Apprentissage supervisé

- $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$  avec une dépendance fonctionnelle entre  $x_i \in X$  et  $y_i \in Y$
  - Problème : trouver  $h \in H / \forall i, h(x_i) = y_i$
- sous-domaine de l'AS caractérisé par le domaine de définition de  $Y, X, H$

20

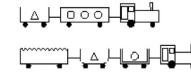
## Systèmes commerciaux (2001)

- $Y$  :
  - continu => régression
  - discret => classement ou **apprentissage de concept** si  $|Y| = 2$
- $X$  : espace des exemples ou des instances
  - vecteur dans  $\mathbb{R}^n$  => apprentissage numérique
  - représentation logique (ordre 0, ordre 1, non classique) => apprentissage symbolique
  - mixte => numérico-symbolique

21

## Exemple de représentation symbolique

- langage attribut-valeur



	#r1	forme1	toit1	#ch1	ch1	#r2	forme2	toit2	#ch2	ch2
t1	2	rect	V	3	circ	2	carré	F	1	tri
t2	2	d_carré	F	1	circ	2	en_U	F	1	tri

22

## Apprentissage supervisé

- $H$  : défini en accord avec  $X$ 
  - apprentissage numérique : SVM, réseaux de neurones, k-NN...
  - apprentissage symbolique : arbres de décision, ensemble de règles, programmes logiques, représentations graphiques...

23

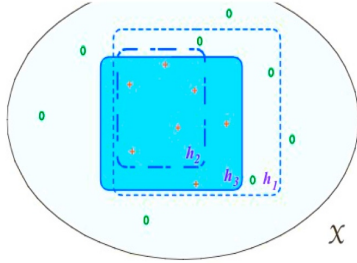
## Système commerciaux (2001)

- SPSS/Clementine (18% du marché et 9% symbolique (AD règles d'association))
- Megaputer (6% symbolique).
- Angoss (3% dont 1% symbolique (AD))
- Business Object (3% symbolique (AD))
- IBM Iminer (3%)

24

## Apprentissage de concepts : terminologie

- Étant donné :
  - l'ensemble d'exemples  $S = S^+ \cup S^-$
  - l'espace d'hypothèse  $H$
- Trouver  $h \in H$  tel que  $h$  soit cohérente avec  $S$



25

## Références

- «Apprentissage Artificiel» A. Cornuejols et L.Miclet, 2002
- «Machine Learning» Tom Mitchell, 1997
- «Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations» I.Witten et E. Frank, 1999
- «Element of Machine Learning» Pat Langley, 1996
- «Artificial Intelligence: A Modern Approach» S. Russel P.Norvig, 1995

26

## Apprentissage inductif

Dans sa forme la plus simple : apprendre une fonction à partir d'exemples

$f$  est la **target fonction cible**

Un **exemple** est une paire  $(x, f(x))$

Problème: trouver une **hypothèse**  $h$   
telle que  $h \approx f$   
étant donné l'ensemble d'apprentissage  $E$

Ceci est un modèle très simplifié de l'apprentissage, car :

- il ignore tout ce qui est connaissance a priori
- il fait l'hypothèse que tous les exemples sont donnés

Erreur apparente: Soit  $E$  un échantillon d'apprentissage  $h$  une hypothèse, l'**erreur apparente** de  $h$  est sur:

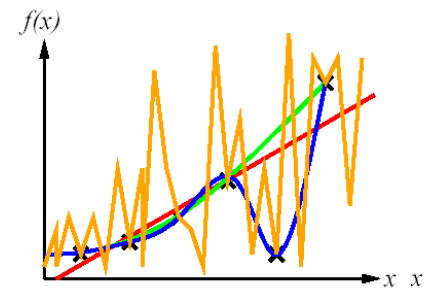
$$|\{x \in S \mid h(x) \neq f(x)\}| / |S|$$

Quand  $|S|$  tend vers l'infini, l'erreur apparente tend vers l'erreur réelle de classification.

27

## Méthodologie en Apprentissage Inductif

- Construire/ajuster  $h$  pour qu'elle s'accorde avec  $f$  sur l'ensemble d'apprentissage
- Par exemple, ajustement de courbes:
- Rasoir d'Ockham: pour un même performance sur l'ensemble d'apprentissage; préférer l'hypothèse la plus simple



28

## Arbre de décision (Quinlan, 83)

Ancêtre des systèmes d'apprentissage du type TDIDT : Top-Down Induction of Decision Trees

Données : un tableau d'observations

Nom	Cheveux	Taille	Poids	Lotion	Résultat
Sarah	blonde	moyen	légère	non	cp de soleil
Dana	blonde	grand	moyenne	oui	RAS
Alex	brune	petit	moyenne	oui	RAS
Annie	blonde	petit	moyenne	non	cp de soleil
Emily	rousse	moyen	lourde	non	cp de soleil
Pete	brune	grand	lourd	non	RAS
John	brune	moyen	lourd	non	RAS
Katie	blonde	petit	légère	oui	RAS

Problème :

Pourquoi certains attrapent un coup de soleil? Comment prédire le résultat pour une nouvelle personne ?

29

## Première solution

Rechercher la personne ayant les mêmes caractéristiques que celle étudiée, et prendre le même résultat.

Pbs :

- peu de chances de trouver l'exemple exact dans les données : ici, il y a 3 valeurs possibles pour les 3 premiers attributs -> nombre de combinaisons =  $3 * 3 * 3 * 2 = 54$

Dans le tableau, il y a 8/54 chances de trouver l'exemple qui coïncide.

- valable pour la classification d'une nouvelle instance seulement

Affaiblissement de cette solution: définir une procédure d'appariement partiel pour toute nouvelle instance présentée = Raisonnement par cas

Pbs :

- Difficulté pour définir une telle procédure (notion de pertinence des attributs vis à vis du problème à résoudre).

- Quels exemples conserver (notion de prototype).

30

## Arbre de décision

Deuxième solution : utiliser un arbre d'identification, c'est-à-dire, un arbre de décision dont chaque nœud représente un attribut et chaque branche une modalité du nœud dont il est issu.

Approche « diviser pour régner »

Chaque feuille de l'arbre conclut sur une classe unique.

Quels attributs sélectionner afin de construire l'arbre ? A priori, on souhaite que le nombre de tests en moyenne pour classifier un nouvel exemple soit minimal (borne sup. = le nombre d'attributs)

31

## Arbres de Décision

- Résolution d'un problème à l'aide de questions/réponses
- Représentation de n'importe quelle fonction booléenne à entrée discrète, avec extension aux fonctions à sortie nominale, continue
- Intelligible pour l'expert

32

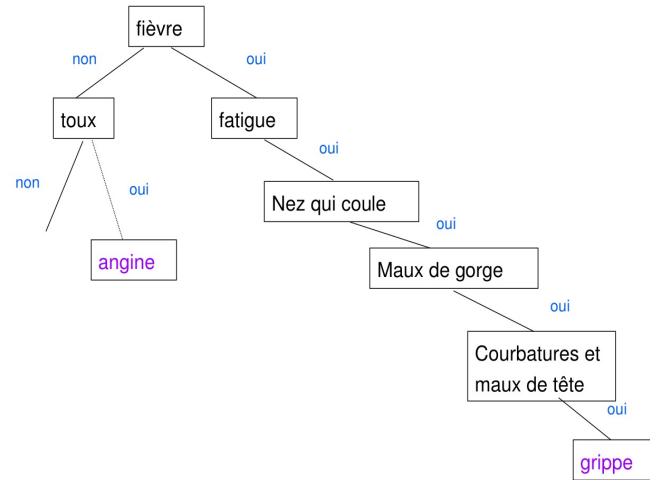


## Un exemple : Diagnostic Médical

- Apparition soudaine de fièvre élevée
- Le patient est fatigué
- Rhinorrhée (nez qui coule)
- Toux
- Douleurs à la gorge
- Enrouement, douleurs dorsales, des membres et céphalées

33

## Un Arbre de Décision



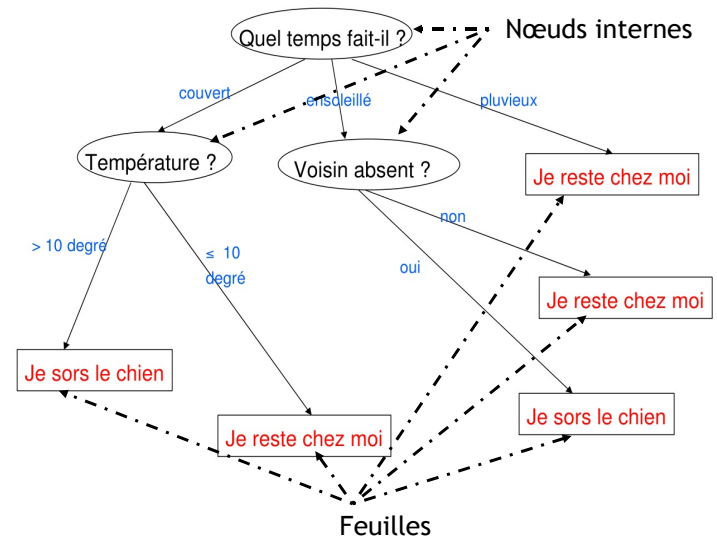
34

## Un exemple : Faut-il sortir le chien ?

- Attributs :
  - le temps qu'il fait : attribut nominal = {pluvieux, ensoleillé, couvert}
  - température extérieure : attribut numérique
  - voisin parti avec son chat : attribut booléen
- Décision à prendre :
  - sortir ou non le chien

35

## Un Arbre de Décision



36

## Méthodes de construction

- Parcours exhaustif de l'espace des arbres
- taille exponentielle en le nombre d'attributs  $n$  et le nombre de valeurs par attribut  $v$

n	v	Arbres possibles
4	2	30
6	2	72385
8	2	$18.10^{18}$

- Construction du plus petit arbre
- NP-difficile

37

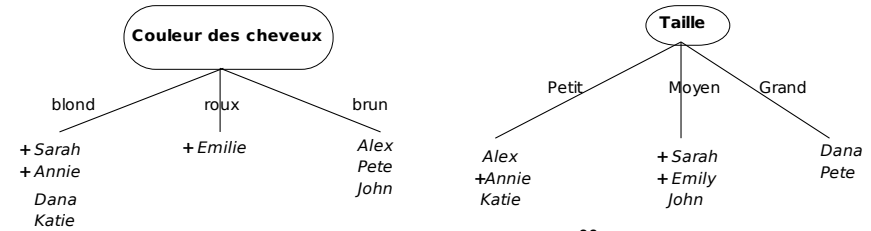
## Arbre de décision (ID3 Quinlan 1983- C4.5 Quinlan 1993)

Recherche descendante récursive (on part de la racine de l'arbre qui contient tous les exemples + et -).

Comment choisir si :

- un nœud est **terminal**
- le meilleur test à appliquer, sinon

Exemple: les candidats "Couleur des cheveux" et "Taille" séparent les exemples de la façon suivante :



```
function DTL(examples, attributes, default) returns a decision tree
if examples is empty then return default
else if all examples have the same classification then return the classification
else if attributes is empty then return MODE(examples)
else
    best ← CHOOSE-ATTRIBUTE(attributes, examples)
    tree ← a new decision tree with root test best
    for each value vi of best do
        examplesi ← {elements of examples with best = vi}
        subtree ← DTL(examplesi, attributes - best, MODE(examplesi))
        add a branch to tree with label vi and subtree subtree
    return tree
```

38

## Propriétés de la fonction de choix d'un test

Principe : On doit comparer des tests, il faut définir une fonction permettant de mesurer le **degré de mélange** des exemples de plusieurs classes à chaque nœud de l'arbre.

La mesure recherchée doit permettre de sélectionner le test minimisant le « désordre » dans les nouveaux nœuds obtenus après application du test : le test choisi doit **séparer** au mieux les exemples de différentes classes (positifs / négatifs).

Ce que l'on veut : une formule qui donne une valeur grande si le test produit des groupes homogènes et une valeur faible s'il conduit à des groupes hétérogènes.

→ mesure d'entropie et critère de Gini

39

## Théorie de l'information

- Pour implémenter **Choose-Attribute** dans l'algorithme DTL
- Entropie au nœud N portant sur l'ensemble d'exemples E (Shannon et Weaver, 49):  

$$\text{entropie}(E) = \sum_{i=1} -P(v_i) \log_2 P(v_i)$$
 où  $P(v_i)$  est la probabilité pour un exemple de E d'appartenir à la classe  $i$ .

- Pour un nœud N contenant  $p$  exemples positifs et  $n$  exemples négatifs:

$$\text{entropie}(N) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

- Exemples :
  - 1 seule classe : entropie(E) =  $-1 \log_2(1) - 0 \log_2(0) = 0$
  - deux classes, contenant le même nombre d'éléments, entropie(E) =  $-0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1$  désordre maximum
  - pour l'exemple, entropie(E) =  $-5/8 \log_2(5/8) - 3/8 \log_2(3/8) = 0.42 + 0.53 = 0.95$

40

## Gain d'information

- Un attribut  $A$  divise l'ensemble d'apprentissage au noeud  $E$  en les sous-ensembles  $E_1, \dots, E_v$  selon leur valeur pour l'attribut  $A$ , où  $A$  a  $v$  valeurs distinctes

- Entropie croisée ou réduction de l'entropie par l'attribut  $A$ :

$$entropie_{croisée}(A) = \sum_{i=1}^v \frac{p_i + n_i}{p + n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

- Gain d'information (GI)

$$IG(E, A) = entropie(E) - entropie_{croisée}(A)$$

- Choisir l'attribut avec le plus grand gain d'information

41

## Gain d'information (suite)

### Exemple :

$$\begin{aligned} IG(E, Taille) &= \text{prob}(\text{petit}) * \text{entropie}(E, \text{petit}) + \\ &\quad \text{prob}(\text{moyen}) * \text{entropie}(E, \text{moyen}) + \\ &\quad \text{prob}(\text{grand}) * \text{entropie}(E, \text{grand}) \\ &= \frac{3}{8} [ -\frac{1}{3} * \log_2(\frac{1}{3}) - \frac{2}{3} * \log_2(\frac{2}{3}) ] + \\ &\quad \frac{3}{8} [ -\frac{1}{3} * \log_2(\frac{1}{3}) - \frac{2}{3} * \log_2(\frac{2}{3}) ] + \\ &\quad \frac{2}{8} [ -1 * \log_2(1) - 0 * \log_2(0) ] \\ &= 0.344 + 0.344 + 0 = 0.69 \end{aligned}$$

42

## Exemple

- \* Recherche de l'attribut minimisant l'entropie croisée

$E$  = ensemble des 8 exemples

$IG(E, Taille) = 0.69$ ;

$IG(E, Cheveux) = 0.5$ ;

$IG(E, Poids) = 0.94$ ;

$IG(E, Lotion) = 0.61$

-> Choix de l'attribut Cheveux

- \* Recherche de l'attribut permettant de discriminer les 4 exemples non discriminés par Cheveux :

$E$  = ensemble des 4 personnes

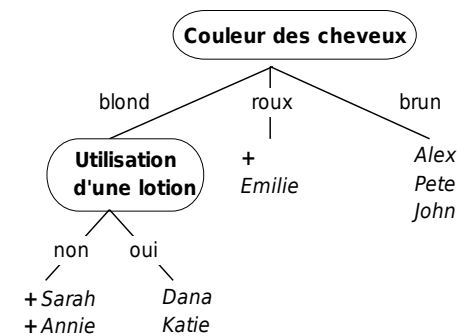
$IG(E, Taille) = 0.5$ ;

$IG(E, Poids) = 1$ ;

$IG(E, Lotion) = 0$  -> Choix de l'attribut Lotion

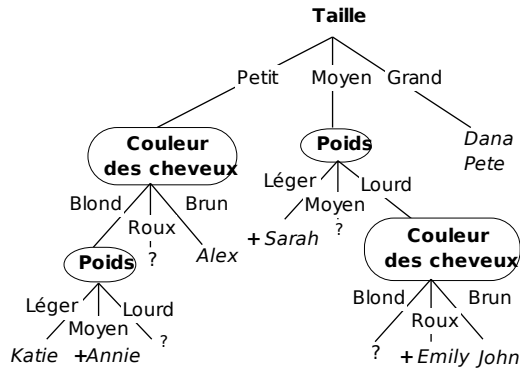
43

## Arbre obtenu



44

Autre arbre possible



- Les attributs avec un gros facteur de branchement vont être favorisés par ce test : les sous-ensembles ont d'avantage de chance d'être homogènes si l'attribut a de nombreuses valeurs -> sur-apprentissage
- Solution: gain ratio (voir TD)

Cart (Breiman et al., 1984)

Critère de Gini

- Inférence d'arbre de décision binaires
- Critère de sélection d'un test: critère de Gini (voir transp. Suivant)
- Arrêt de l'expansion de l'arbre si  $Gini(E) \leq g_0$  ou  $N(E) \leq n_0$
- On assigne une feuille à la valeur de classe majoritaire

$$GINI(E) = 1 - \sum_j [p(j|E)]^2$$

où  $p(j|E)$  est la probabilité pour un exemple de E d'appartenir à la classe j

Si E est séparé par l'attribut A en k classes  $C_1 \dots C_k$

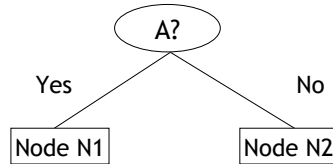
$$GINI_A = \sum_{i=1}^k \frac{n_i}{n} GINI(C_i)$$

Gain (réduction de l'impureté)

$$Gain(E, A) = gini(E) - gini_A(E)$$

## Erreur de classification Vs. Index de Gini

	Parent
C1	7
C2	3
<b>Gini = 0.42</b>	



$$\begin{aligned} \text{Gini}(N1) &= 1 - (3/3)^2 - (0/3)^2 \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Gini}(N2) &= 1 - (4/7)^2 - (3/7)^2 \\ &= 0.490 \end{aligned}$$

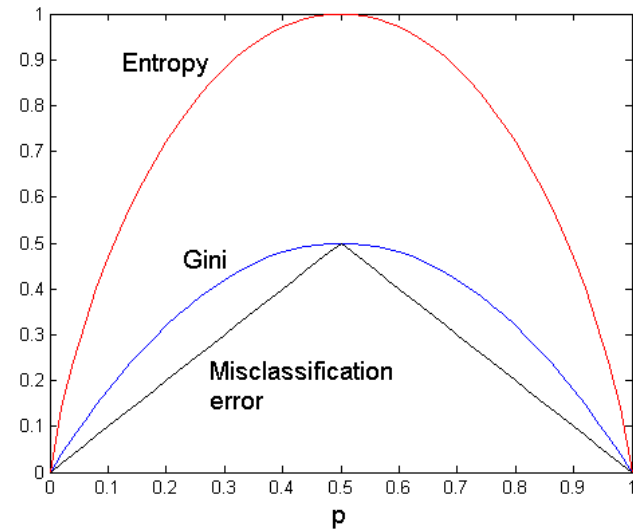
$$\begin{aligned} \text{Gini(Children)} &= 3/10 * 0 \\ &+ 7/10 * 0.49 \\ &= 0.343 \end{aligned}$$

	N1	N2
C1	3	4
C2	0	3
<b>Gini=0.361</b>		

L'index de Gini baisse de .42 à .343 alors que l'erreur de classification reste à 30%.

49

## Comparaison



## Autres mesures de distances entre distributions

- Beaucoup d'autres mesures ont été proposées dans la littérature, par exemple, le test du  $\chi^2$ 
  - Soit E : n exemples appartenant à C classes, chaque classe ayant  $n_j$  éléments.
  - Soit a un attribut booléen. Le test a=vrai sépare chaque classe j en  $l_j$  (resp.  $r_j$  points) : les exemples de la classe j ayant l'attribut a à vrai (resp. à faux).
  - On mesure la pertinence du test en mesurant la différence pour chaque partition / tirage aléatoire dans E.

$$\chi^2(C|a) = \sum_{j=1}^C \left( \frac{l_j - (l * n_j / n)}{\sqrt{l * n_j / n}} \right)^2 + \left( \frac{r_j - (r * n_j / n)}{\sqrt{r * n_j / n}} \right)^2$$

$$\text{avec } l = \sum_{j=1}^C (l_j) \quad \text{et } r = \sum_{j=1}^C (r_j)$$

Plus la valeur du test de  $\chi^2(C|a)$  est faible, moins l'attribut a est pertinent.

51

## ID3 : De l'arbre à un ensemble de règles

Pour chaque chemin dans l'arbre d'identification, en partant de la racine vers les feuilles, prendre la conjonction de chaque test sur ce chemin comme antécédent de la règle et la classification associée à la feuille comme conséquence de la règle.

Exemple : à partir de l'arbre trouvé, contenant 4 feuilles, on obtient 4 règles correspondantes :

- Si  
Alors                    la personne a des cheveux blonds et si elle utilise une lotion elle ne prend pas de coup de soleil
  
- Si  
lotion  
Alors                    la personne a des cheveux blonds et si elle n'utilise pas une lotion elle prend un coup de soleil
  
- ...

Etape supplémentaire : simplification et optimisation du système de règles. On teste si l'on peut éliminer un antécédent de la règle sans changer le comportement de la règle sur les exemples.

- 1- Si  
Alors                    la personne utilise une lotion elle ne prend pas de coup de soleil

52

- **Avantages**

- ✓ Système ancien, commercialisé
- ✓ Complexité algorithmique faible : (cas pire :  $O$  (Exemple \* Attribut<sup>2</sup>))
- ✓ Nombreuses extensions existantes permettant :
- ✓ Incrémentalité: **ID5, ...**  
**Post-élagage statistique : C4, ASSISTANT 86, ...**  
**Nombres réels : Newld, ...**

- **Inconvénients**

- ✓ Hypothèse de non-corrélation entre attributs (les nœuds ne peuvent tester qu'un seul attribut)
- ✓ Pas de chaînage entre les règles (règles peu lisibles et moins bien indexées)
- ✓ Impossibilité de classer un nouvel exemple dont certains attributs sont inconnus ou imprécis

- L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Belmont, CA : Wadsworth.
- A. Cornuéjols et L. Miclet. : Apprentissage Artificiel : Concepts et Algorithmes, Eyrolles, 2002.
- J.R. Quinlan : « Learning efficient classification procedures and their application to chess end game », in R.S. Michalski, J.G. Carbonell and T.M. Mitchell (Eds.), Machine Learning : an Artificial Intelligence Approach. San Mateo, Morgan Kaufman, 1983.
- J.R. Quinlan : « Induction of Decision Trees », Machine Learning, 1 (1), 81-106, 1986.
- J.R. Quinlan : « C4.5. Programs for Machine Learning », San Mateo, Morgan Kaufman, 1993 .