APPRENTISSAGE ARTIFICIEL: ARBRE DE DÉCISIONS

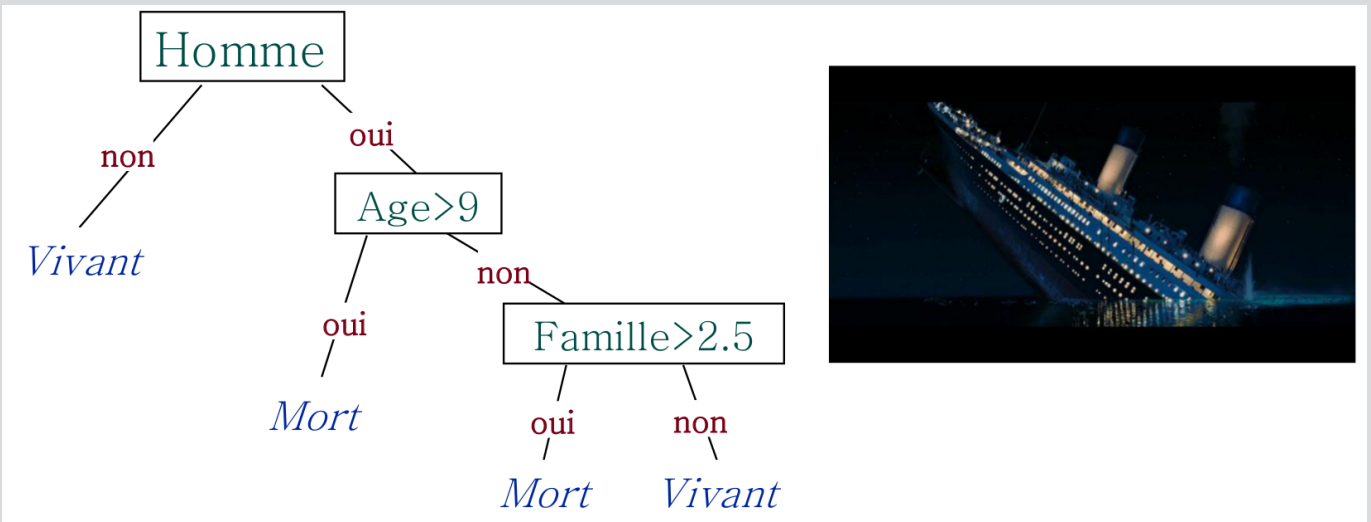
M. Serrurier
IRIT, Toulouse, France

December 8,2015

ARBRES DE DÉCISION
// EXEMPLES

- Les arbres de décision sont des classifieurs pour des instances représentées dans un formalisme attribut/valeur
  - les noeuds de l'arbre testent les attributs
  - Il y a une branche pour chaque valeur de l'attribut testé
  - Les feuilles spécifient les classes (deux ou plus)

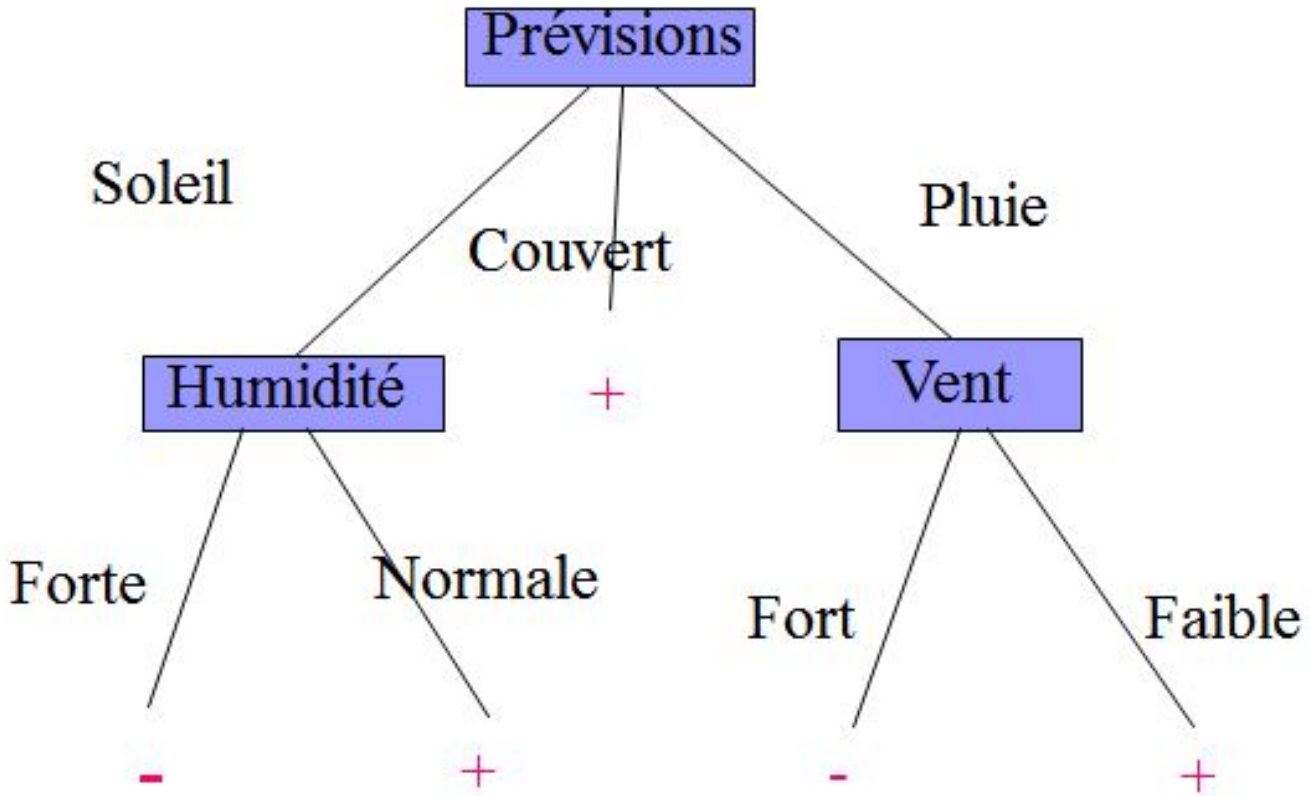
exemple



ARBRES DE DÉCISION
// EXEMPLE

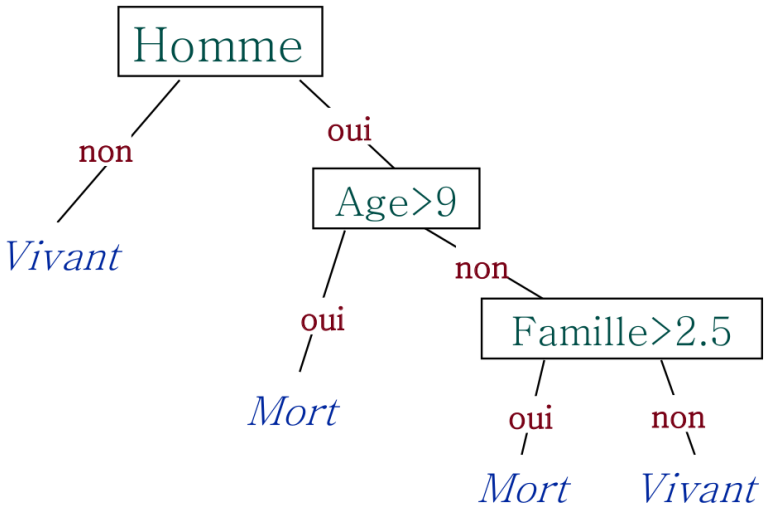
Cas	Prévisions	Température	Humidité	Vent	**Sport**
E1	Soleil	Chaude	Elevée	Faible	-
E2	Soleil	Chaude	Elevée	Fort	-
E3	Couvert	Chaude	Elevée	Faible	+
E4	Pluie	Douce	Elevée	Faible	+
E5	Pluie	Froide	Normale	Faible	+
E6	Pluie	Froide	Normale	Fort	-
E7	Couvert	Froide	Normale	Fort	+
E8	Soleil	Douce	Elevée	Faible	-
E9	Soleil	Froide	Normale	Faible	+
E10	Pluie	Douce	Normale	Faible	+
E11	Soleil	Douce	Normale	Fort	+
E12	Couvert	Douce	Elevée	Fort	+
E13	Couvert	Chaude	Normale	Faible	+

ARBRES DE DÉCISION
// EXEMPLES



- SI Prévisions= Soleil et Humidité = Forte ALORS Sport=
- SI Prévisions= Couvert ALORS Sport=
- ....

ARBRES DE DÉCISION
// EXERCICE



- Transformer cet arbre en un ensemble de règles
- Simplifier l'ensemble de règles
- Peut-on retransformer cet ensemble de règles en un arbre ?

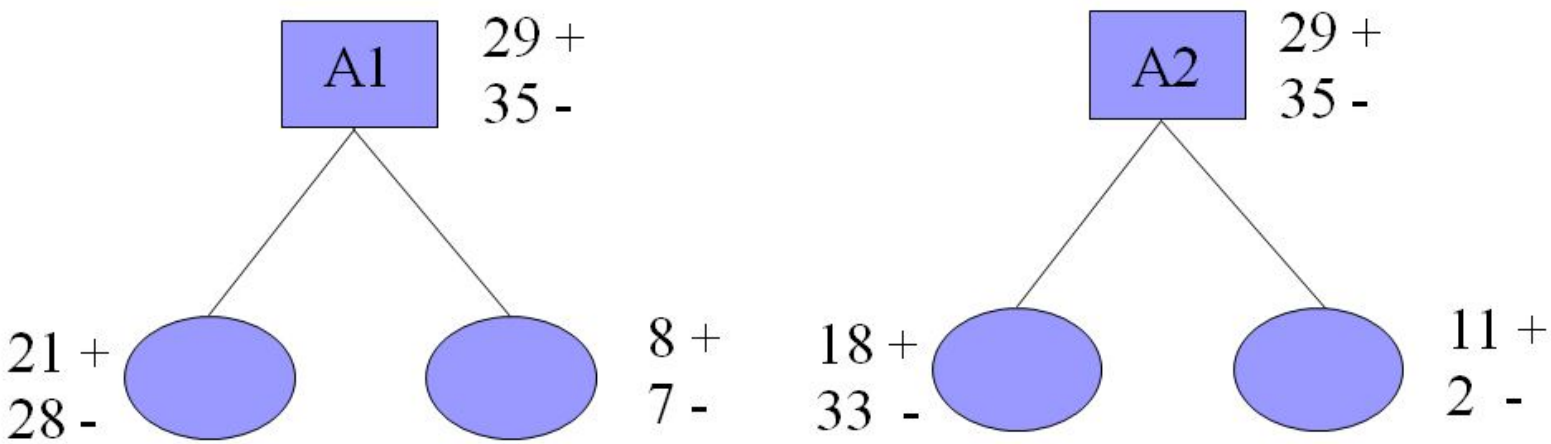
ARBRE DE DÉCISION

- Méthode de classification et de prédiction
- Les attributs apparaissant dans l'arbre sont les attributs pertinents pour le problème considéré
- Un arbre est équivalent à un ensemble de règles de décision

CONSTRUCTION D'ARBRES DE DÉCISION
// ALGORITHME TOP-DOWN

- construireNoeud(E)
  - Si tous les exemples de E sont dans la même classe C_i
    - affecter l'étiquette C_i au noeud courant
  - Sinon sélectionner un attribut A avec les valeurs $v_1 \ldots v_n$
    Partitionner E selon $v_1 \ldots v_n$ en $E_1, \ldots, E_n$
    - Pour chacune des branches n construireNoeud(E_j).

CHOIX DU MEILLEUR ATTRIBUT ?



ENTROPIE

- S est un ensemble d'exemples
- p^+ proportion d'ex. positifs dans S
- p^- proportion d'ex. négatifs dans S
- L'entropie mesure l'impureté de S
- $\text{Ent}(S) = -p^+ \log_2 p^+ - p^- \log_2 p^-$
- Ent(S) = Nombre de bits nécessaire pour coder la classe(+ ou -) d'un élément tiré au hasard

ARBRE DE DÉCISION
// GAIN

- $\text{Gain}(S, A) =$ réduction d'entropie due au test de l'attribut A

$$\text{Gain}(S, A) = \text{Ent}(s) - \sum_{v \in \text{Valeurs}(A)} \frac{|S_v|}{|S|} \text{Ent}(S_v)$$

- Si on travaille avec plus de 2 classes, la formule d'entropie peut être généralisée

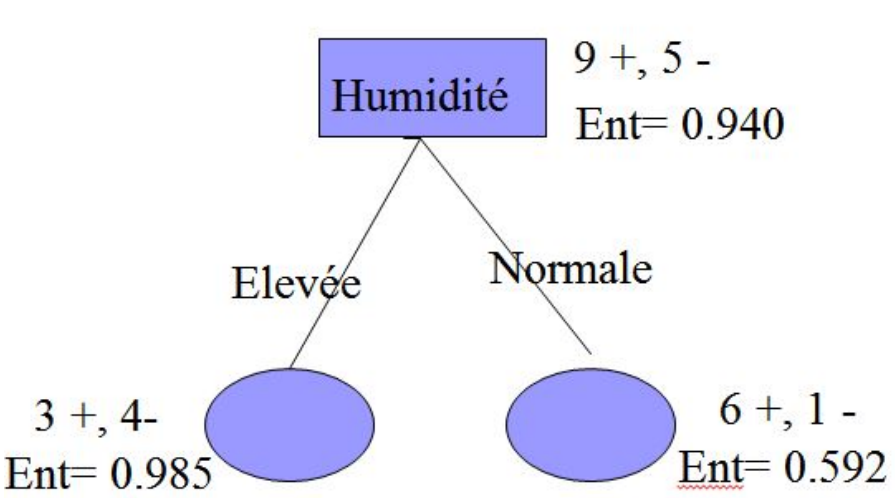
$$\text{Ent}(s) = -\sum_{s} f_s \log_{|s|} f_s$$

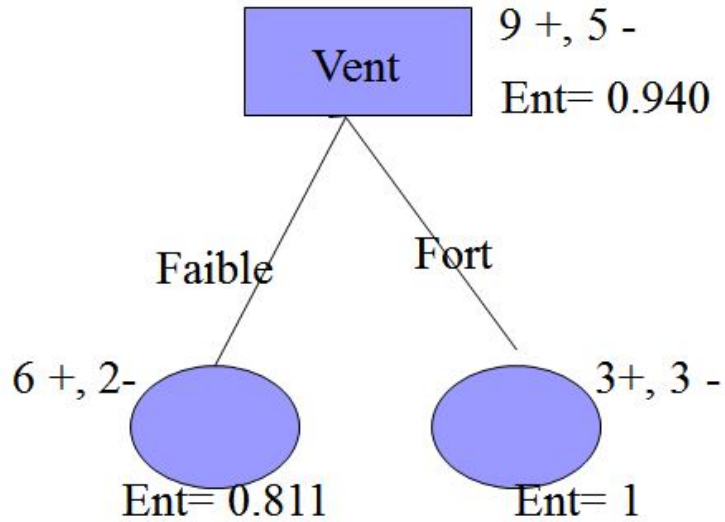
- où fs est la proportion de la classe s

Implémentations

- C4.5, ID3 (basé sur l'entropie)
- CART (basé sur le Gini index)

GAIN
// EXEMPLE





- $\text{Gain}(\text{S}, \text{Hum.}) = 0.940 - 7/14 * 0.985 - 7/14 * 0.592 = 0.151$
- $\text{Gain}(\text{S}, \text{Vent}) = 0.940 - 8/14 * 0.811 - 6/14 * 1. = 0.048$

ENTROPIE ET GAIN
// EXERCICE

Soit S $= \{8^+; 8^-\}$ un ensemble d'exemples Calculer :

- L'entropie de S
- Soit A l'attribut binaire qui sépare les en $A_1\{8^+, 0^-\}$ et $A_2\{0^+, 8^-\}$. Calculer l'entropie de A_1 et de A_2. Calculer le gain.
- Soit B l'attribut binaire qui sépare les en $B_1\{3^+, 3^-\}$ et $B_2\{5^+, 5^-\}$. Calculer l'entropie de B_1 et de B_2 Calculer le gain.

ARBRE DE DÉCISION
// SPAM

blacklist	connu	type	fautes	spam
F	T	texte	F	-
T	F	image	F	+
T	T	texte	F	+
F	T	imettexte	T	-
T	T	imettexte	T	+
F	F	texte	T	+
F	F	texte	F	-

Exercice

Sachant que le premier attribut sélectionné est type, continuer la construction de l'arbre de décision correspondant à cette base en utilisant l'algorithme vu en cours et le critère d'entropie.

ARBRE DE DÉCISION
// COUPS DE SOLEIL

cheveux	taille	poids	crème solaire	coup de soleil
blond	moyenne	léger	non	+
blond	grande	moyen	oui	-
brun	petite	moyen	oui	-
blond	petite	moyen	non	+
roux	moyenne	lourd	non	+
brun	grande	lourd	non	-
brun	moyenne	lourd	non	-
blond	petite	léger	oui	-

Exercice

Construire l'arbre de décision correspondant à cette base

ARBRE DE DÉCISION
// LIMITES

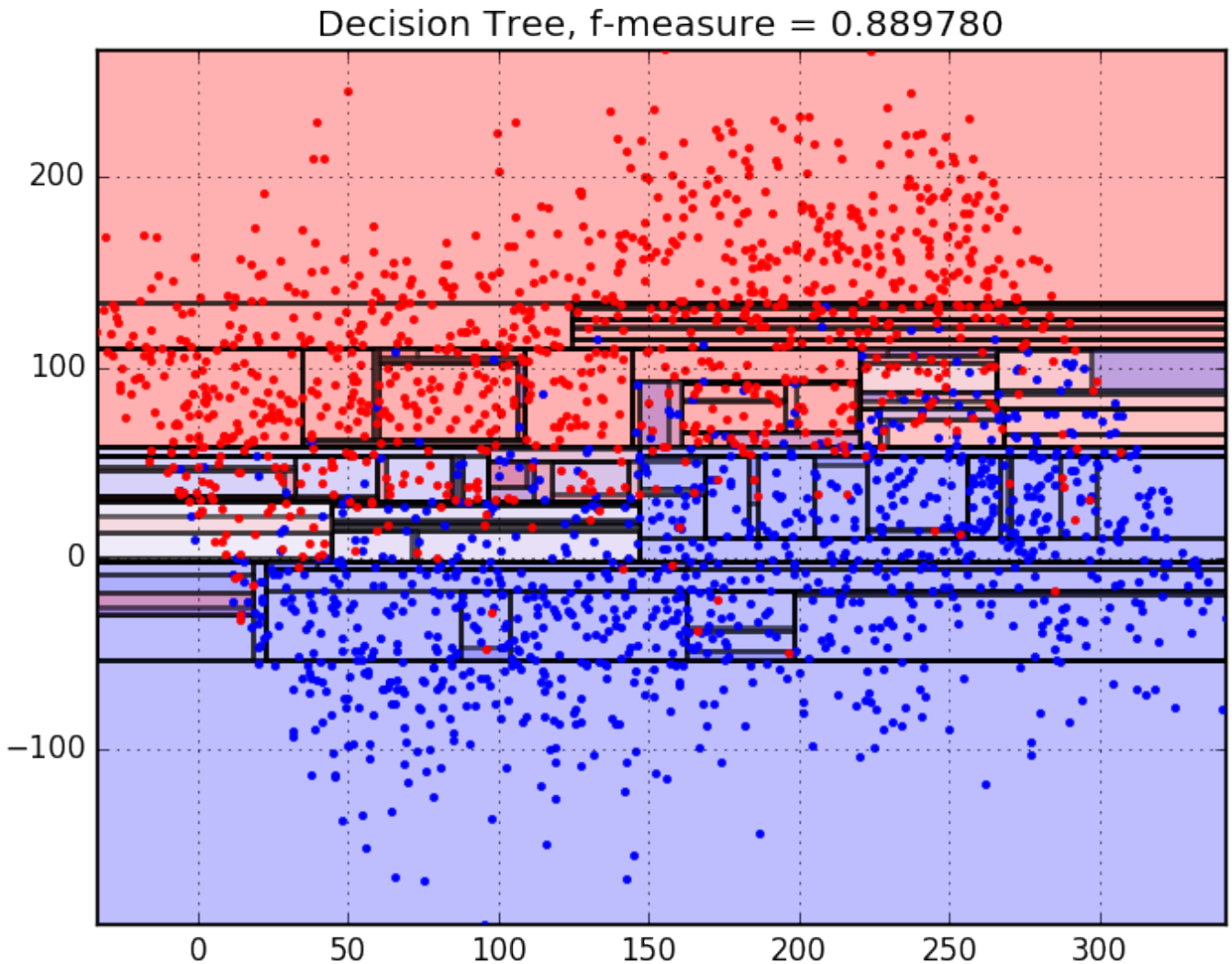
- Attributs à valeur continue
- Attributs à facteurs de branchement différents
- Valeurs manquantes
- Sur-apprentissage
- Recherche gloutonne
- Variance des résultats :
  - arbres différents à partir de données peu différentes

ATTRIBUTS À VALEURS CONTINUES

- Créer dynamiquement un attribut binaire à partir des valeurs observées, dans le nœud considéré, pour un attribut A continu
- On trie les valeurs observées pour A
- On cherche une seuil
  - s(A $<$ s ou A $>$ s) ?
  - qui donne le meilleur gain d'information

ATTRIBUTS À VALEURS CONTINUES

// ILLUSTATION



SURAJUSTEMENT DE L'ARBRE DE DÉCISION

- Surajustement= arbre trop "proche" des données
- Arbre peu performant pour la prédiction Le bruit dans les exemples peut conduire à un surajustement de l 'arbre
  - E15= (Soleil,Chaude,Normale, Fort, NON)
  - Conséquences pour l'arbre construit ?
- Elagage de l'arbre

SURAJUSTEMENT DE L'ARBRE DE DÉCISION

// SOLUTION

- Comment éviter le surajustement ?
  - Arrêter de construire l 'arbre quand la séparation des exemples n'est plus significative
  - Construire l 'arbre complet puis l'élaguer
  - Transformer l 'arbre en ensemble de règles et simplifier les règles