Fondements de l'Intelligence Artificielle

Cours 5 – mercredi 9 novembre 2022

Adrien Revault d'Allonnes

ara@up8.edu

Université Paris 8 – Vincennes à Saint-Denis

FIA – sept. à déc., 2022

Contexte : classification

- Modèles du jour : les arbres de décision
- Notations
  - N exemples, notés x_i, décrits par P attributs, qualitatifs ou quantitatifs
  - C catégories possibles
- Différences
  - classification multiclasse (monolabel)
  - données peuvent être qualitatives (pas que vectorielles)

Exemple : données

	Toux	Fièvre	Poids	Douleur
André	non	oui	normal	gorge
Béatrice	non	oui	normal	abdomen
Charles	oui	oui	maigre	aucune
David	oui	non	obèse	poitrine

Exemple : ensemble d'apprentissage

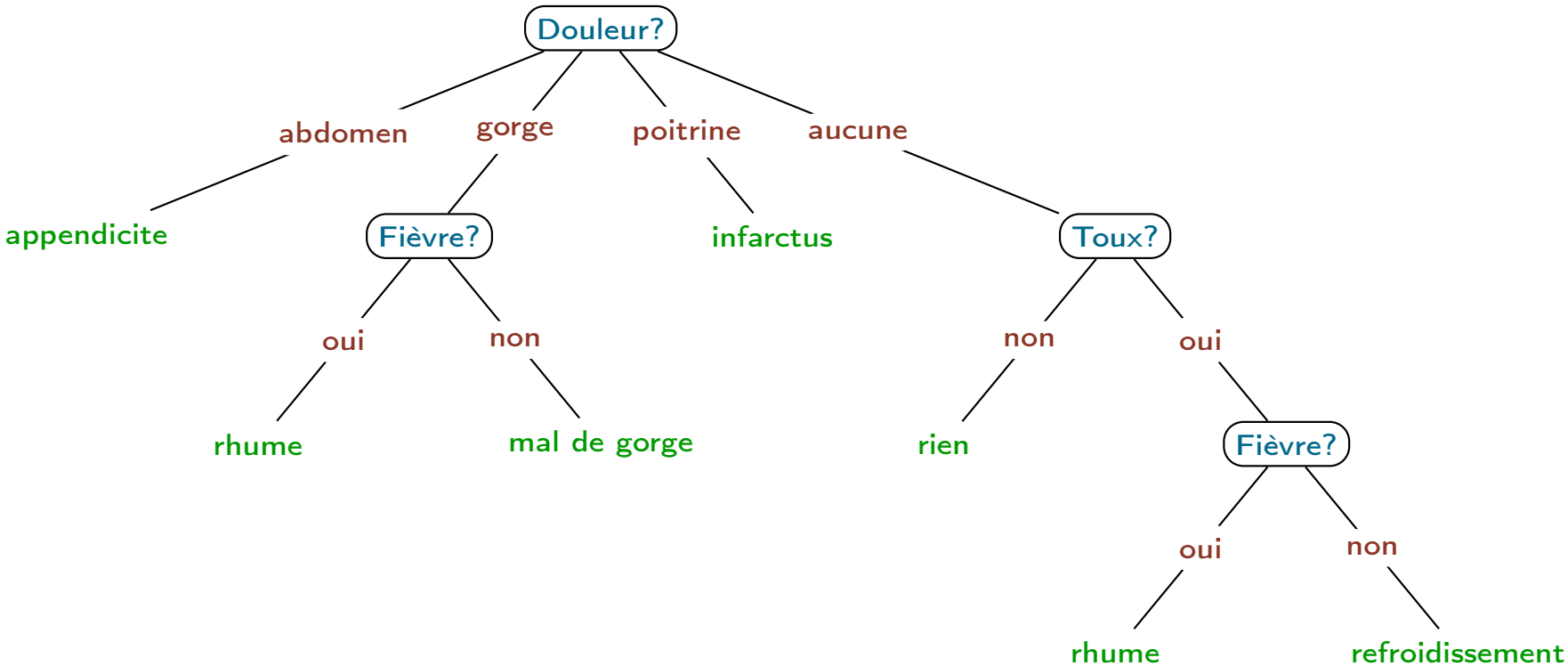
	Toux	Fièvre	Poids	Douleur	**Diagnostic**
André	non	oui	normal	gorge	rhume
Béatrice	non	oui	normal	abdomen	appendicite

...

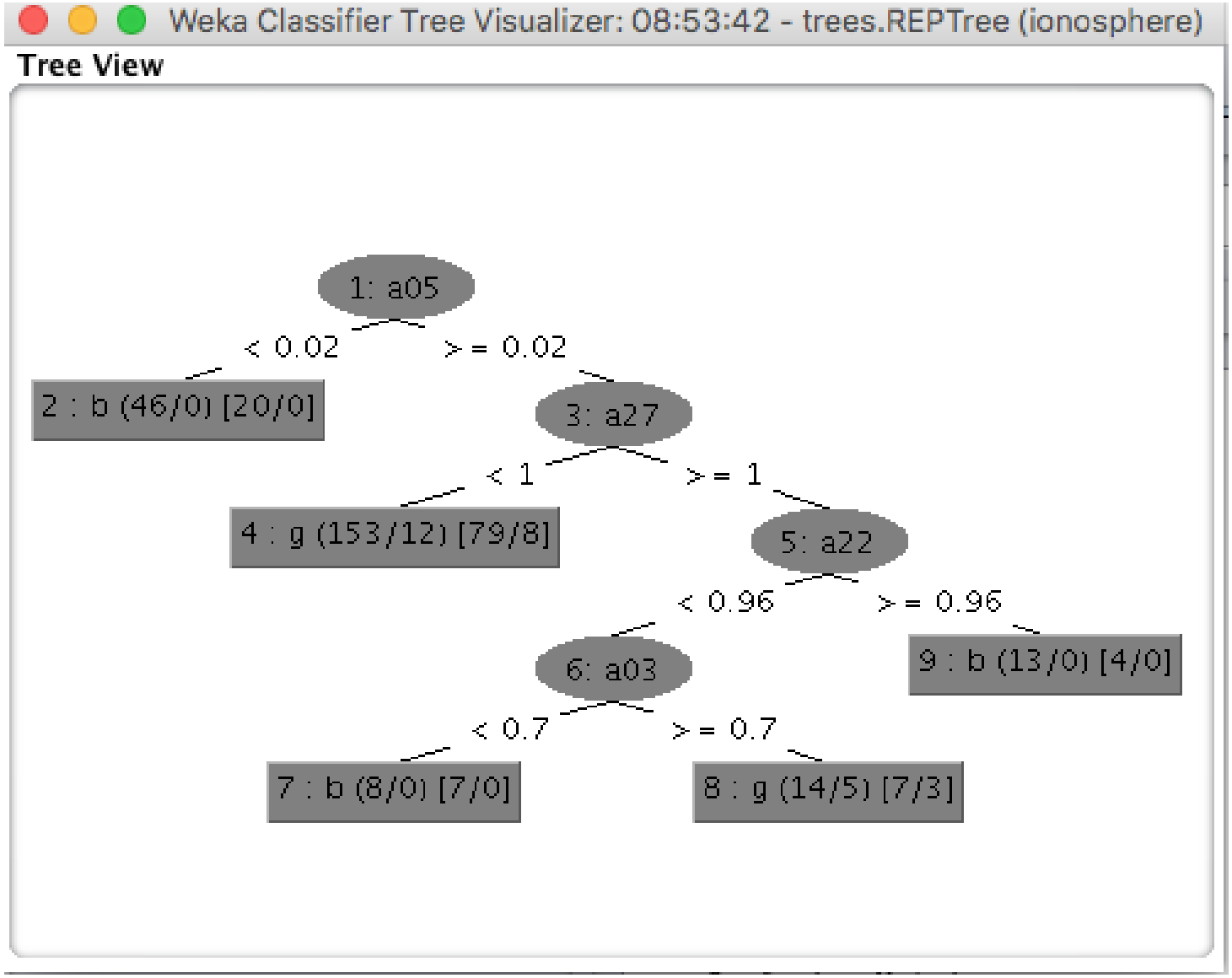
Arbres de décision : principes

- Un arbre de décision est un classifieur, représenté sous forme d'arbre, tel que :
  - les nœuds de l'arbre testent les attributs
  - une branche par modalité de l'attribut testé
  - les feuilles spécifient la catégorie

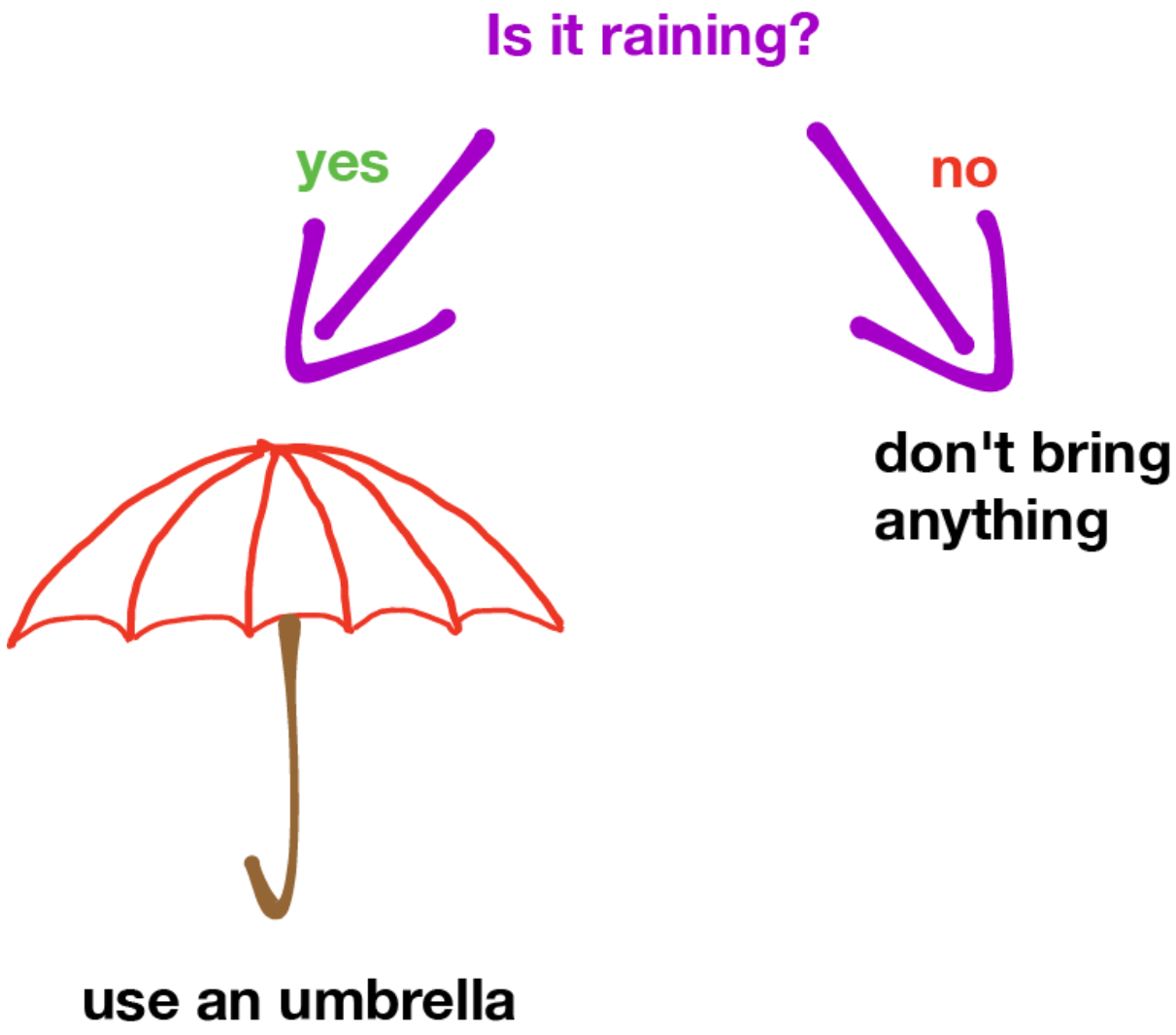
Exemple


Douleur?
abdomen
gorge
poitrine
aucune
appendicite
Fièvre?
infarctus
Toux?
oui
non
non
oui
rhume
mal de gorge
rien
Fièvre?
oui
non
rhume
refroidissement

Exemple

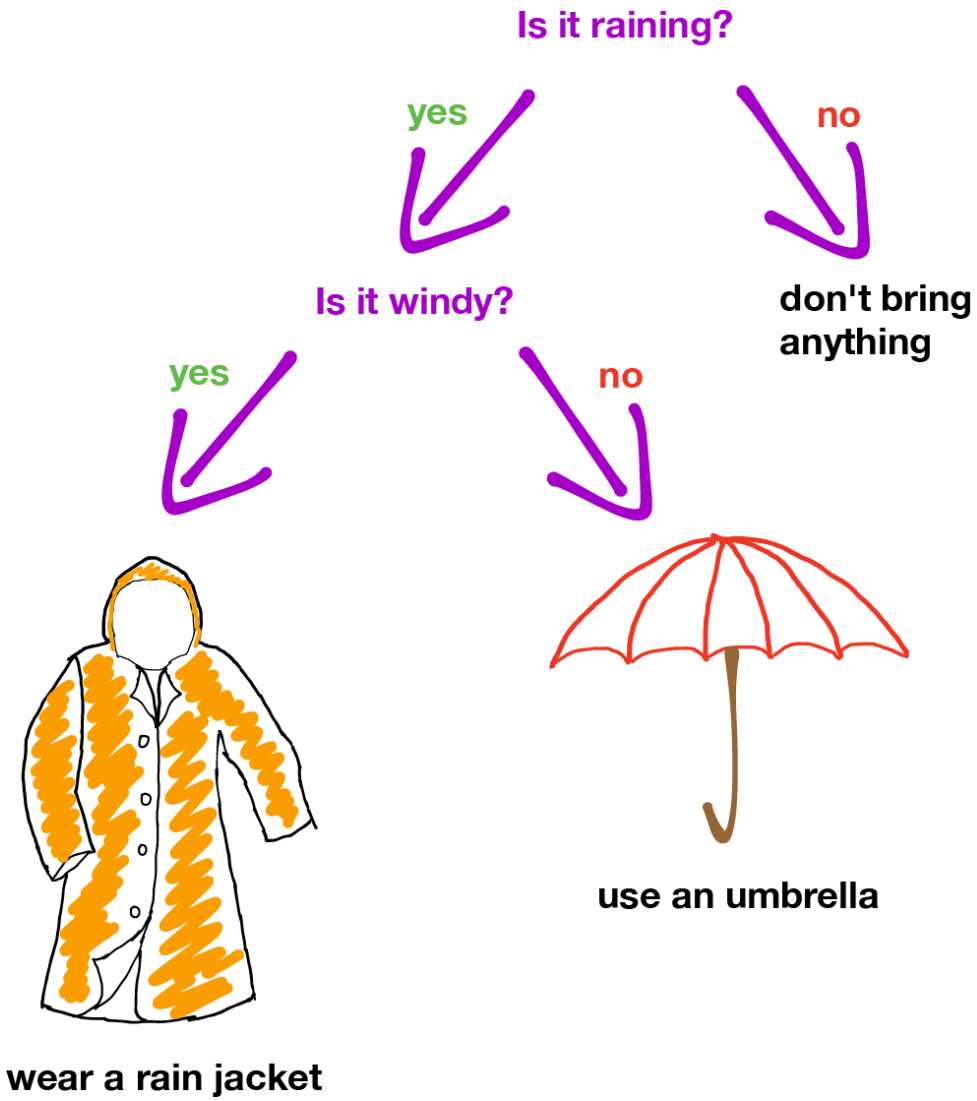

Weka Classifier Tree Visualizer: 08:53:42 - trees.REPTree (ionosphere)
Tree View
1: a05
< 0.02
>= 0.02
2 : b (46/0) [20/0]
3: a27
< 1
>= 1
4 : g (153/12) [79/8]
5: a22
< 0.96
>= 0.96
9 : b (13/0) [4/0]
6: a03
< 0.7
>= 0.7
7 : b (8/0) [7/0]
8 : g (14/5) [7/3]

Exemple

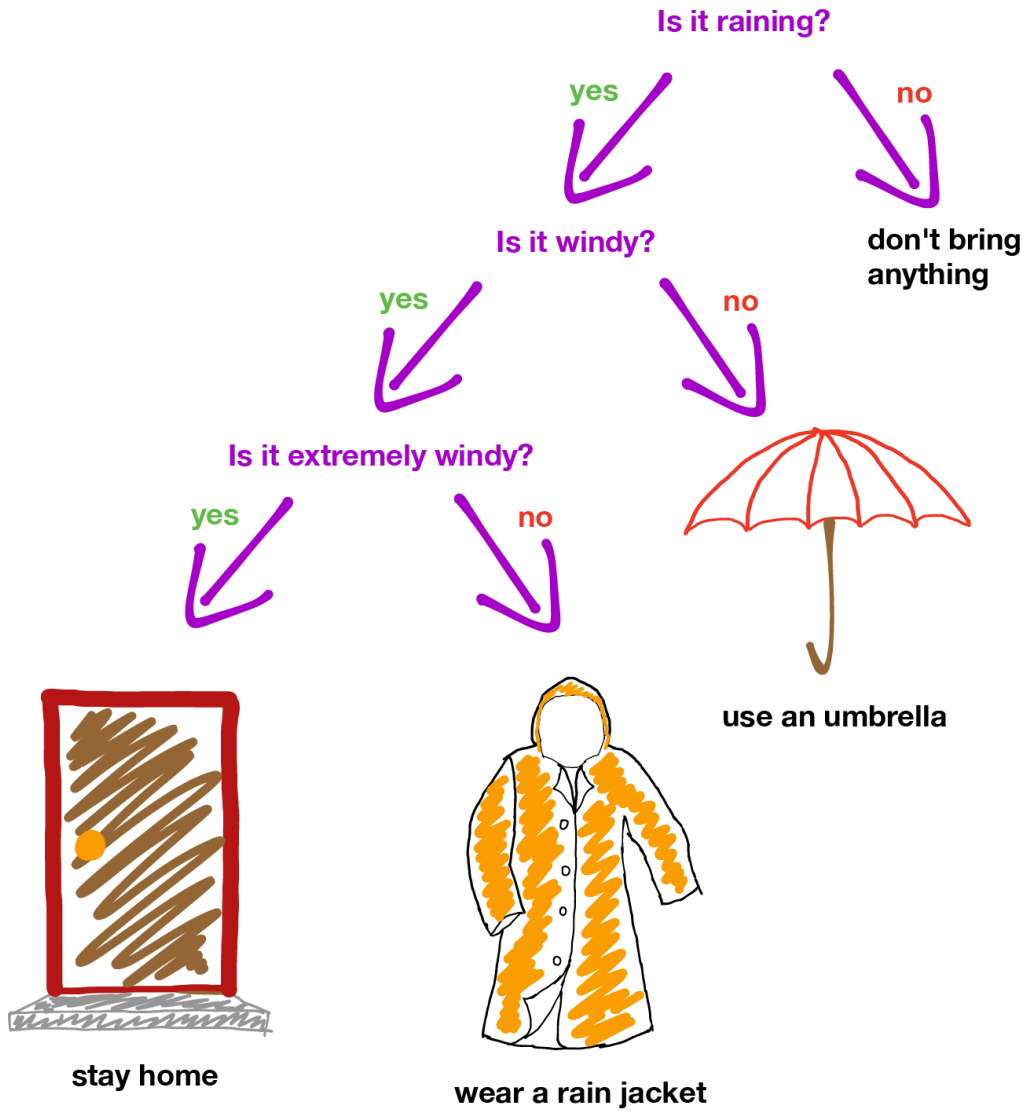


© Machine Learning @ Berkeley

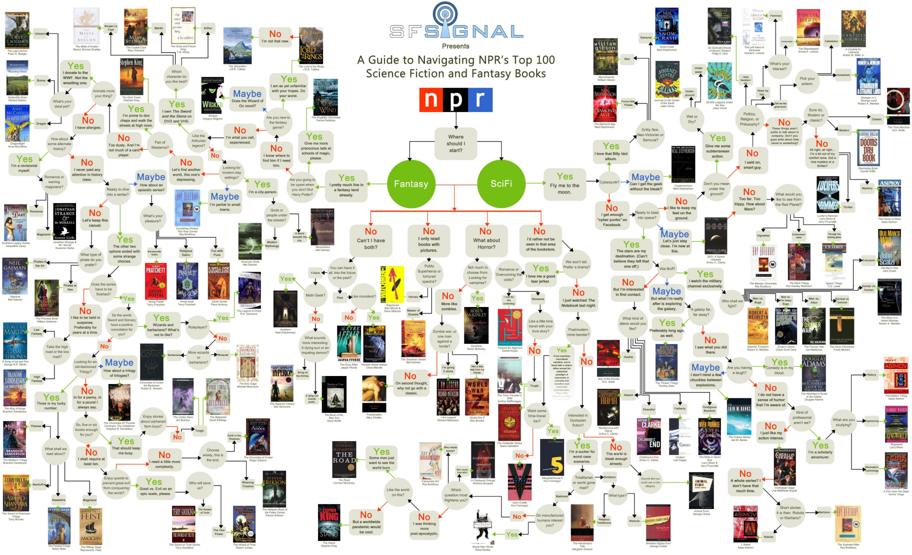
Exemple



Exemple


Is it raining?
yes
no
Is it windy?
don't bring anything
yes
no
Is it extremely windy?
yes
no
use an umbrella
stay home
wear a rain jacket
© Machine Learning @ Berkeley

Exemple


SF SIGNAL
Presents
A Guide to Navigating NPR's Top 100
Science Fiction and Fantasy Books
npr
Where should I start?
Fantasy
SciFi

Arbres de décision : intérêts

- Avantages
  - classifieurs **interprétables**
    - $\neq$ perceptron
    - $\neq$ kppv
  - traitent données qualitatives
  - fonctionnent bien

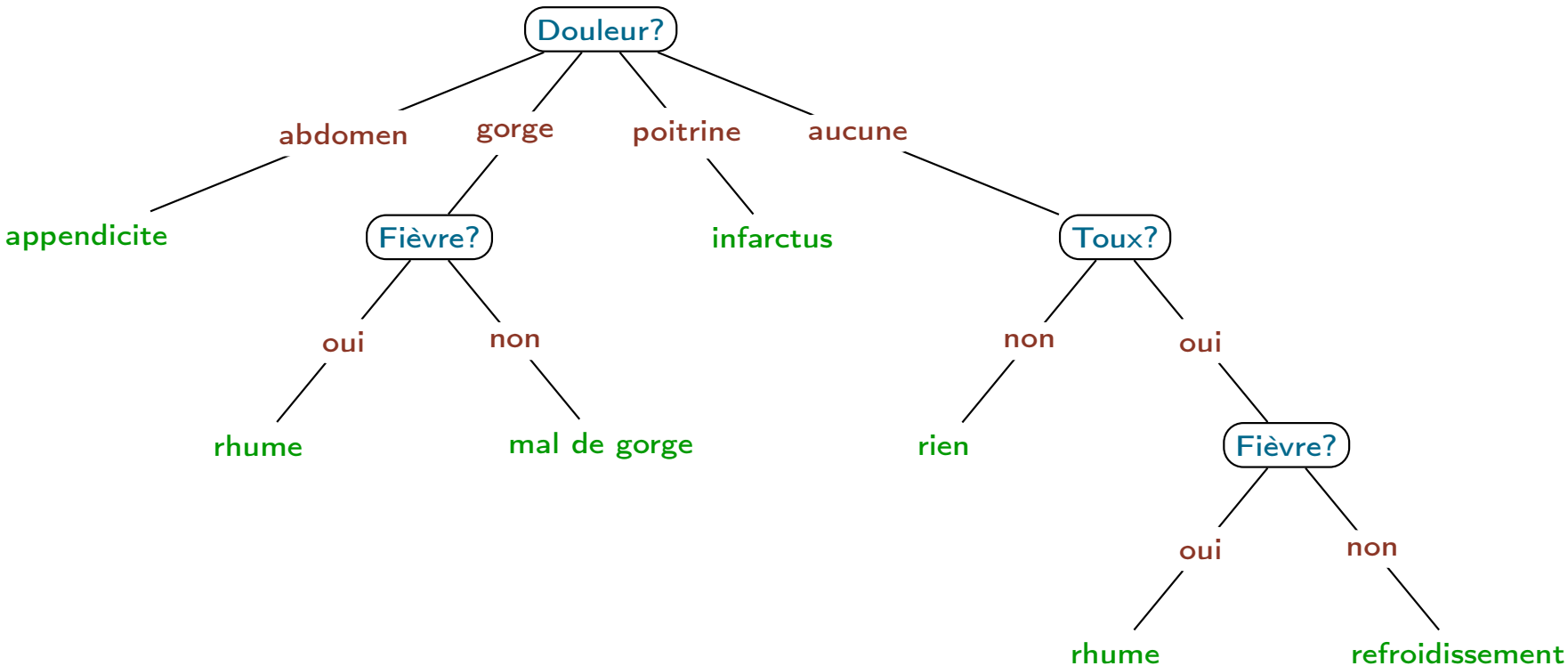
    NB si nombre attributs pas trop grand

- Inconvénients
  - pas toujours si interprétables que ça
  - lents et instables pendant l'apprentissage

Arbres de décision : interprétabilité

- Avantage
  - fonctions de décision lisibles « par un humain »

$\Rightarrow$ d'où l'utilisation de ces arbres pour la découverte de propriétés



Arbres de décision : interprétabilité

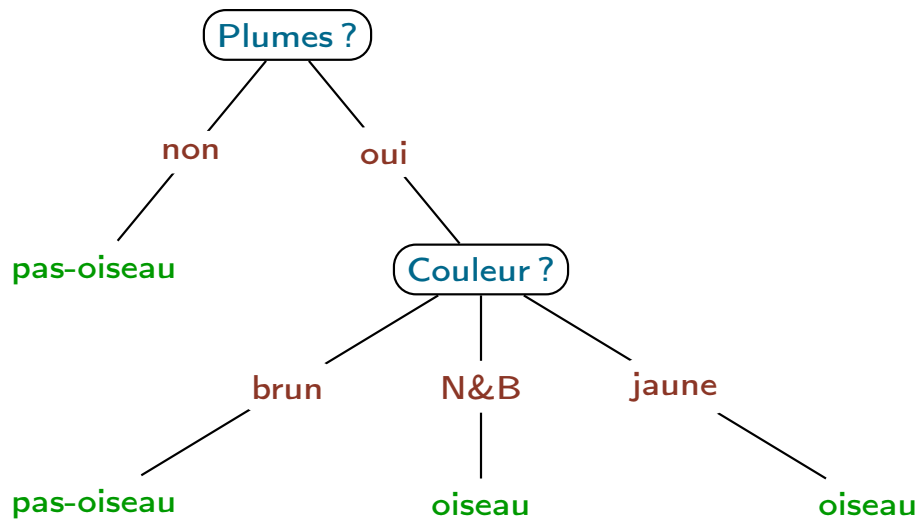
- Toute fonction booléenne peut s'écrire comme un arbre de décision
- Selon les fonctions, les arbres sont plus ou moins grands
- Un arbre peut se représenter comme une disjonction de règles

```
   (Si Plumes=non                          Alors Classe=pas-oiseau)
ou (Si Plumes=oui ET couleur=brun          Alors Classe=pas-oiseau)
ou (Si Plumes=oui ET couleur=N&B           Alors Classe=oiseau)
ou (Si Plumes=oui ET couleur=jaune         Alors Classe=oiseau)
```

Arbres de décision : interprétabilité

- Toute fonction booléenne peut s'écrire comme un arbre de décision
- Selon les fonctions, les arbres sont plus ou moins grands
- Un arbre peut se représenter comme une disjonction de règles

```
   (Si Plumes=non                          Alors Classe=pas-oiseau)
ou (Si Plumes=oui ET couleur=brun          Alors Classe=pas-oiseau)
ou (Si Plumes=oui ET couleur=N&B           Alors Classe=oiseau)
ou (Si Plumes=oui ET couleur=jaune         Alors Classe=oiseau)
```



Arbres de décision : interprétabilité...

- Toute fonction booléenne peut s'écrire comme un arbre de décision
  - Rappel :
    - avec 6 attributs booléens
    - on peut définir environ 2 milliards de fonctions booléennes...

- Selon les fonctions, les arbres sont plus ou moins grands
  - la taille de l'arbre peut grandir exponentiellement !

- Un arbre peut se représenter comme une disjonction de règles
  - limité à la logique des propositions (pas de relations)

Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage,
  comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables

Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage,
  comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage,
  comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

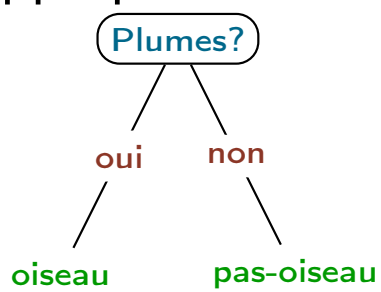
- Quel arbre est le plus approprié ?

Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage,
  comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

- Quel arbre est le plus approprié ?

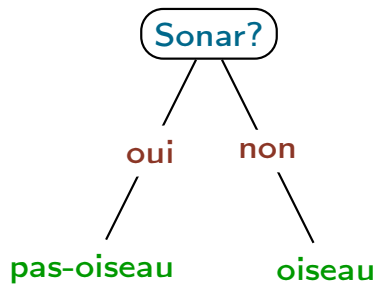


Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage, comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

- Quel arbre est le plus approprié ?

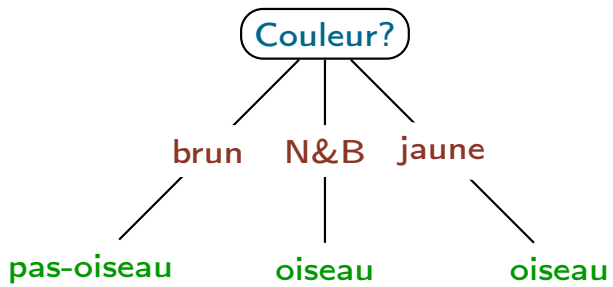


Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage, comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

- Quel arbre est le plus approprié ?

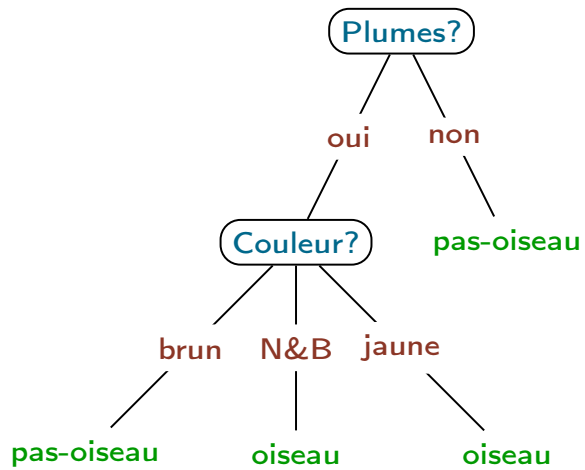


Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage, comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

- Quel arbre est le plus approprié ?

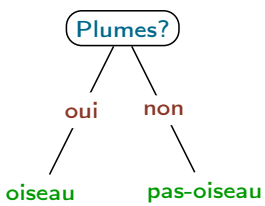


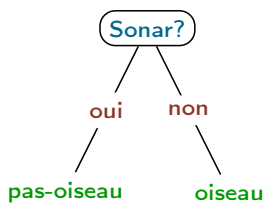
Arbres de décision : apprentissage

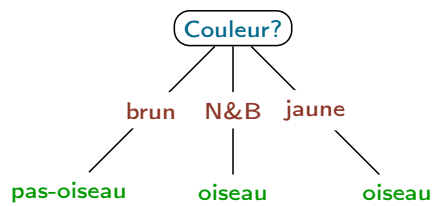
- À partir d'un ensemble d'apprentissage, comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables
- Soit la base de données suivante :

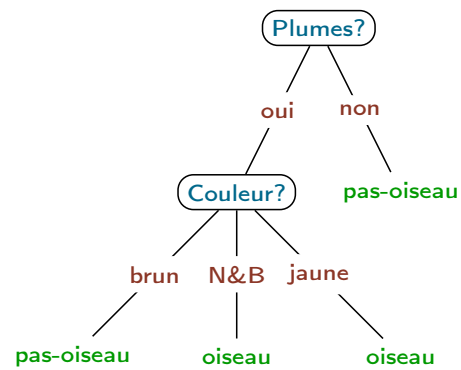
	Couleur	Ailes	Plumes	Sonar	**Concept**
Faucon	jaune	oui	oui	non	oiseau
Pigeon	N&B	oui	oui	non	oiseau
Chauve-souris	brun	oui	non	oui	pas-oiseau

- Quel arbre est le plus approprié ?









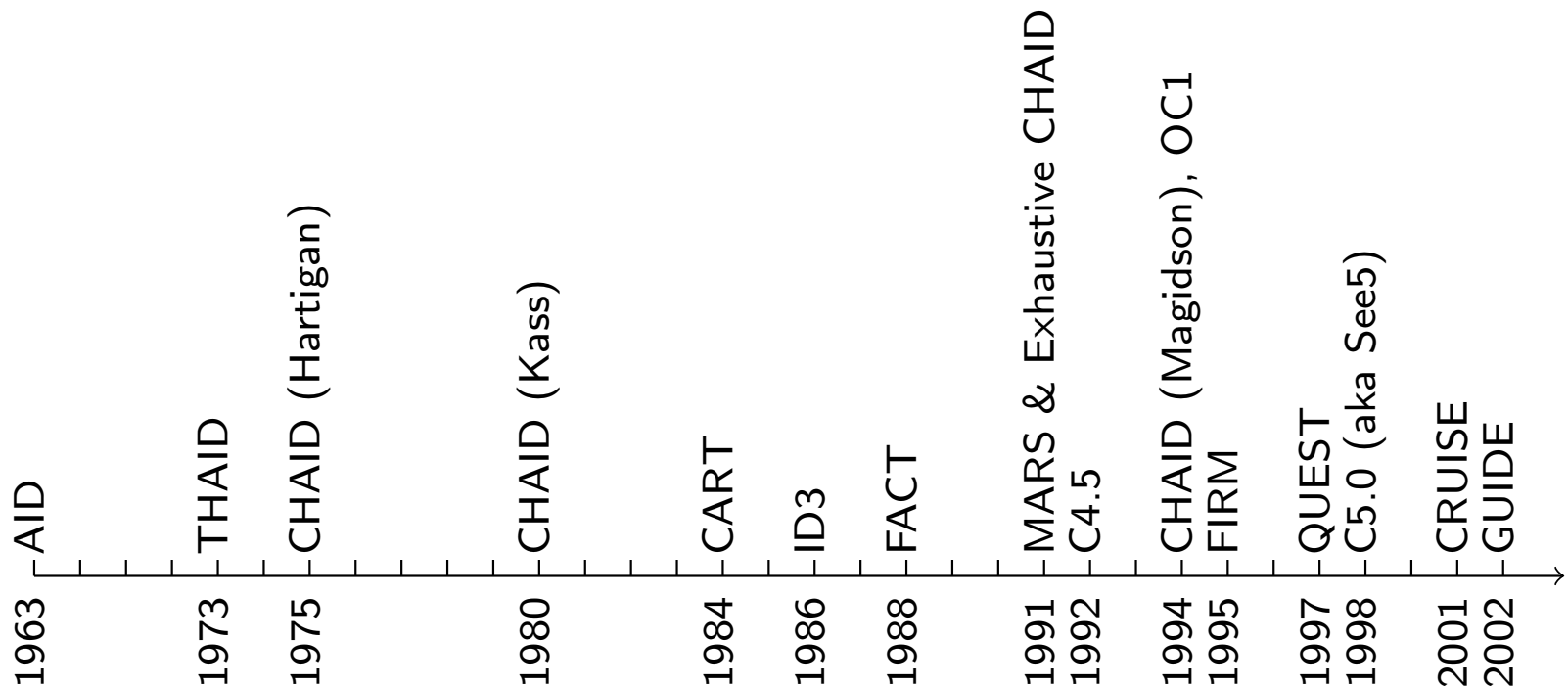
Arbres de décision : apprentissage

- À partir d'un ensemble d'apprentissage,
  comment construire un arbre de décision efficace ?
  - le plus souvent, plusieurs arbres possibles **et corrects**
  - énumération exhaustive impossible (NP-complet)
    - 4 attributs et 3 modalités = 55 296 arbres réalisables

- Solution :
  - construction itérative de l'arbre

  $\rightarrow$ méthode gloutonne

Apprentissage d'arbres de décision : les algorithmes



Histoire des algorithmes d'arbres de décision

Principes

- Construction de la racine vers les feuilles
  - ajout de l'attribut le **plus informatif** comme nœud de l'arbre
- Utilisation d'une **mesure de dicrimination**
  - mesures classiques : entropie de Shannon, index de Gini, ...
- Critères d'une bonne mesure de discrimination :
  - obtenir des nœuds cohérents
  - minimiser la taille de l'arbre
  - obtenir de bons résultats en classification

Principes

- Construction de la racine vers les feuilles
  - ajout de l'attribut le **plus informatif** comme nœud de l'arbre
- Utilisation d'une **mesure de dicrimination**
  - mesures classiques : entropie de Shannon, index de Gini, ...
- Critères d'une bonne mesure de discrimination :
  - obtenir des nœuds cohérents
  - minimiser la taille de l'arbre
  - obtenir de bons résultats en classification
- Rôle de la mesure de discrimination :
  - mesure de la « prédictabilité » de c_k, de C, d'une valeur v_i de A
  - mesure du **pouvoir de discrimination** de l'attribut A sur la classe C

Mesures de discrimination classiques

- Utilisation de l'entropie de Shannon :

$$H_S(C|A) = -\sum_i P(v_i) \sum_k P(c_k|v_i) \log(P(c_k|v_i))$$

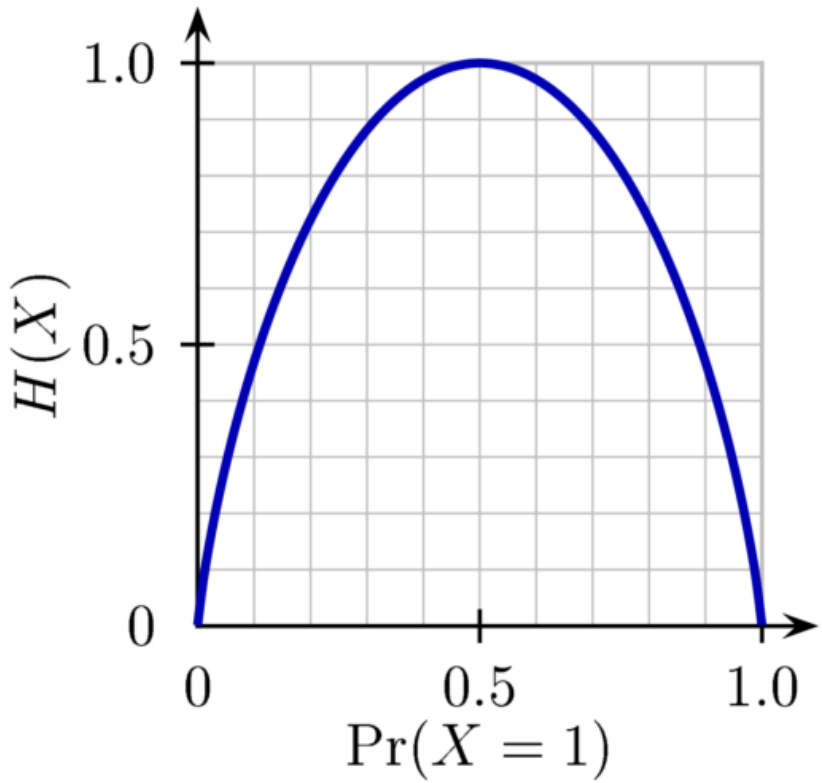
- Mesure issue de la théorie de l'information
  - initiée par C. E. Shannon en 1948
- Mesure un **taux de désordre**

  $\Rightarrow$ à minimiser

Mesures de discrimination classiques : 2 classes

- Utilisation de l'entropie de Shannon : 2 classes

$$H_S(C|A) = -P_+ \log(P_+) - P_- \log(P_-)$$



Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - à noter. . .ça se simplifie pas mal

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - à noter. . .ça se simplifie pas mal

$$I_G(p) \quad = \quad \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - à noter. . .ça se simplifie pas mal

$$I_G(p) \quad = \quad \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k \quad = \quad \sum_{i=1}^{n} p_i \times (1 - p_i)$$

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - à noter. . .ça se simplifie pas mal

$$\begin{aligned} I_G(p) &= \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k = \sum_{i=1}^{n} p_i \times (1 - p_i) \\ &= \sum_{i=1}^{n} p_i - p_i^2 \end{aligned}$$

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - à noter. . .ça se simplifie pas mal

$$\begin{aligned} I_G(p) &= \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k &= \sum_{i=1}^{n} p_i \times (1 - p_i) \\ &= \sum_{i=1}^{n} p_i - p_i^2 &= \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i^2 \end{aligned}$$

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - à noter. . .ça se simplifie pas mal

$$\begin{aligned} I_G(p) &= \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k = \sum_{i=1}^{n} p_i \times (1 - p_i) \\ &= \sum_{i=1}^{n} p_i - p_i^2 = \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i^2 \\ &= 1 - \sum_{i=1}^{n} p_i^2 \end{aligned}$$

Mesures de discrimination classiques

- Utilisation de l'indice de Gini :

$$I_G(p) = \sum_{i=1}^{n} p_i \sum_{k \neq I} p_k$$

- Mesure adaptée du coefficient de Gini, mesure économique
- Mesure « l'**impureté** » d'un ensemble

  $\Rightarrow$ à minimiser
  - le critère est donc

$$H_G(C|A) = \sum_{a_i \in A} P(a_i) \times (1 - \sum_{c_k \in C} P(c_k|a_i)^2)$$

Construction de l'arbre : cas général

- Algorithme d'apprentissage
  1. Pour chaque attribut A_j, calculer $H(C|A_j)$
  2. Choisir l'attribut A_j qui minimise $H(C|A_j)$
     - ajouter un nœud à l'arbre de décision sur l'attribut A_j
  3. Partitionner la base d'apprentissage, selon les modalités d'A_j

Construction de l'arbre : cas général

- Algorithme d'apprentissage
  1. Pour chaque attribut A_j, calculer $H(C|A_j)$
  2. Choisir l'attribut A_j qui minimise $H(C|A_j)$
     - ajouter un nœud à l'arbre de décision sur l'attribut A_j
  3. Partitionner la base d'apprentissage, selon les modalités d'A_j
- Exemple

Exemple

	Devoirs finis	Parents de bonne humeur	Beau temps	Goûter pris	**Jouer dehors**
1	Vrai	Faux	Vrai	Faux	OUI
2	Faux	Vrai	Faux	Vrai	OUI
3	Vrai	Vrai	Vrai	Faux	OUI
4	Vrai	Faux	Vrai	Vrai	OUI
5	Faux	Vrai	Vrai	Vrai	NON
6	Faux	Vrai	Faux	Faux	NON
7	Vrai	Faux	Faux	Vrai	NON
8	Vrai	Vrai	Faux	Faux	NON

Exemple

	DF	BH	BT	GP	JD
1	V	F	V	F	OUI
2	F	V	F	V	OUI
3	V	V	V	F	OUI
4	V	F	V	V	OUI
5	F	V	V	V	NON
6	F	V	F	F	NON
7	V	F	F	V	NON
8	V	V	F	F	NON

Exemple

	DF	BH	BT	GP	JD
1	V	F	V	F	OUI
2	F	V	F	V	OUI
3	V	V	V	F	OUI
4	V	F	V	V	OUI
5	F	V	V	V	NON
6	F	V	F	F	NON
7	V	F	F	V	NON
8	V	V	F	F	NON

- Calcul de $H(C|DF)$, $H(C|BH)$, $H(C|BT)$ et $H(C|GP)$

Exemple

	DF	BH	BT	GP	JD
1	V	F	V	F	OUI
2	F	V	F	V	OUI
3	V	V	V	F	OUI
4	V	F	V	V	OUI
5	F	V	V	V	NON
6	F	V	F	F	NON
7	V	F	F	V	NON
8	V	V	F	F	NON

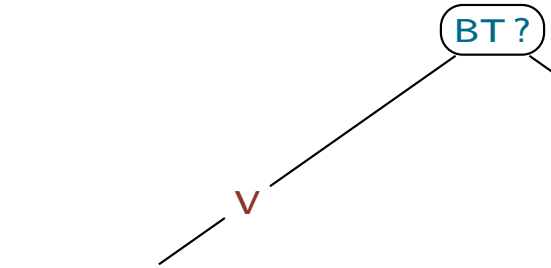
- Calcul de $H(C|DF)$
  - $H(C|DF) = \frac{5}{8}J(DF = V) + \frac{3}{8}J(DF = F)$
  - $J(DF = V) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5})$
  - $J(DF = F) = -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})$

Exemple

	DF	BH	BT	GP	JD
1	V	F	V	F	OUI
2	F	V	F	V	OUI
3	V	V	V	F	OUI
4	V	F	V	V	OUI
5	F	V	V	V	NON
6	F	V	F	F	NON
7	V	F	F	V	NON
8	V	V	F	F	NON

- Calcul de $H(C|DF)$, $H(C|BH), H(C|BT)$ et $H(C|GP)$
  - $H(C|DF) = \frac{5}{8}J(DF = V) + \frac{3}{8}J(DF = F)$
  - $J(DF = V) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2(\frac{2}{5})$
  - $J(DF = F) = -\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})$
- $H(C|DF) = 0.95, H(C|BH) = 0.95, H(C|BT) = 0.8$ et $H(C|GP) = 1$

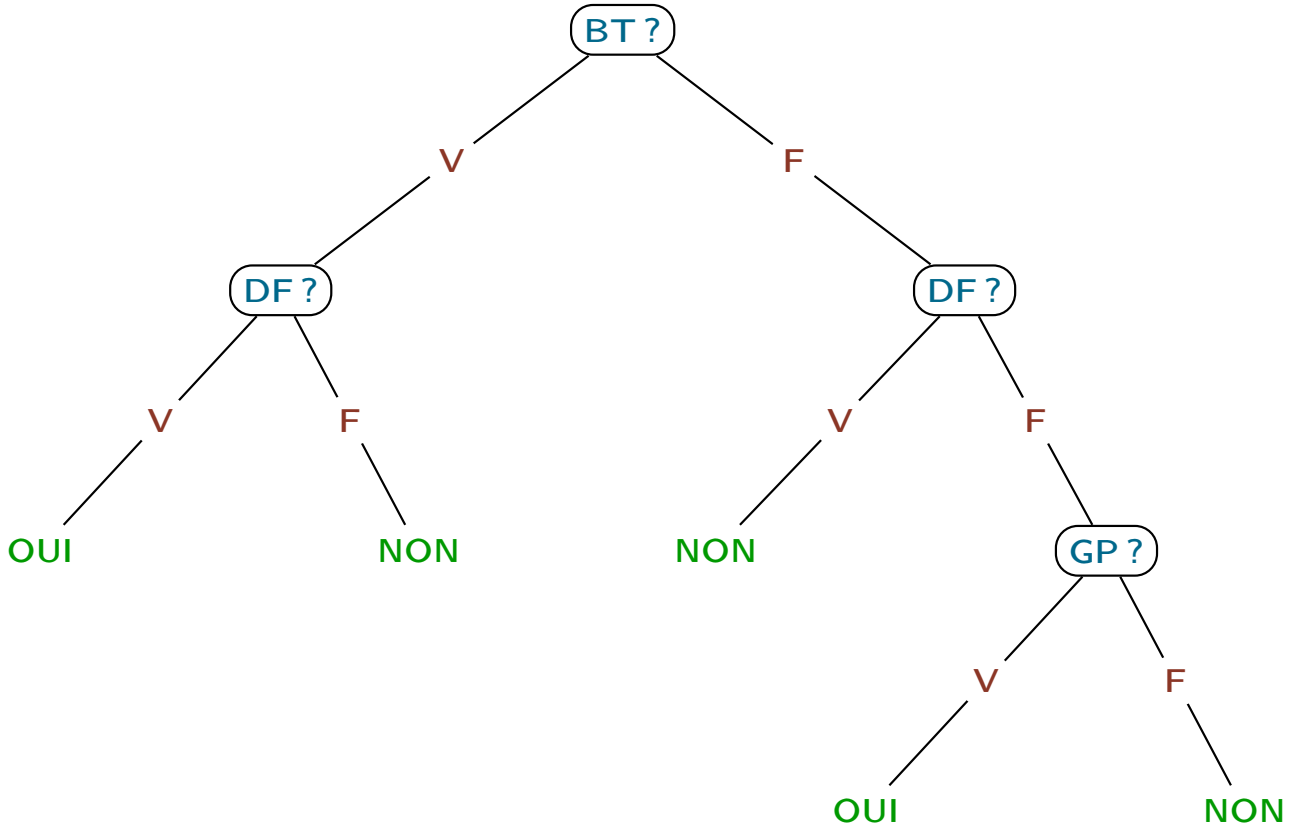
Exemple



	DF	BH	BT	GP	JD
1	V	F	V	F	OUI
3	V	V	V	F	OUI
4	V	F	V	V	OUI
5	F	V	V	V	NON

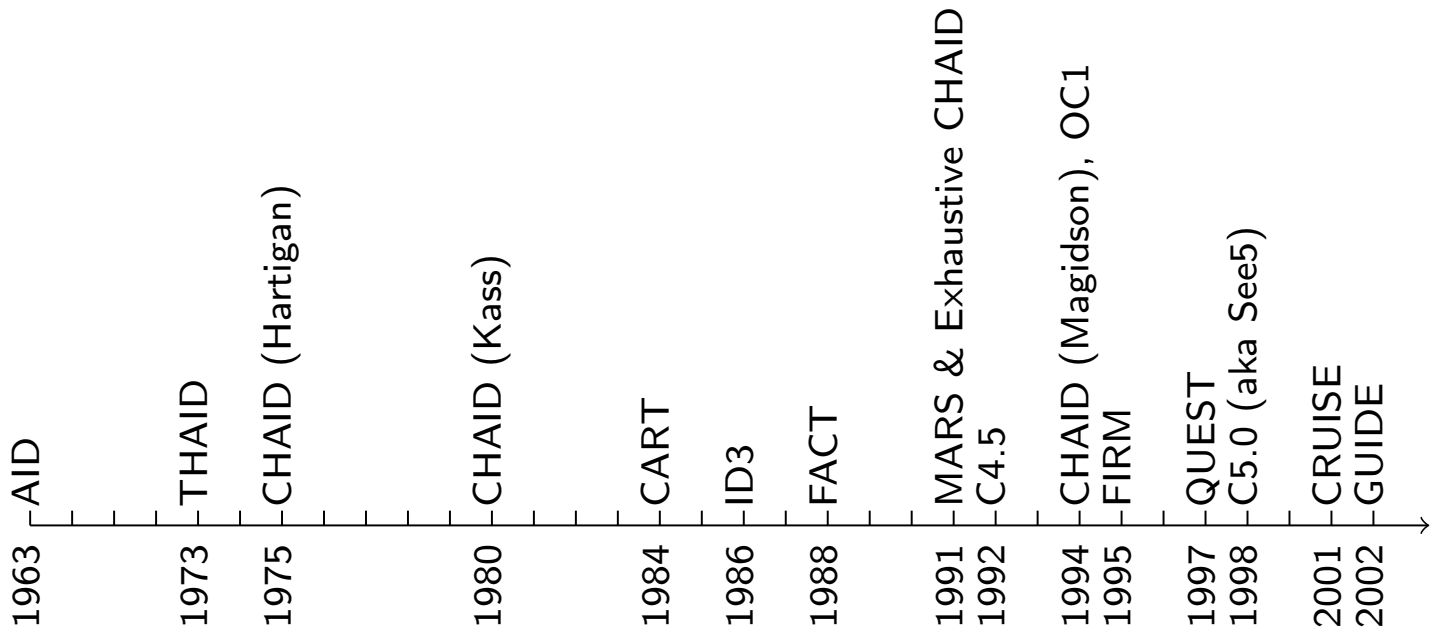
	DF	BH	BT	GP	JD
2	F	V	F	V	OUI
6	F	V	F	F	NON
7	V	F	F	V	NON
8	V	V	F	F	NON

Exemple



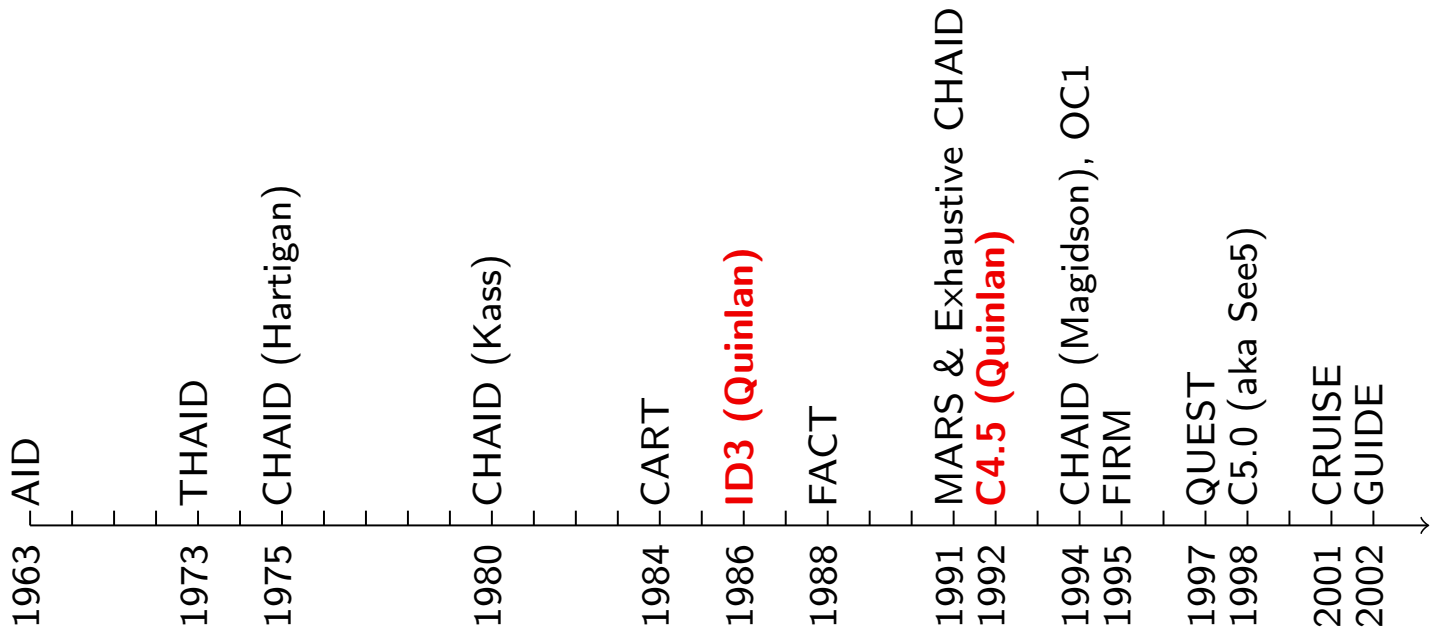
Données continues

- Problématique
  - que faire si on a des attributs continus ?



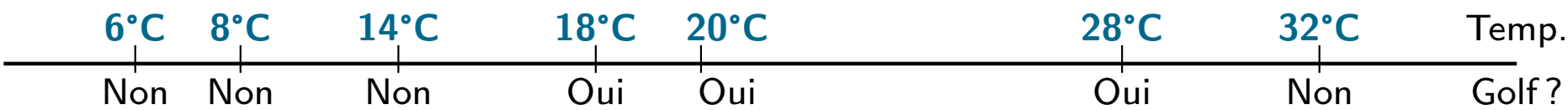
Données continues

- Problématique
  - que faire si on a des attributs continus ?



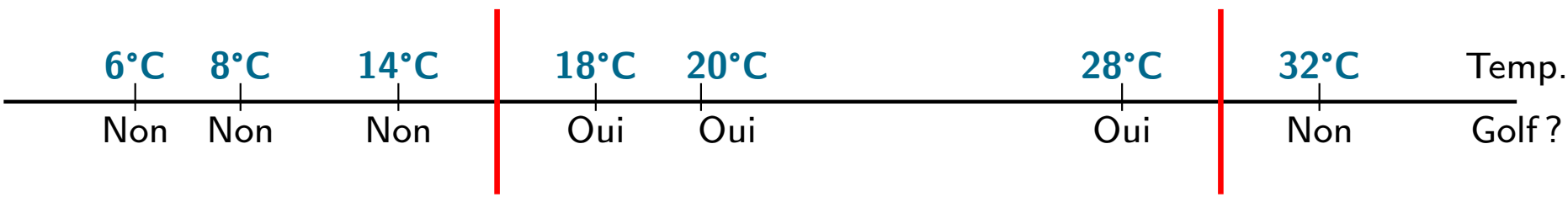
Discrétisation

- Problématique
  - que faire si on a des attributs continus ?
- Solution
  - on discrétise : transformation d'une variable continue en une (ou plusieurs) variable discrète

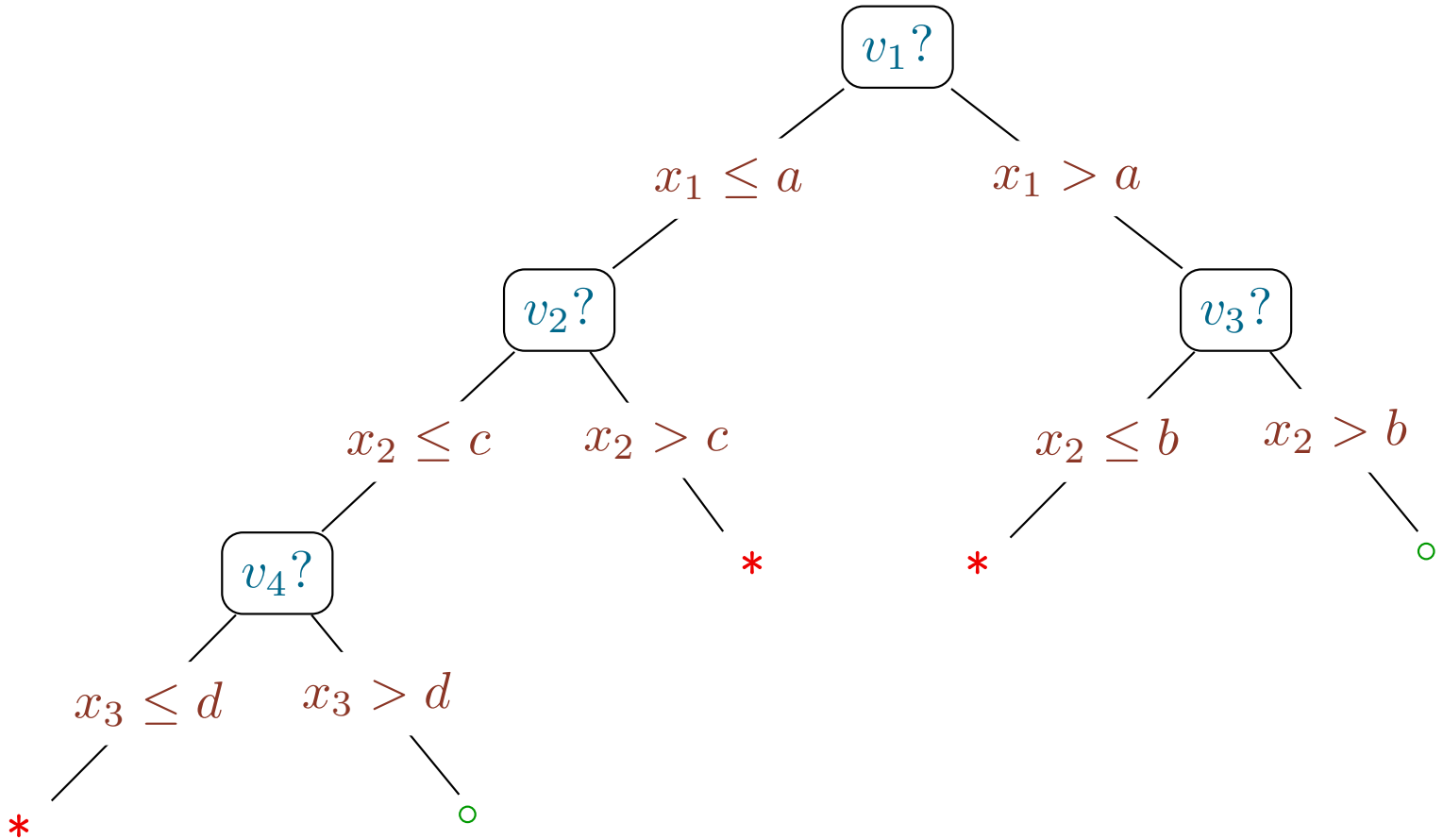


Discrétisation

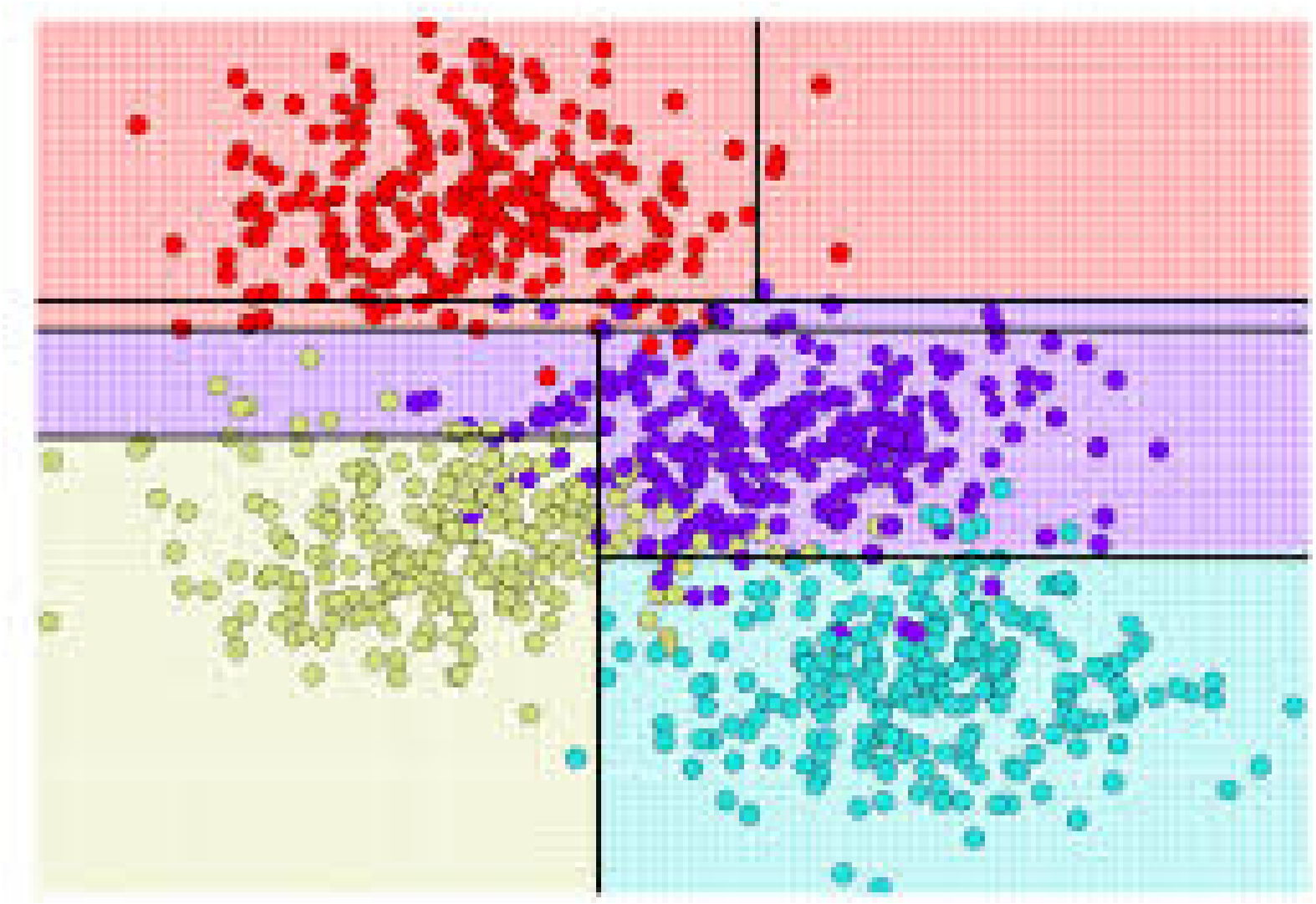
- Problématique
  - que faire si on a des attributs continus ?
- Solution
  - on discrétise : transformation d'une variable continue en une (ou plusieurs) variable discrète



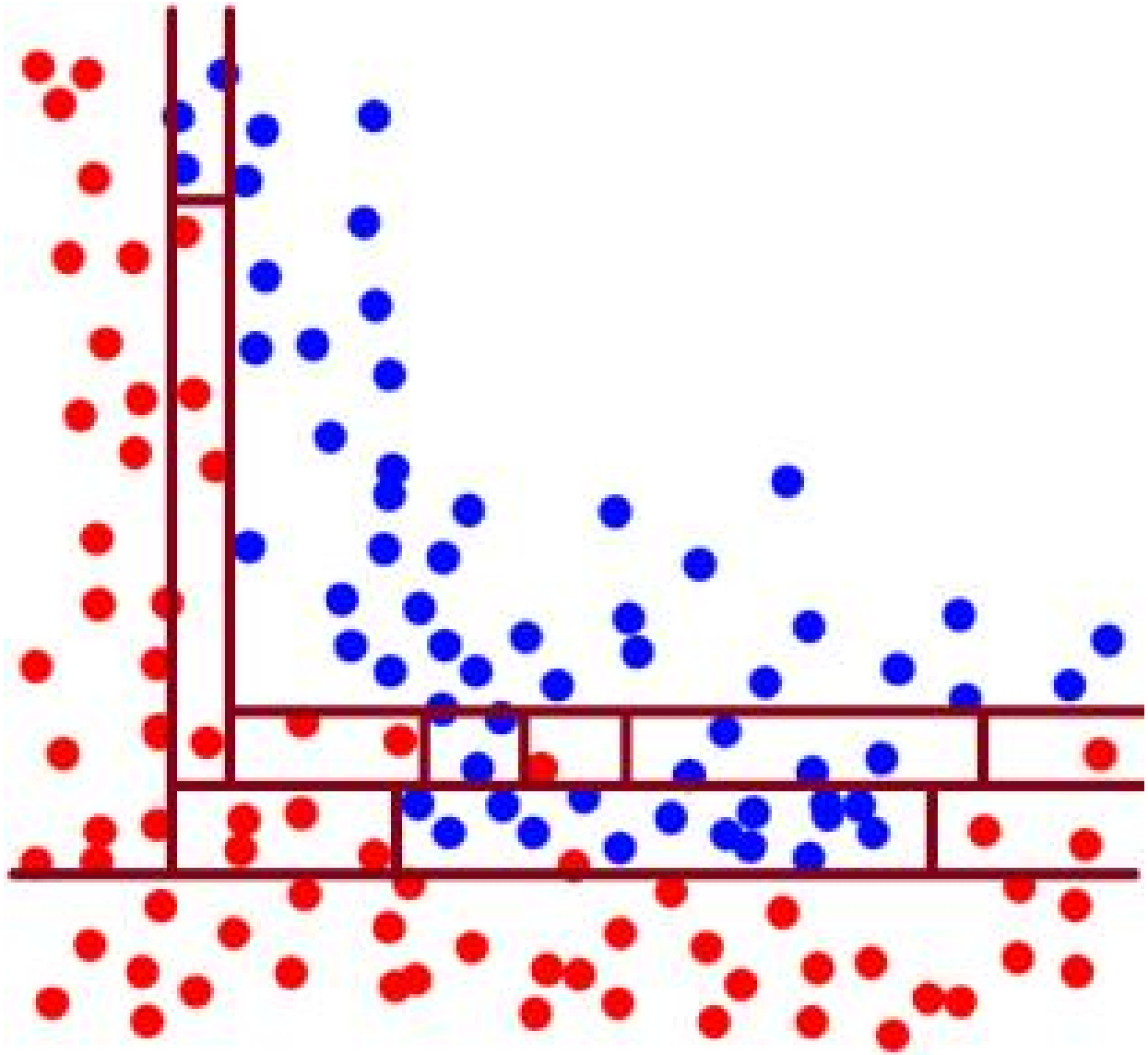
Exemple



Exemple



Exemple



Algorithme C4.5

1. Check for base cases
   - all samples in list belong to the same class
   - none of the features provide any information gain
   - instance of previously unseen class encountered
2. For each attribute A
   - find the normalised information gain ratio from splitting on A
3. Let A^* be the attribute with the highest information gain
4. Create a decision node that splits on A^*
5. Recur on the sublists obtained by splitting on A^*

Surapprentissage

- Élagage a posteriori
  - idée : élaguer après la construction de l'arbre entier
    - remplacer le sous-arbre optimisant un critère d'élagage par un nœud
    - nombreuses méthodes
    - encore beaucoup de recherche

- Minimal Cost-Complexity Pruning (MCCP) [Breiman et al., 84]
- Reduced Error Pruning (REP) [Quinlan, 87 & 93]
- Minimum Error Pruning (MEP) [Niblett & Bratko, 86]
- Critical Value Pruning (CVP) [Mingers, 87]
- Pessimistic Error Pruning (PEP) [Quinlan, 87]
- Error-Based Pruning (EBP) [Quinlan, 93]
- ...

Conclusion

- Arbres de décision appropriés pour :
  - classification de formes décrites en attributs-valeurs
  - attributs à valeurs discrètes
  - résistant au bruit
- Stratégie
  - recherche par construction incrémentale d'hypothèses
  - critère local (gradient) fondé sur critère statistique
- Engendre
  - arbre de décision interprétable (e.g. règles de production)
- Nécessite contrôle de la taille de l'arbre