eXplainable Artificial Intelligence

Cours 6 – mardi 24 octobre 2023

Marie-Jeanne Lesot
Christophe Marsala
Jean-Noël Vittaut
Gauvain Bourgne

LIP6 – Sorbonne Université

XAI – 2023-2024

Plan du cours

Interprétabilité

Quelques rappels
- notations
- représentation symbolique ou numérique

Approches interprétables

Arbres de décision

Forêts d'arbres de décision

Plan du cours

Explicabilité et interprétabilité

- Soit un modèle de prédiction f
- Différents niveaux d'interprétabilité
  - méthode de construction de f
  - modèle f
  - construction d'une prédiction faite par f
- Explications fournies pour comprendre une prédiction faite par f
- Construction d'un modèle
  - approches "boîtes noires"
    - réseaux de neurones, DL,...
  - approches interprétables "by design"
    - k plus proches voisins
    - raisonnement à partir de cas
    - système à base de règles
    - arbre de décision

Plan du cours

Interprétabilité

Quelques rappels
- notations
- représentation symbolique ou numérique

Approches interprétables

Arbres de décision

Forêts d'arbres de décision

Rappels : notations (1)

- Ensemble de n exemples (ou cas, ou individus) : $\mathbf{x}_1, \ldots, \mathbf{x}_n$
  - chaque individu $\mathbf{x}_i$ est décrit par d variables.
    $x_{i,j} \in \mathcal{U}_j$ (ou x_{ij}) est la valeur de la variable j pour l'exemple $\mathbf{x}_i$
- Base d'apprentissage (cas numérique)
  - ensemble d'exemples $\mathbf{X} \in \mathcal{U}_1 \times \ldots \times \mathcal{U}_d$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,d} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,d} \end{pmatrix}$$

  - apprentissage supervisé : les $\mathbf{x}_i$ sont associés à un label $y_i \in \mathcal{U}_\mathbf{Y}$
    - ensemble de labels associés à $\mathbf{X}$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Rappels : notations (2)

- Pour un seul exemple : $\mathbf{x} = (x_1, x_2, \ldots, x_d)$
- Terminologie : un label y_i = une classe
- Classifieur f : $f(\mathbf{x})$ est la classe donnée par f à l'exemple $\mathbf{x}$
  - cas binaire :
    - $f : \mathcal{U}_1 \times \ldots \times \mathcal{U}_d \longrightarrow \{-1, +1\}$
      $\mathbf{x} \longmapsto f(\mathbf{x})$
    - ou aussi : $f : \mathcal{U}_1 \times \ldots \times \mathcal{U}_d \longrightarrow \{l_1, l_2\}$ avec l_1 et l_2 deux labels
  - cas multiclasses :
    - $f : \mathcal{U}_1 \times \ldots \times \mathcal{U}_d \longrightarrow \{l_1, l_2, \ldots, l_k\}$

Types d'attributs

- Attributs catégoriels (aussi dits symboliques)
  - valeur binaire : {vrai, faux}, {féminin, masculin}, {+1, -1}, {0, 1}
  - nationalité : {français, chinois, marocain, kenyan, brésilien...}
  - tranche d'impôts : {1, 2, 3, 4, 5}
  - ...
- Attributs numériques
  - âge (d'une personne) : valeur (an) dans $[0, 120]$
  - longueur d'onde de la lumière visible : valeur (nm) dans $[380, 780]$
  - prix d'achat d'un livre de poche : valeur (euros) dans $[1.5, 15]$
  - ...

Ex.	âge	cheveux		groupe	Classe
		couleur	longueur		
$\mathbf{x}_1$	25	noir	18.7	2	+1
$\mathbf{x}_2$	37	roux	5.42	1	+1
$\mathbf{x}_3$	29	châtain	32.23	1	-1

Du catégoriel au numérique

- Comment utiliser des données catégorielles avec des classifieurs numériques ?
  - par exemple : perceptron, knn,...
- Transformer le catégoriel en numérique $\Longrightarrow$ encodage one hot
  - chaque attribut catégoriel est transformé
    - on remplace les catégories par autant de variables binaires $\{0, 1\}$
- Par exemple :
  - Pays = {France, Allemagne, Maroc, Japon}
  - création de 4 variables binaires : une pour France, etc...

Ex.	Pop.(m)	p_France	p_Allemagne	p_Maroc	p_Japon	Classe
$\mathbf{x}_1$	66.99	1	0	0	0	Europe
$\mathbf{x}_2$	83.02	0	1	0	0	Europe
$\mathbf{x}_3$	36.03	0	0	1	0	Afrique
$\mathbf{x}_4$	126.5	0	0	0	1	Asie

Plan du cours

Interprétabilité

Quelques rappels

Approches interprétables

- k plus proches voisins
- raisonnement à partir de cas
- système à base de règles

Arbres de décision

Forêts d'arbres de décision

Le plus simple des classifieurs interprétables

- L'algorithme des k plus proches voisins
  - pas de construction : on mémorise la base d'apprentissage $\mathbf{X}$
  - prédiction de la classe de $\mathbf{x}$
    1. trouver les k exemples de $\mathbf{X}$ les "plus proches" de $\mathbf{x}$
    2. agréger les classes des k exemples pour en déduire celle de $\mathbf{x}$
  - approche : proximité des exemples
    - utilisation d'une distance
- Interprétabilité : construction, compréhension et prédiction
  - classe de $\mathbf{x}$: donnée par les exemples qui lui sont le plus proche

Version générale des k-ppv

- Raisonnement à partir de cas (case-based reasoning)
- Attributs mixtes (numériques et/ou catégoriels)
  - comment calculer une distance ?
  - remplacer la distance par une mesure de similarité

    $s : \mathbb{X} \times \mathbb{X} \to [0, 1]$

    1) $s(\mathbf{x}, \mathbf{x}') = 1 \iff \mathbf{x} = \mathbf{x}'$ identité

    2) $s(\mathbf{x}, \mathbf{x}') = s(\mathbf{x}', \mathbf{x})$ symétrie

    avec $\mathbb{X}$: ensemble des exemples possibles (pas seulement ceux de $\mathbf{X}$)
- Apprentissage : mémoriser les exemples (base de cas)
  - plus général : exemple associé à une décision ou une "solution"
- Utilisation : solution pour $\mathbf{x}$ (prédiction, classification,...)
  1. trouver les cas les plus similaires
  2. adapter leurs solutions pour fournir un résultat pour $\mathbf{x}$
- Interprétabilité : construction, compréhension et prédiction
  - la solution pour $\mathbf{x}$ est une adaptation des solutions des éléments qui lui ressemblent

Systèmes à base de règles

- Règle d'inférence classique pour faire de la déduction
  - à partir de règles et d'observations, on déduit des conclusions

règle	:	V est A	$\longrightarrow$	U est C
observation	:	V est A		
conclusion	:			U est C

- Mécanisme du Modus Ponens
  - lien d'implication
  - l'observation implique l'inférence de la conclusion
- Énoncé : "si V est A alors U est C"
  - observer "V est A" : la proposition "V est A" est vraie
  - conclure "U est C" : la proposition "U est C" est vraie
- Vision classification :
  - V est A : $\mathbf{x}$ est A
    prémisse composée ($k \leq d$) : x_1 est A_1 et . . .et x_k est A_k
  - U est C : $\mathbf{Y}$ est y

Construction de bases de règles

- Connaissances expertes
- Apprendre à partir d'une base d'apprentissage $(\mathbf{X}, \mathbf{Y})$
  - cadre symbolique (IA historique)
  - cadre numérique surtout pour des modèles flous
- Quelques approches
  - règles d'association
  - tables de décision
  - arbres de décision
- Interprétabilité : construction, compréhension et prédiction
  - construction : approches par regroupements ou divisions
  - compréhension : une base de règles est interprétable
  - prédiction : lecture des règles qui se déclenchent
- Mais cela peut dépendre aussi de la taille de la base et de la complexité des règles...

Plan du cours

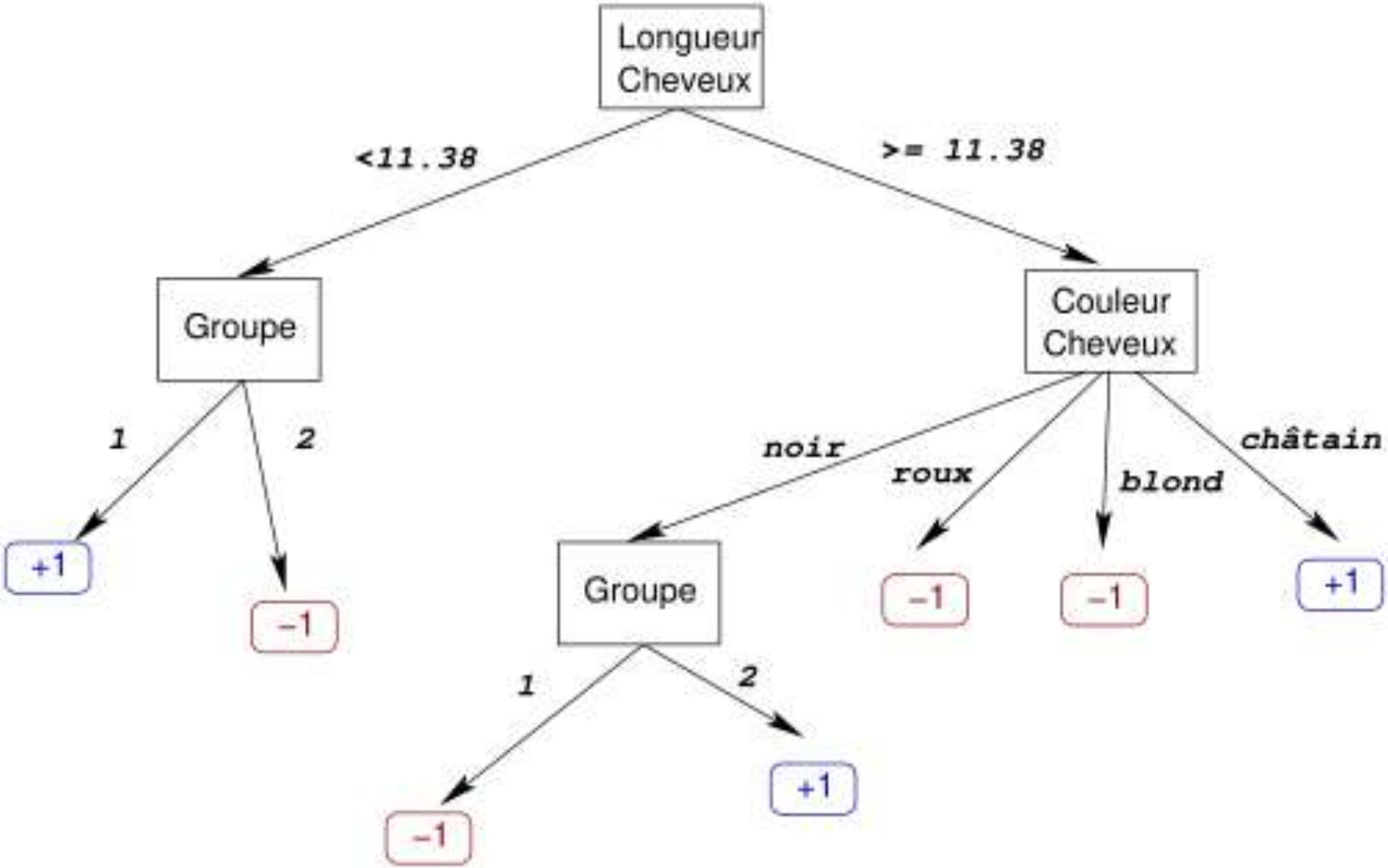
Arbres de décision

- Une forme de représentation des connaissances
- Représentation graphique et hiérarchique d'une base de règles
  - prémisses : nœuds internes d'une branche
  - conclusion : feuilles de l'arbre (décision/classe)
- Machine learning : méthodes inductives de construction d'arbres de décision
  - algorithme CART de Breiman's, Friedman's et al.'s

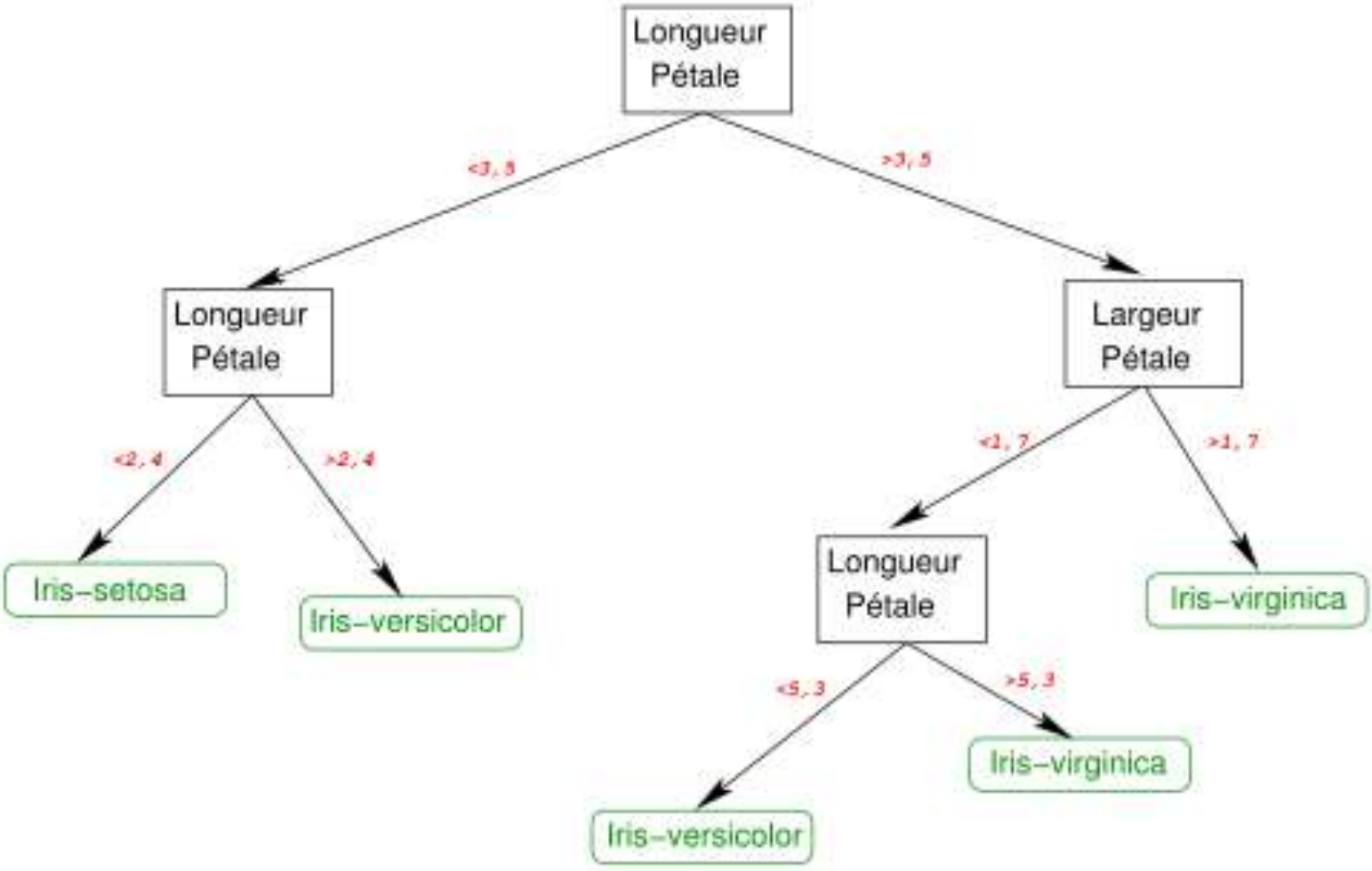

  - algorithme ID3 (puis C4.5) de Quinlan

- Caractéristiques de ces algorithmes
  - simplicité, rapidité
  - approche formelle

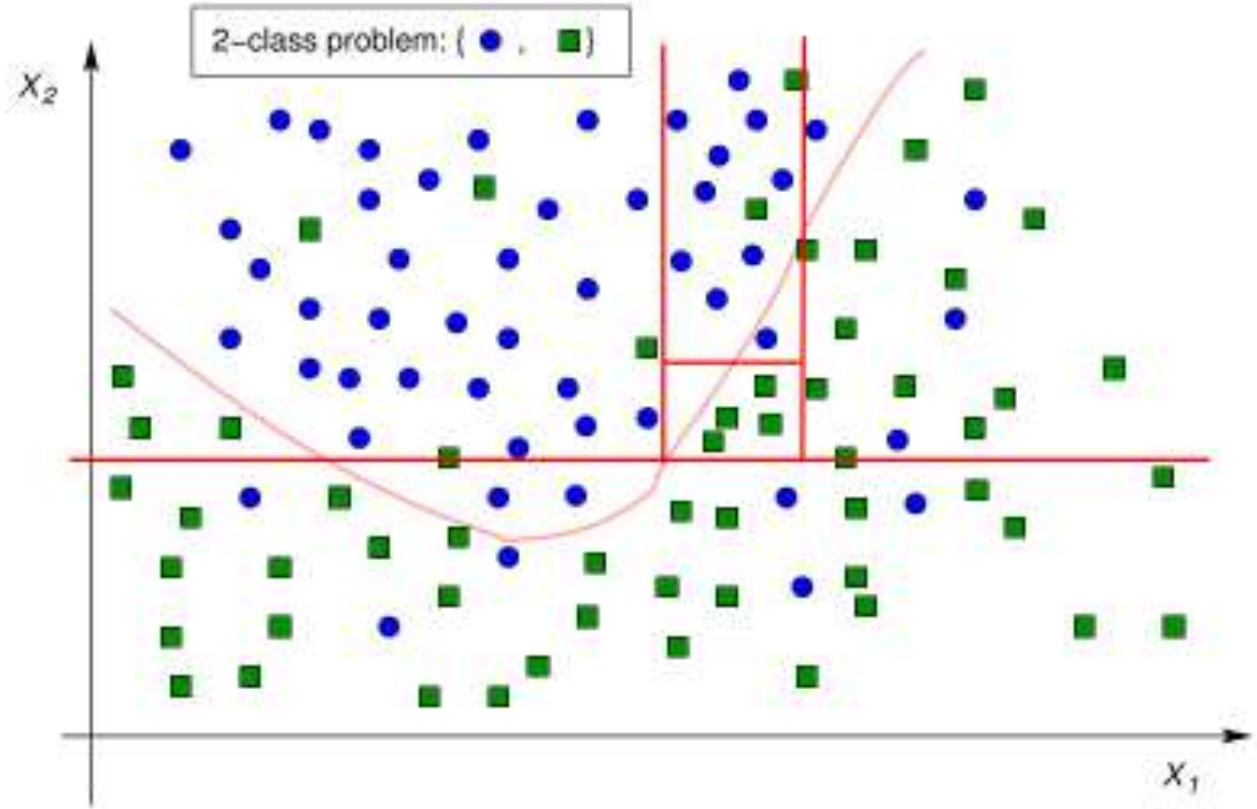
Exemple d'arbre de décision (général)



Exemple d'arbre de décision (numérique)

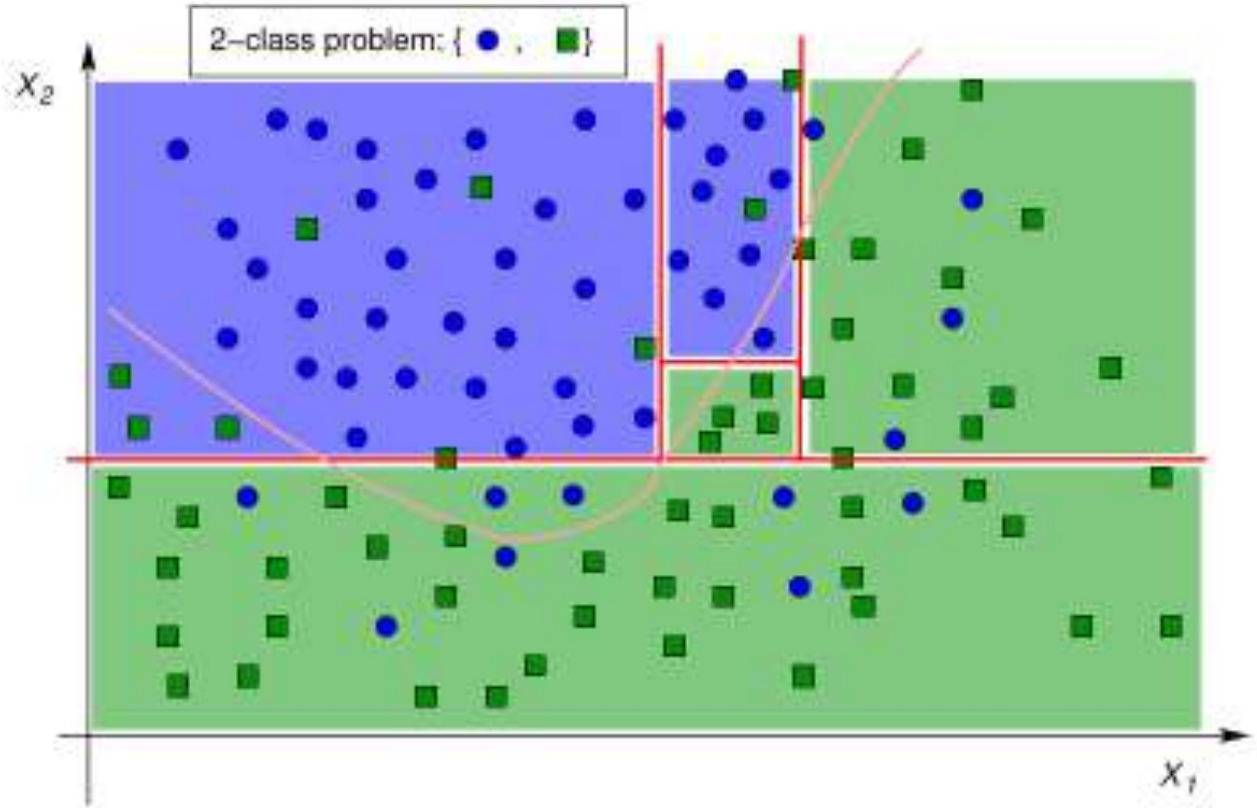


Frontières fournies par un arbre de décision (1)



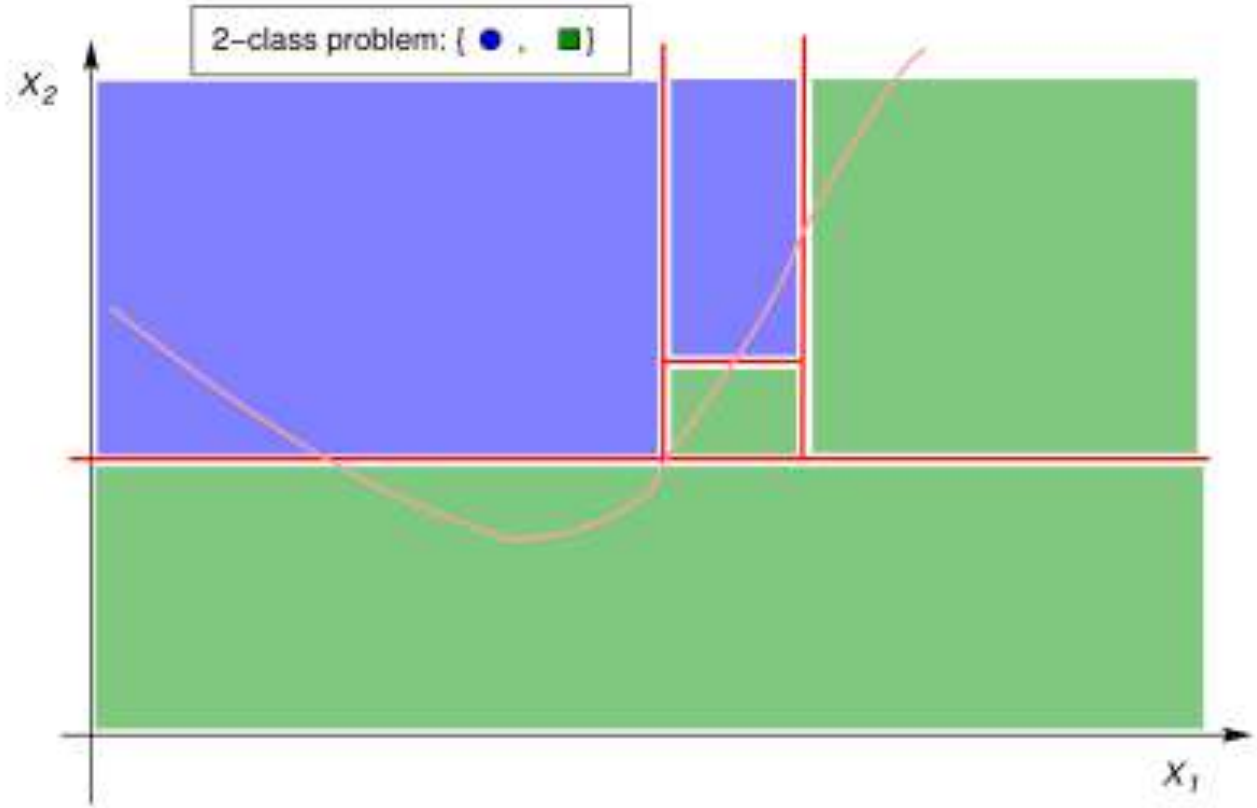
- Un arbre de décision définit un découpage par des frontières parallèles aux axes
  - chaque frontière est définie par un test d'un nœud de l'arbre

Frontières fournies par un arbre de décision (2)



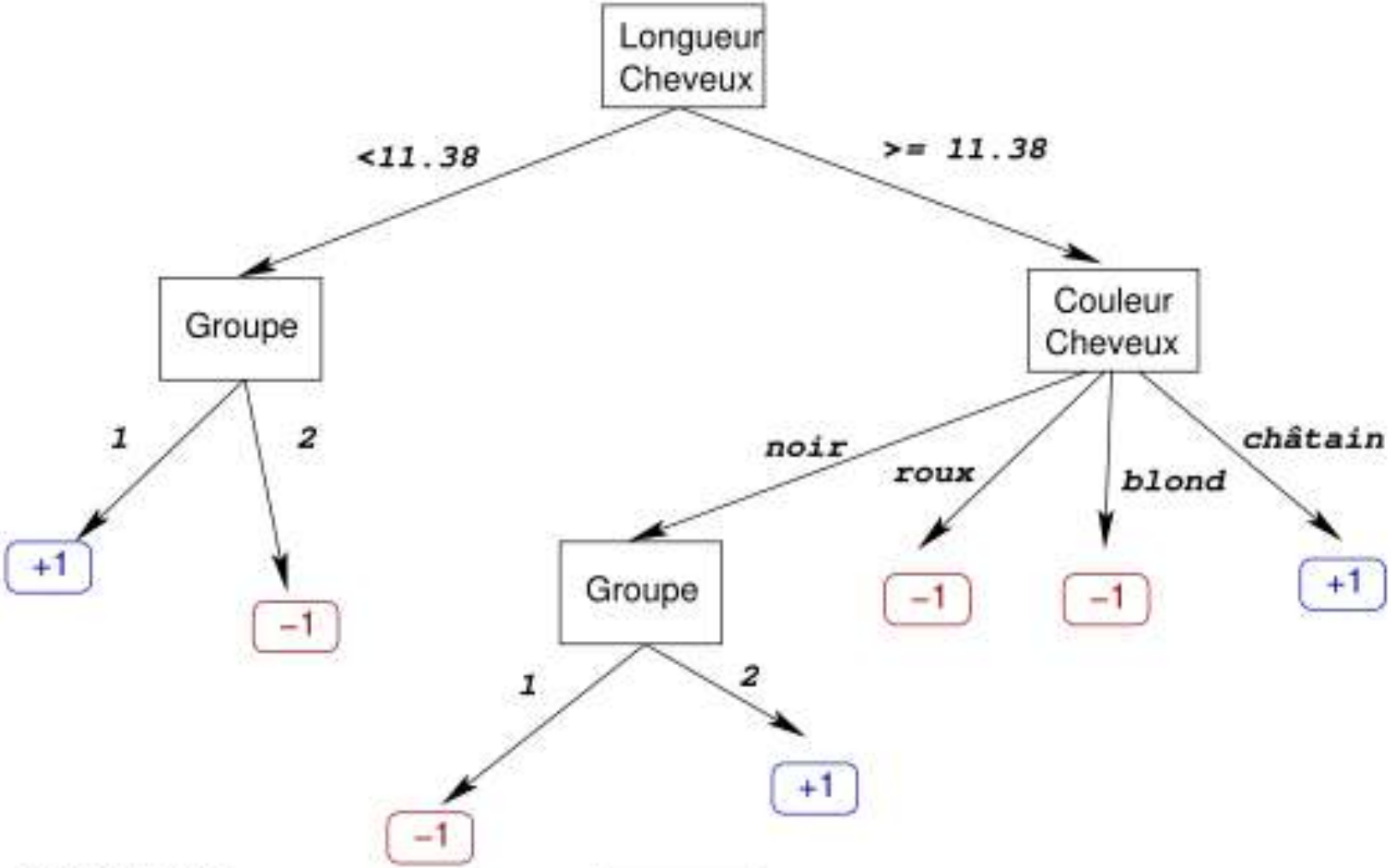
- Les frontières définissent des régions de décision
  - découpage précis des classes

Frontières fournies par un arbre de décision (3)

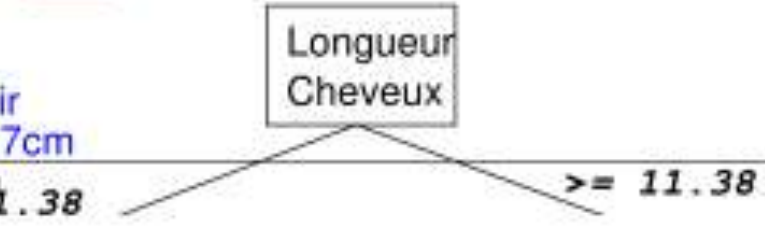


- Les régions servent à classer de nouveaux exemples

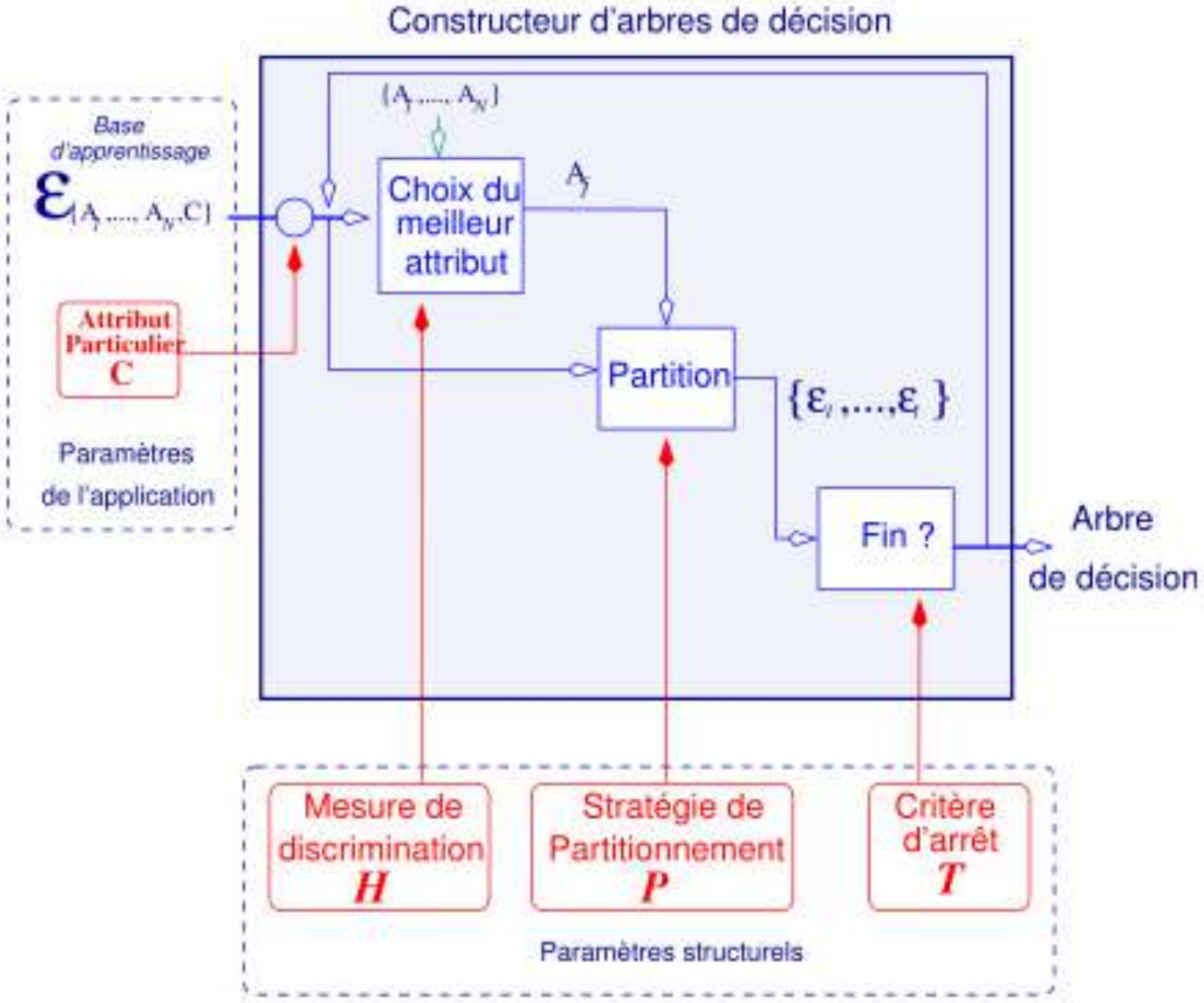
Classification avec un arbre de décision



nouveau cas:
âge : 31 ans
couleur cheveux : noir
longueur cheveux : 17cm
groupe : ?



Apprentissage d'un arbre de décision ; algorithme générique



Sélection d'attributs : mesure de discrimination

- Paramètre clé des algorithmes top down
  - ordonnancement optimal des questions pour déterminer la classe
- Choix d'une bonne mesure de discrimination
  - pour obtenir des nœuds cohérents
  - pour minimiser la taille de l'arbre (heuristique)
  - pour obtenir de bons résultats en classification
- Arbres de décision
  - mesure d'impureté (CART) : index de Gini, entropie de Shannon
  - théorie de l'information (ID3) : entropie de Shannon
- Entropie de Shannon : mesure un taux de désordre

$$H_S(X) = -\sum_{x \in X} p(x) \log(p(x))$$

  - mesure issue de la théorie de l'information
  - initiée par C.E. Shannon en 1948

Mesure d'entropie conditionnelle

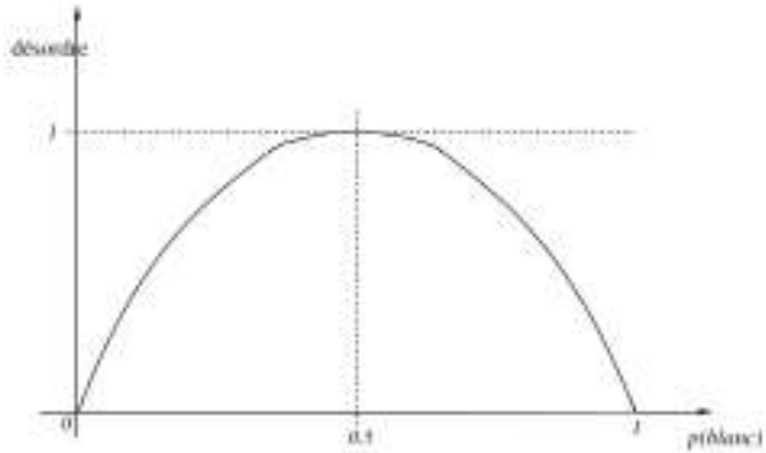
- Avec la base d'apprentissage $(\mathbf{X}, \mathbf{Y})$
  - $X = \{v_1, \ldots, v_m\}$: ensemble des valeurs prises dans $\mathbf{X}$ par un attribut (= une colonne de $\mathbf{X}$)
  - $Y = \{c_1, \ldots, c_K\}$: ensemble des valeurs différentes de $\mathbf{Y}$
- Utilisation de l'entropie de Shannon : forme conditionnelle

$$H_S(Y|X) = -\sum_i p(v_i) \sum_k p(c_k|v_i) \log(p(c_k|v_i))$$

- $H_S(Y|X)$ mesure du pouvoir de discrimination de l'attribut A envers la classe Y
  - X est discriminant pour Y si pour tout i, la connaissance de la valeur v_i de X permet d'en déduire une valeur unique c_k de Y

Relation entre probabilité et désordre

- Cas binaire : 2 classes (blanc ou noir)
  - $p(\text{noir}) = 1 - p(\text{blanc})$

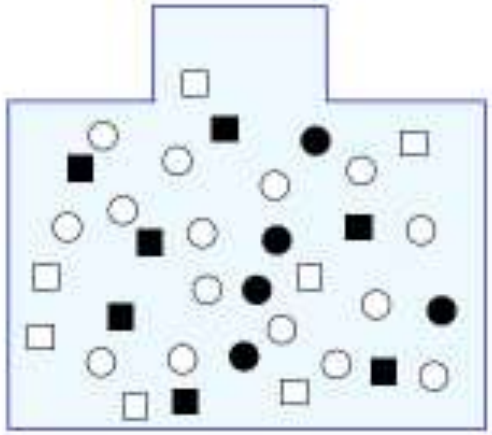


- Entropie de Shannon : $H_S(X) = -\sum_{x \in X} p(x) \log(p(x))$

$$H_S(\text{urne}) = -p(\text{blanc}) \log(p(\text{blanc})) - p(\text{noir}) \log(p(\text{noir}))$$

  - $H_S(\text{urne}) = 0$ quand $p(\text{blanc}) = 1$ ou quand $p(\text{blanc}) = 0$
  - $H_S(\text{urne}) = 1$ quand $p(\text{blanc}) = p(\text{noir}) = 0.5$

Désordre moyen et choix d'un attribut

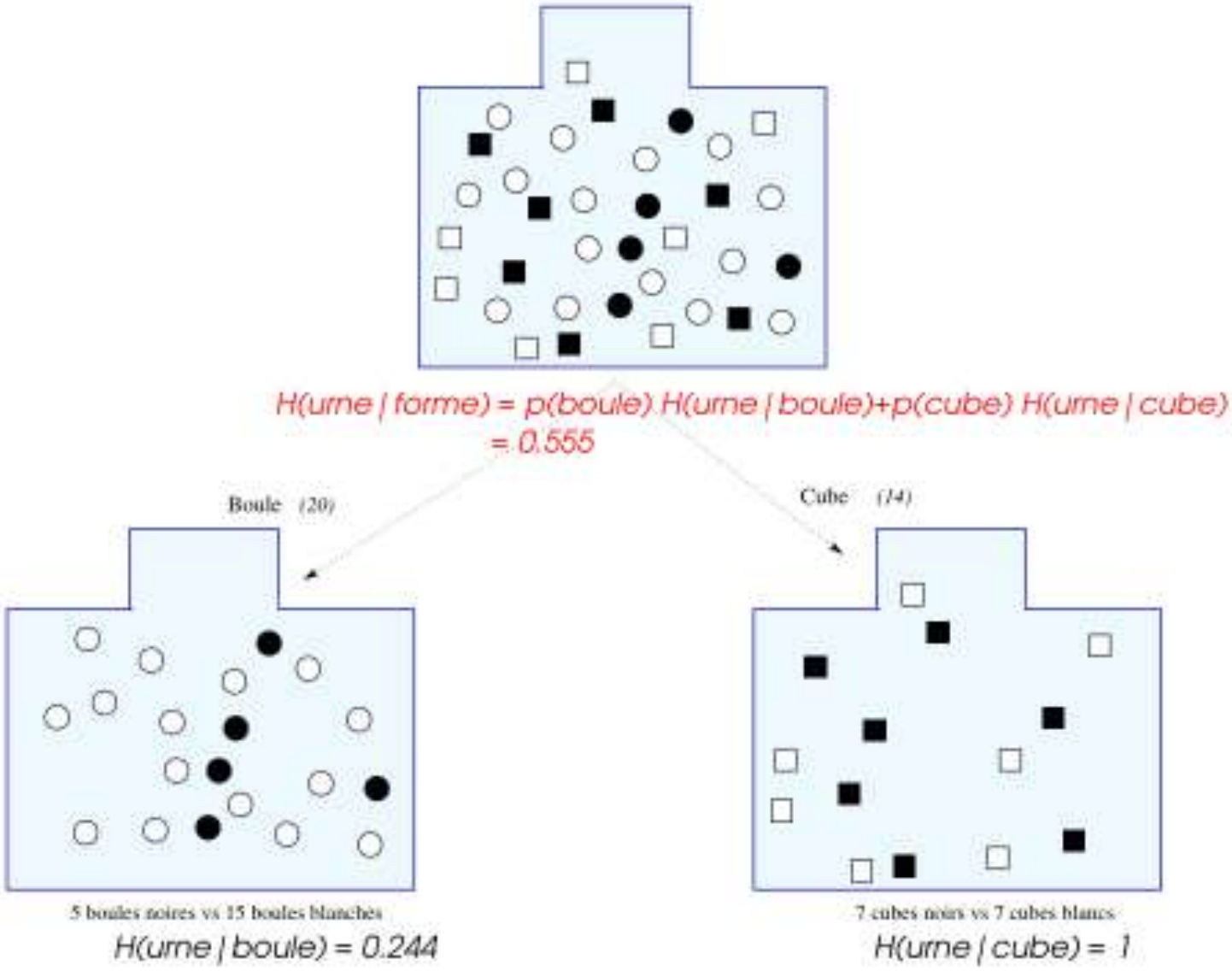


- Objectif : prédire la couleur de la boule / cube tiré
- Quelle stratégie pour mieux prédire ?
  - tirer "quelque chose" est prédire sa couleur
  - tirer une boule est prédire sa couleur
  - tirer un cube est prédire sa couleur
- Entropie de l'urne :

$$H(\text{urne}) = -p(\text{blanc}) \log(p(\text{blanc})) - p(\text{noir}) \log(p(\text{noir}))$$

soit $H(\text{urne}) = -\frac{22}{34} \log \frac{22}{34} - \frac{12}{34} \log \frac{12}{34} = 0.649$

Désordre moyen et choix d'un attribut



Désordre moyen et choix d'un attribut : bilan

- Entropie de l'urne : 0.649
- Entropie de l'urne connaissant la forme : 0.555
- Gain d'information apporté par la connaissance de la forme $0.649 - 0.555 = 0.094$
- Il est intéressant d'utiliser la forme pour prédire !

Gain d'information (1)

- Choix du meilleur attribut pour partitionner la base
  - la partition se fait sur ses valeurs
  - chaque valeur de l'attribut définit un sous-ensemble des exemples
- À l'aide d'une mesure de discrimination
  - choisir l'attribut qui apporte le plus d'information pour améliorer la connaissance de la classe
  - c'est-à-dire celui qui maximise le gain d'information

$$I_S(X, Y) = H_S(Y) - H_S(Y|X)$$

  - $H_S(Y)$: entropie de la base selon les valeurs de la classe
    - vaut 0 si tous les exemples de la base ont la même classe
    - vaut 1 si équi-répartition des différentes valeurs de la classe
  - $H_S(Y|X)$: pouvoir de discrimination de X relativement à Y
  - $I_S(X, Y)$: gain apporté par un découpage de la base selon les valeurs de X

Gain d'information (2)

$$I_S(X, Y) = H_S(Y) - H_S(Y|X)$$

- On cherche l'attribut X qui maximise $I_S(X, Y)$
- En pratique : $H_S(Y)$ est le même pour tous les attributs
- On cherche donc l'attribut X qui minimise $H_S(Y|X)$

Exemples de mesures de discrimination (1)

- Étant donné une base d'apprentissage $(\mathbf{X}, \mathbf{Y})$
  - $X = \{v_1, \ldots, v_m\}$: attribut de description (colonne de $\mathbf{X}$)
  - $Y = \{c_1, \ldots, c_K\}$: ensemble des valeurs possibles pour la classe/le label dans $\mathbf{Y}$ (par exemple, $C = \{-1, +1\}$)
- Entropie de Shannon [Shannon, 1948]

$$H_S(Y|X) = -\sum_{j=1}^{m} p(v_j) \cdot \sum_{k=1}^{K} p(c_k|v_j) \log p(c_k|v_j)$$

- Mesure d'information

Exemples de mesures de discrimination (2)

- Rapport de gain (Gain Ratio) [Quinlan, 1986]

$$H_R(Y|X) = \frac{H_S(Y|X)}{H_S(X)}$$

- Mesure d'information
  - utilisée en présence d'attributs symboliques
  - pénalise un grand nombre de valeurs pour X

Exemples de mesures de discrimination (3)

- Indice de diversité de Gini [Gini, début 19e]

$$H_G(Y|X) = \sum_{j=1}^{m} p(v_j) \cdot (1 - \sum_{k=1}^{K} p(c_k|v_j)^2)$$

- Mesure du niveau d'adéquation entre les 2 répartitions X et Y

Exemples de mesures de discrimination (4)

- Mesure d'ambiguïté [Yuan & Shaw, 1995]

$$H_Y(Y|X) = \sum_{j=1}^{m} p(v_j) \cdot g(\Pi(Y|v_j))$$

- avec g une mesure de non-spécificité

$$g(\Pi(C|v_j)) = \sum_{i=2}^{K} \pi_i \cdot (\log(i) - \log(i-1))$$

- où π est obtenue de la façon suivante :
  1. on ordonne les $p(c_k|v_j)$ dans l'ordre décroissant :
     on note p_1 la plus grande probabilité, puis p_2 la suivante, etc.
     jusqu'à p_K la plus petite des probabilités
  2. on définit $\pi_i = \frac{p_i}{p_1}$ pour tout $i = 1, \ldots, K$

Validation d'une mesure de discrimination

- Comment caractériser une bonne mesure de discrimination ?
- Cadre classique : théorie de l'information
  - mesure de l'information apportée par un événement, etc.
  - mesure de basée sur l'impureté des sous-ensembles

Construction de l'arbre : algorithme classique (symbolique)

- Mettre la base d'apprentissage dans la pile à traiter
- Tant qu'il y a des ensembles dans la pile à traiter : prendre un ensemble
  - si le critère d'arrêt est atteint alors créer une feuille
  - sinon
    1. calculer $H(Y|X_j)$ pour tous les attributs X_j
    2. choisir l'attribut X_j qui minimise $H(Y|X_j)$
    3. créer un nœud dans l'arbre de décision avec X_j
    4. partitionner la base avec les valeurs de X_j
    5. mettre les ensembles dans la pile à traiter

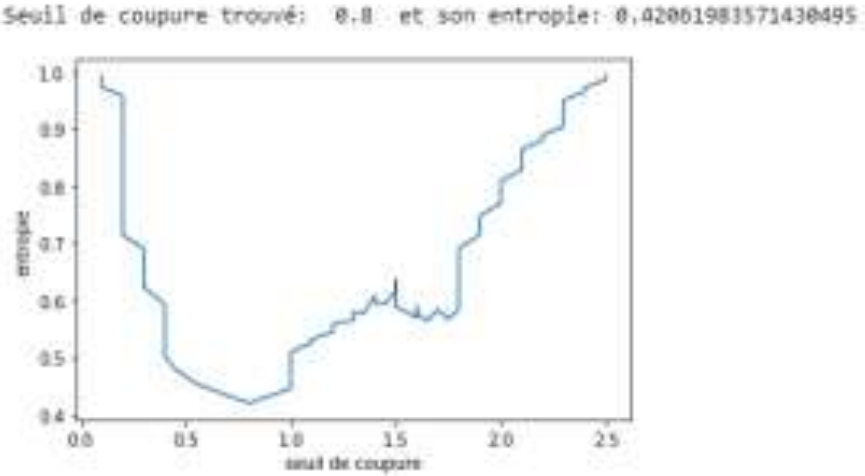
Critère d'arrêt de la construction de l'arbre

- Quelques exemples de critères d'arrêt
  - tous les exemples de la base ont la même classe
  - utilisation d'une tolérance : la plupart des exemples ont la même classe
    - utilisation d'un seuil $\varepsilon \in [0, 1]$
    - on calcule $H(Y) = -\sum_{k=1}^{K} p(c_k) \log p(c_k)$ et on s'arrête si $H(Y) \leq \varepsilon$
  - tous les attributs ont été utilisés une fois dans le cas symbolique
  - trop peu d'exemples dans l'ensemble traité

- Création d'une feuille de l'arbre de décision

Traitement des attributs numériques

- X_j, attribut numérique
  - utilisation d'une valeur de coupure v_j
  - construction de 2 intervalles : $]-\infty, v_j[$ et $[v_j, +\infty[$
  - on note : $\{X_j, v_j\}$ cette décomposition
- On détermine la valeur v_j qui minimise $H(Y|\{X_j, v_j\})$
  - phase de discrétisation : recherche exhaustive
  - on ensuite traite l'attribut comme un attribut catégoriel



Un même attribut numérique peut intervenir plusieurs fois dans l'arbre final avec des seuils de coupure différents

Conclusion sur les arbres de décision

- Ce dont on n'a pas parlé
  - élagage post-construction (pruning)
    - très performant pour les arbres "classiques"
    - négligeable pour les arbres flous
  - approches de discrétisation des attributs numériques
  - arbres de régression
    - la sortie est une variable continue
- Modèle très pratique et très utilisé
  - modèle interprétable (selon sa taille...)
    - construction de l'arbre et utilisation de l'arbre
    - représentation des connaissances
  - modèle proche des modèles humain
  - bon compromis efficacité / utilisabilité
- Quelques critiques...
  - taux de reconnaissance
  - frontières parallèles aux axes
  - difficultés avec les classes déséquilibrées

Le problème des classes déséquilibrées

- Difficulté à gérer des distributions déséquilibrées entre les classes
  - par exemple, application médicale : 10% de malades, 90% de non malades
    - prédire "non malade" : taux de bonne classification de 90%
    - mais... c'est très peu informatif en fait...
  - si on construit un arbre de décision : selon le critère d'arrêt choisi, un seul nœud peut être suffisant !

- Il faut donc généralement
  - soit avoir un modèle d'apprentissage qui en tienne compte
  - soit équilibrer les classes avant l'apprentissage

Adapter les arbres aux classes déséquilibrées

- Extraire des échantillons équilibrés de la base d'apprentissage
  - sélection aléatoire de N exemples pour chaque valeur de la classe
  - construire un arbre avec cet échantillon
- Ce processus peut être répété plusieurs fois
  - ensembles d'arbres de décision (forêts)

Plan du cours

Interprétabilité

Quelques rappels

Approches interprétables

Arbres de décision

Forêts d'arbres de décision
- différentes approches
- résumer des forêts

Multiplier pour améliorer

- Idée : améliorer les performances en multipliant les modèles
  - ensemble learning
- Constituer un ensemble d'apprenants
  - plusieurs apprenants spécialistes localement
- Inciter la diversité des apprenants de l'ensemble
  - obtenir une couverture de l'espace de représentation efficace
- Contrainte : maintenir l'interprétabilité
- Dans ce qui suit :
  - un apprenant : arbre de décision ou classifieur
  - construction d'ensemble d'apprenants
  - utilisation de l'ensemble construit

Ensembles d'arbres de décision (1)

- Boostrap
  - approche utilisée pour quantifier l'incertitude d'une prédiction
  - sélection aléatoire d'échantillon d'exemples pour évaluer le modèle
  - l'échantillonnage est réalisé un grand nombre de fois
- Bagging : boostrap aggregating
  - apprentissage : tirage aléatoire de sous-bases d'apprentissage
    - construction d'un arbre
  - combinaison des résultats de tous les arbres
    - classification d'un exemple par l'ensemble
- Boosting
  - combinaison de classifieurs "faibles"
  - construction itérative en tenant compte des exemples mal classés

Ensembles d'arbres de décision (2)

- Random forest
  - but : diversifier le plus possible les arbres construits
  - sélection aléatoire dans des sous-espaces de l'espace de description des exemples
    - sélection aléatoire des dimensions utilisables à chaque niveau de la construction de l'arbre
    - parmi les dimensions choisies : choix classique
- Variante "extrême" : extremely random ensemble
  - tous les choix sont faits aléatoirement

Approche bagging d'arbres

- Construction de la forêt : approche bootstrap
  - sélection aléatoire de N exemples
    - classique : tirage avec remise
    - variante : tirage sans remise
  - construction d'un arbre avec cet échantillon
  - on réitère ce processus pour multiplier les arbres de la forêt
- Utilisation de la forêt pour classer $\mathbf{x}$
  - chaque arbre fournit une classe pour $\mathbf{x}$
  - agrégation de ces degrés pour obtenir la classe de $\mathbf{x}$
- Interprétabilité
  - construction : intuitive mais peut être vu comme complexe
  - compréhension : dépend de la taille de la forêt
  - prédiction : complexe au vu de la taille

Résumer des forêts

- Efficacité des approches d'ensembles [Breiman 1996, Geurts et al. 2006, ...]
  - mais... stockage coûteux, classification complexe,
  - perte d'interprétabilité ?
  - pourtant "plusieurs peuvent être meilleurs que tous" .. ;
- Solution : résumer un ensemble mais en préservant la diversité
  - mesurer la qualité d'un classifieur en prenant en compte la précision et la diversité
  - précision : taux de bonne classification
  - diversité : couverture de l'espace de recherche

Mesure de la diversité de classifieurs : mesures "pairwise"

(pour en savoir plus et réfs : [Zhou, 2012])

- On considère une base de n exemples donnée
- Soit un ensemble de classifieurs $f_1, \ldots, f_m$
  - pour un exemple x donné : $f_i(x)$ classe de x trouvée par i
- Mesures "pairwise" :
  - utiliser une comparaison des classifieurs 2 à 2
  - table de contingence :

	$f_j(x) = 0$	$f_j(x) = 1$
$f_i(x) = 0$	n_{00}	n_{01}
$f_i(x) = 1$	n_{10}	n_{11}

  - mesure de désaccord entre f_i et f_j : $\frac{n_{10} + n_{01}}{n}$
  - Q-statistique : $\frac{n_{00}n_{11} - n_{10}n_{01}}{n_{00}n_{11} + n_{10}n_{01}}$
  - test du Kappa, etc.

Exemple : test du κ

[Cohen, 1960]

- Accord de 2 classifieurs (par rapport au hasard)
- Table de contingence :

	$f_j(x) = 0$	$f_j(x) = 1$
$f_i(x) = 0$	n_{00}	n_{01}
$f_i(x) = 1$	n_{10}	n_{11}

on note $n = n_{00} + n_{01} + n_{10} + n_{11}$

- Mesure de l'accord entre f_i et f_j : $\text{Accord}(f_i, f_j) = \frac{n_{00} + n_{11}}{n}$
- Probabilités de classer $c \in \{0, 1\}$ pour f_i et f_j :

$$p_i(c) = \frac{1}{n}(n_{c0} + n_{c1}) \text{ et } p_j(c) = \frac{1}{n}(n_{0c} + n_{1c})$$

- Probabilités que les 2 classifieurs soient d'accord pour $c \in \{0, 1\}$: $p_{ij}(c) = p_i(c)p_j(c)$

Exemple : test du κ

[Cohen, 1960]

- Concordance entre f_i et f_j :

$$\kappa = \frac{\text{Accord}(f_i, f_j) - (p_{ij}(0) + p_{ij}(1))}{1 - (p_{ij}(0) + p_{ij}(1))}$$

- Interprétation du résultat [Landis& Koch, 1977]

Mesure de la diversité de classifieurs : mesures globales

(pour en savoir plus et réfs : [Zhou, 2012])

- Soit un ensemble de classifieurs $f_1, \ldots, f_m$
  - pour un exemple x donné : $f_i(x)$ classe de x trouvée par i
- Mesures globales :
  - mesure directe de la diversité globale de l'ensemble
  - exemple : mesure d'entropie
    - on ne tient pas compte de la correction de la classification
    - $P(c_l|x_k)$: proportion de classifieurs qui prédisent c_l pour x_k
    - $-\frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{q} P(c_l|x_k)\log P(c_l|x_k)$
  - exemple : [Kuncheva and Whitaker, 2003]
    - $N_C(x)$: nombre de classifiers qui trouvent la bonne classe de x
    - mesure $\frac{1}{nm^2}\sum_{k+1}^{n} N_C(x_k)(m - N_C(x_k))$
    - plus la valeur est élevée, plus l'ensemble est diversifié

Résumé un ensemble : optimiser le nombre de classifieurs

(pour en savoir plus et réfs : [Zhou, 2012])

- Éliminer les classifieurs superflus
  - comment les trouver ?
- Différentes approches
  - par ordonnancement selon une mesure donnée
    - réduction de l'erreur, Kappa,...
  - par clustering
    - représentation vectorielle et distance
    - clustering hiérarchique, k-moyennes
    - prototypes issus des clusters
  - par optimisation
    - algorithmes génétiques,...