Support de cours en Fouille de données
Chapitre 5 : Les arbres de décision (classification)

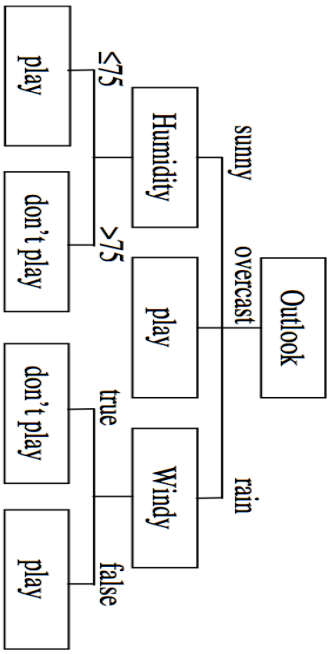
1. **Introduction :** les arbres de décision sont un moyen d'apprentissage supervisé qui permet de séparer des individus dans des groupes selon des règles ou de prévoir la valeur d'une variable continue (cible) à partir de variables en entrée. L'apprentissage est supervisé car les classes ou variables à prévoir sont prédéfinies. On cherche alors à induire un ensemble de règles à appliquer à une nouvelle instance afin de déterminer sa classe d'appartenance ou à connaître la valeur de la variable. L'appellation *arbre* provient du fait que les règles s'enchaînent de sorte que chaque règle correspond à un test donnant lieu à un nœud et les alternatives de réponse au test donnent lieu aux branches. Lorsqu'un arbre de décision est utilisé pour prédire une variable qualitative, on parle *d'arbre de classification*. Par contre, lorsqu'il est utilisé pour prédire des variables continues, on parle *d'arbres de régression*. Dans ce cours, nous nous intéressons au premier cas.

2. **Exemples d'utilisation:**

    a. Connaître le degré risque d'un demandeur de prêt ?
    b. Connaître le risque qu'une population attrape une maladie
    c. Reconnaître n objet détecté par un radar (véhicule, individu, immeuble, arbre).
    d. Etablir le degré de ressemblance de personnes à un criminel recherché

3. **Exemple illustratif :** on prend l'exemple classique du joueur de golf qui décide de jouer ou pas sur la base du temps qu'il fait. Le tableau repris dans ce cas est illustré ci-dessous à gauche. Dans cet exemple, chaque nouveau jour représente un individu sur lequel on veut décider comme jour de jeu ou de non jeu (classes *jouer* et *ne pas jouer)*. Les autres variables sont des variables en entrée. La figure à droite illustre un arbre de classification qu'on peut facilement interpréter. On remarque que selon les données qui ont servi à induire l'arbre, la variable température n'a pas été utilisée.

4. **Formulation du problème :** les données nécessaires à l'induction de règles de décision sont appelées ensemble d'apprentissage (training set). Elles se présentent sous une forme tabulaire *individu*/*attribut* et peuvent être qualitatives (e.g. Outlook, Windy, Class) ou numériques (Temp et Humidity). On distingue une variable pour être la variable à prédire. Dans l'exemple ci-dessous, il s'agit de la variable Class avec deux modalités : *play* et *don't play*. Le problème consiste à élaborer un arbre où chaque nœud (appelé nœud de décision) consiste en un test sur la valeur d'un attribut. Les feuilles de l'arbre correspondent aux valeurs de l'attribut de prédiction. Les tests de valeur d'attributs diffèrent selon la nature des attributs (qualitatifs ou numériques).



Outlook	Temp (°F)	Humidity (%)	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

5. **Algorithme de base pour la construction d'un arbre de décision :** avant de présenter l'algorithme, les données d'apprentissage doivent satisfaire les conditions suivantes (i) avoir une forme tabulaire individu/attribut, (ii) la classe cible doit être prédéfinie (iii) des données suffisantes (centaine, millier d'individus ou plus).

   a. **Principe de l'algorithme :** l'algorithme est récursif. A partir d'un nœud initial, on choisit à chaque étape un attribut pour la division de l'ensemble d'apprentissage en sous-ensembles : chaque sous-ensemble correspond à l'attribut sélectionné et donne lieu à un nœud fils. Chaque branche du père vers le fils correspond à une valeur de l'attribut sélectionné. L'algorithme s'arrête lorsqu'il n'y a plus d'attributs de sélection.

   b. **Présentation de l'algorithme :** Dans l'algorithme CLS de Hoveland et Hunt (1950), les valeurs de la classe de prédiction sont (+) et (-).
      i. A l'ensemble d'apprentissage E. Créer un nœud de A.
      ii. Si tous les exemples de E ont la valeur positive pour la classe de prédiction, alors créer un nœud P sous la racine et s'arrêter.
      iii. Si tous les exemples de E ont la valeur négative, alors créer un nœud N sous la racine et s'arrêter.
      iv. Sélectionner un attribut X ayant les modalités v_1, v_2, …v_N et partitionner E en N sous-ensembles E_i, chacun correspondant à une modalité. Pour chaque sous-ensemble, créer un nœud ayant comme libellé la modalité v_i.
      v. Pour chaque sous-ensemble E_i, E E_i, aller à ii.

6. **Problème de sélection des meilleurs (attributs) classifieurs :** l'algorithme précédent ne spécifie pas le critère de sélection des attributs à chaque étape. Ce choix peut être aléatoire, ou basé sur un critère particulier. Les arbres obtenus peuvent être différents en profondeur. Le choix du classifieur détermine l'efficacité de l'induction à partir de l'arbre de décision. On cite deux choix : l'utilisation de l'entropie et gain informationnel, et l'utilisation des fréquences.

7. **Sélection à base d'entropie / Algorithme ID3 (iterative dichotomiser):** l'entropie telle que définie par Shannon mesure la quantité d'information attendue lors de l'envoi d'un message. Elle est liée à l'incertitude par rapport à un phénomène donné. L'exemple courant et celui du jet d'une pièce de monnaie avec deux valeurs possibles (pile ou face) : l'entropie est maximale lorsque la pièce n'est pas truquée, i.e. les deux faces ont la même probabilité. L'entropie est nulle lorsqu'on possède une certitude (si la pièce contient deux faces *piles*). L'entropie est utilisée dans l'estimation du gain information (ou réduction d'incertitude) par rapport à une valeur de la classe à prédire. L'attribut choisi à une étape est celui qui minimise l'incertitude par rapport aux valeurs de la classe à prédire. La réduction d'incertitude (d'entropie) est également appelée le gain informationnel.

   a. **Entropie de Shannon :** soit un ensemble de mots à coder m_1, m_2, …sur un canal binaire C, ayant des probabilités d'apparition respective p_1, p_2, …. L'entropie est donnée par la formule

   $$\text{Entropie(C)} = \sum_{i=1}^{c} - p_i \log_2 p_i$$

   b. **Analogie avec les arbres de décision :** par rapport aux arbres de décision, chaque valeur d'un attribut donné correspond à un mot dans l'entropie de Shannon. La probabilité de chaque valeur est calculée par rapport à un sous-ensemble de données courant, i.e. celui obtenu par la division des donnée.

   c. **Utilisation de l'entropie / gain informationnel pour la sélection du meilleur classifieur :** pour décider quel attribut choisir comme attribut de décision à une étape donnée, on calcule pour chaque attribut le gain informationnel. Soit S l'ensemble courant de données et soit A un attribut. Le gain en information induit par le choix de A comme attribut de décision est donné par la formule

$$\text{Gain } (S, A) = \text{Entropy } (S) - \sum_{v \in \text{ Values } (A)} \frac{|S_v|}{|S|} \text{ Entropy } (S_v)$$

Où Entropy(S) est l'entropie de l'ensemble de données courant, $|S_v|$ est le nombre d'exemples ayant la valeur *v* dans l'ensemble S et |S| désigne la cardinalité de l'ensemble S. L'attribut choisi est celui qui correspond au **plus grand gain informationnel**.

Exemple illustratif : on reprend l'exemple du joueur du golf. L'ensemble S initial contient 14 jours. Notons que les attributs *Temp* et *Humidity* sont continus et donc nous classons le premier dans 3 classes : hot (>= 80), mild (<=70) et cool (70 < temp < 80) et le second dans deux classes : normal (<85) et high (>=85). Nous condons aussi les valeurs de l'attribut Class en Yes ou Play et No

autrement. Nous obtenons le nouveau tableau suivant :

Outlook	Temp	Humidity	Windy	Class
Sunny	Cool	Normal	True	Y
Sunny	Hot	High	True	N
Sunny	Hot	High	False	N
Sunny	Cool	High	False	N
Sunny	Mild	Normal	False	Y
Overcast	Cool	High	True	Y
Overcast	Hot	Normal	False	Y
Overcast	Mild	Normal	True	Y
Overcast	Hot	Normal	False	Y
Rain	Cool	Normal	True	N
Rain	Mild	Normal	True	N
Rain	Cool	Normal	False	Y
Rain	Mild	Normal	False	Y
Rain	Mild	High	False	Y

Etape 1 : choix du premier attribut : on calcule le gain informationnel des choix Outlook, Temp, Humidity, Windy. Soir S l'ensemble initial. Valeurs (Outlook) : Sunny, Overcast Rain.

Gain (S,Outlook) = Entropie (S) -SOMME $_{v \text{ dans } \{sunny,\ overcast,\ rain\}}(|S_v|/|S| * \text{Entropie}(S_v))$
= Entropie (S) - (5/14)*entropie (S_{sunny}) – (4/14)*entropie($S_{overcast}$) – (5/14)*entropie(S_{rain})= 0.94 - (5/14)*0,97-0-(5/14)*0,97= **0.25.**
Par la même manière, on trouve Gain (S, Temp) = **0.04** , Gain(S, Humidity)=**0.005** et Gain(S, Windy) = **0.05** □ on choisit Outlook comme premier attribut de partitionnement. La racine de l'arbre est l'attribut Outlook avec trois valeurs : *Sunny*, *Overcast* et *Rain*. En plus, on aura les trois tableaux suivants

Temp	Humidity	Windy	Class
Cool	Normal	True	Y
Hot	High	True	N
Hot	High	False	N
Cool	High	False	N
Mild	Normal	False	Y

emp	Humidity	Windy	Class
Cool	High	True	Y
Hot	Normal	False	Y
Mild	Normal	True	Y
Hot	Normal	False	Y

Outlook *Sunny*

Outlook *Overcast*

Temp	Humidity	Windy	Class
Cool	Normal	True	N
Mild	Normal	True	N
Cool	Normal	False	Y
Mild	Normal	False	Y
Mild	High	False	Y

Outlook *Rain*

On réitère l'algorithme pour chacun des sous-ensemble obtenus.