Arbres de décision

1 Construction d'un arbre de décision

Les échantillons suivants correspondent à un ensemble de conditions météorologiques qui permettent (P pour positif) ou non (N pour négatif) la pratique du golf :

num	Outlook	Temperature	Humidity	Windy	Play
1	overcast	hot	85.0	false	P
2	overcast	cool	70.0	true	P
3	overcast	mild	85.0	true	P
4	overcast	hot	70.0	false	P
5	rainy	mild	85.0	false	P
6	rainy	cool	80.0	false	P
7	rainy	cool	70.0	true	N
8	rainy	mild	80.0	false	P
9	rainy	mild	85.0	false	P
10	sunny	hot	85.0	false	N
11	sunny	hot	85.0	true	N
12	sunny	mild	85.0	false	N
13	sunny	cool	70.0	false	P
14	sunny	mild	70.0	true	P

TABLE 1 – Exemples d'apprentissage ($\mathcal{D}_{\text{train}}$).

num	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	78	false	N
2	sunny	hot	90	true	N
3	overcast	hot	80	false	P
4	rainy	mild	96	false	P
5	rainy	cool	76	false	P
6	rainy	cool	75	false	P
7	overcast	cool	65	true	P
8	overcast	mild	83	false	N
9	sunny	cool	72	false	P
10	rainy	mild	76	false	P

TABLE 2 – Exemples de validation ($\mathcal{D}_{\text{val}}$).

Q. 1 : Sur la base de l'algorithme présenté en cours (figure 1), construisez un arbre de décision n-aire utilisant l'indice d'impureté de Gini à partir de l'échantillon d'apprentissage en prenant comme critère d'arrêt de l'expansion une impureté tolérée i_0 de 0,33.

Construire tous les arbres de décision possibles afin d'identifier le meilleur n'est pas une solution envisageable.
On cherche donc à construire intelligemment l'arbre de décision par une **induction descendante** (top-down induction of decision tree):

Procédure : **ConstruireArbre(X)** *(X est l'ensemble des exemples)*
Si tous les exemples de X appartiennent à la même classe
Alors créer une feuille portant le nom de cette classe
Sinon
  Si un critère d'arrêt est vérifié (taux minimum de bonne classification, seuil minimum de mélange, nombre minimum d'exemples...) ou s'il n'y a plus de tests de discrimination disponibles
  Alors créer une feuille portant le nom de la classe dominante dans X
  Sinon
    Déterminer le test de séparation est le plus discriminant en fonction des valeurs de chaque attribut
    Créer un nœud étiqueté par l'attribut utilisé par le test
    Pour chaque séparation X_i de X **faire** ConstruireArbre(X_i)

FIGURE 1 – Pseudo algorithme de construction d'un arbre de décision.

Q. 2 : Calculer le taux d'erreur obtenu sur l'ensemble d'apprentissage (table 1).
Q. 3 : Calculer le taux erreur obtenu sur l'ensemble de validation (table 2).
Q. 4 : Reprendre la construction précédente dans le cas où on n'utilise pas de critère d'arrêt particulier.
Q. 5 : Que constatez-vous par rapport à l'évolution des taux d'erreur obtenus sur le corpus d'apprentissage et sur le corpus de validation ?
Q. 6 : Calculez la matrice de confusion du meilleur modèle.

2 Élagage d'un arbre de décision

Reprenons le problème de l'exercice 1 et les échantillons des tableaux 1 et 2 qui correspondent à un ensemble de conditions météorologiques qui permettent (P pour positif) ou non (N pour négatif) la pratique du golf.

Q. 1 : Appliquer l'algorithme d'élagage (figure 2) du cours sur l'arbre entièrement développé de l'exercice 1 (Q. 4), lequel correspond donc à $T_{\max}$.
Q. 2 : En conséquence, quel est l'arbre « optimal » de cette famille $T_{\max}$, T_1, T_2, ... ?

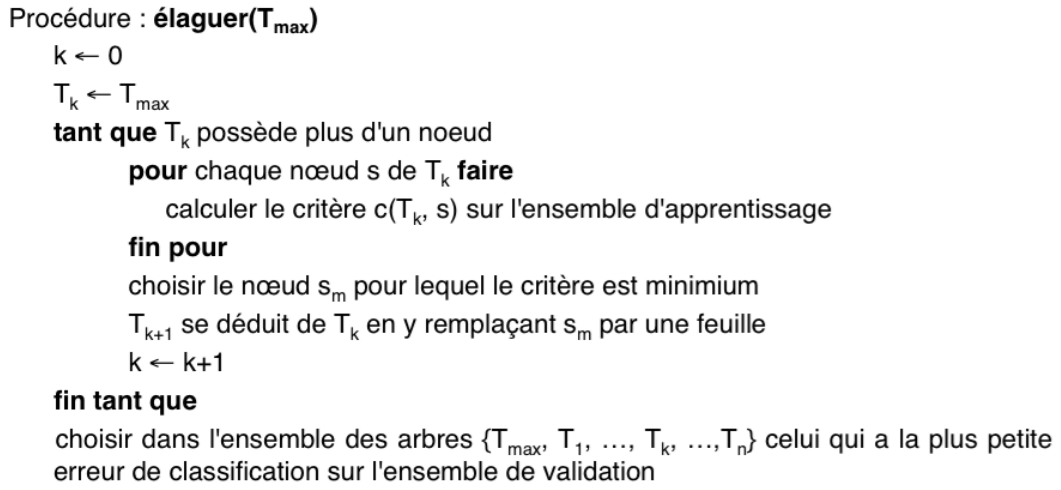

FIGURE 2 – Algorithme d'élagage d'un arbre de décision. Le critère d'évaluation est $c(T_k, s) = \frac{\mathrm{MC}(\mathrm{elag}(T_k,s))-\mathrm{MC}(T_k)}{\mathrm{NF}(T_k)(\mathrm{NF}(\mathrm{elag}(T_k,s))-1)}$ avec $\mathrm{MC}(T_k)$ le nombre d'exemples de l'ensemble d'apprentissage mal classés par T_k, $\mathrm{MC}(\mathrm{elag}(T_k, s))$ le nombre d'exemples de l'ensemble d'apprentissage mal classés si l'on transforme s en feuille, $\mathrm{NF}(T_k)$ le nombre de feuilles de T_k, et $\mathrm{NF}(\mathrm{elag}(T_k, s))$ le nombre de feuilles du sous-arbre situé sous le nœud s de T_k.