

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 pas (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

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	rain	mild	74	false	P

TABLE 2 – Exemples de test

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. 1 : Sur la base du pseudo-algorithme présenté en cours (voir 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.

Q. 2 : Calculer le taux d'erreur obtenu sur l'ensemble d'apprentissage.

Q. 3 : Calculer le taux erreur obtenu sur l'ensemble de test.

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 du taux de bonne classification sur le corpus d'apprentissage et sur le corpus de test ?

2 Elagage d'un arbre de décision

Soient les échantillons des tableaux 1 et 2.

Travail demandé

1. Appliquer l'algorithme d'élagage (voir figure 3) sur l'arbre de la figure 2. On considèrera qu'une humidité supérieure à 80 est haute et qu'en dessous elle est normale.
2. Déterminer l'arbre "optimal" de cette famille.

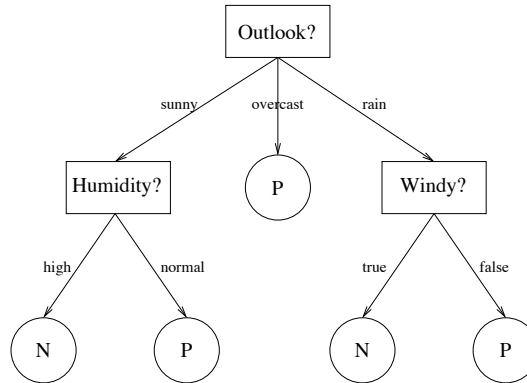


FIGURE 2 – Arbre de décision obtenu par ID3.

Procédure : **élaguer**($T_{\max}$)

$k \leftarrow 0$

$T_k \leftarrow T_{\max}$

tant que T_k possède plus d'un nœud

pour chaque nœud s de T_k **faire**

 calculer le critère $c(T_k, s)$ sur l'ensemble d'apprentissage

fin pour

 choisir le nœud s_m pour lequel le critère est minimum

T_{k+1} se déduit de T_k en y remplaçant s_m par une feuille

$k \leftarrow k+1$

fin tant que

choisir dans l'ensemble des arbres $\{T_{\max}, T_1, \dots, T_k, \dots, T_n\}$ celui qui a la plus petite erreur de classification sur l'ensemble de validation

FIGURE 3 – Pseudo algorithme d'élagage d'un arbre de décision. Le critère d'évaluation est $c(T_k, s) = \frac{MC_{elag}(T_k, s) - MC(T_k)}{n(T_k)(nt(T_k, s) - 1)}$ avec $MC_{elag}(T_k, s)$ le nombre d'exemples de l'ensemble d'apprentissage mal classés par le nœud s de T_k si on transforme s en feuille, $MC(T_k)$ le nombre d'exemples de l'ensemble d'apprentissage mal classés sous le nœud s de T_k . $n(T_k)$ le nombre de feuilles de T_k et $nt(T_k, s)$ le nombre de feuilles du sous-arbre situé sous le nœud s de T_k .