# TP 9 : Algorithme ID3

13 janvier 2025

## 1 Exemple

On considère un ensemble $E$ de 16 individus $x_i$ ayant 4 attributs binaires et une classe $y_i$ numérotée entre 0 et 2. On note $x_{i,j}$ la valeur booléenne du $j$-ème attribut pour le $i$-ème individu.

Le tableau ci-dessous donne pour chaque individu les valeurs booléennes de ses quatre attributs, ainsi que sa classe :

| Individu $x_i$ | Attribut 1 | Attribut 2 | Attribut 3 | Attribut 4 | Classe $y_i$ |
|---|---|---|---|---|---|
| $x_1$ | 0 | 0 | 1 | 0 | 0 |
| $x_2$ | 0 | 0 | 0 | 0 | 0 |
| $x_3$ | 1 | 1 | 0 | 0 | 1 |
| $x_4$ | 0 | 0 | 1 | 1 | 2 |
| $x_5$ | 0 | 1 | 1 | 0 | 1 |
| $x_6$ | 0 | 0 | 0 | 1 | 2 |
| $x_7$ | 1 | 0 | 0 | 0 | 0 |
| $x_8$ | 0 | 0 | 1 | 1 | 2 |
| $x_9$ | 0 | 0 | 1 | 1 | 2 |
| $x_{10}$ | 0 | 0 | 0 | 1 | 2 |
| $x_{11}$ | 1 | 1 | 0 | 0 | 1 |
| $x_{12}$ | 0 | 0 | 1 | 1 | 2 |
| $x_{13}$ | 0 | 1 | 1 | 0 | 1 |
| $x_{14}$ | 0 | 0 | 0 | 0 | 0 |
| $x_{15}$ | 1 | 0 | 0 | 1 | 2 |
| $x_{16}$ | 0 | 0 | 1 | 0 | 0 |

On rappelle que :

- Pour un ensemble $A$ dont les éléments de classes 0,...,$n-1$ sont répartis dans des classes $C_0$,...,$C_{n-1}$, on note $f_i$ la proportion $\frac{|C_i|}{|A|}$ d'éléments dans la classe $C_i$ parmi les éléments de $A$
- Pour un tel ensemble, on appelle **entropie** de $A$ la valeur $H(A) = \sum_{k=0}^{n-1} f_i log(f_i)$

— Pour un ensemble $E$ divisé en plusieurs ensembles $E_{i,k}$ selon un critère $k$, la probabilité qu'un élément choisi uniformément dans $E$ se trouve dans $E_{i,k}$ est notée $p_i = \frac{|E_{i,k}|}{|E|}$

— Pour un tel ensemble $E$, on note le gain $G(E,k) = H(E) - \sum_{k=0}^{m-1} p_i H(E_{i,k})$ où $m$ est le nombre de sous-ensembles de $E$.

En particulier, on considérera ici que chaque critère booléen sépare $E$ en deux sous-ensembles $E_0$ et $E_1$.

1) Pour chaque attribut $k$, calculer le gain $G(E,k) = H(E) - \sum_{k=1}^{n} p_i H(E_{i,k})$. Quel attribut $k_0$ maximise le gain ?

2) En déduire les fils $E_{0,k_0}$ et $E_{1,k_0}$ de la racine dans l'arbre construit par l'algorithme ID3.

3) Effectuer le même calcul sur les sous-noeuds jusqu'à parvenir à des feuilles correspondant à des ensembles d'éléments de même classe. Quel est l'arbre de décision obtenu par application de l'algorithme ID3 ?

4) Quelle est la classe prédite pour un nouvel élément d'attributs $(0,1,0,0)$ ?

# 2 Implémentation

## 2.1 Arbres de Décision

On cherche ici à implémenter l'algorithme ID3 en C. On commence par définir un type et des fonctions pour manipuler des arbres de décision.

On définit le type suivant :

```
struct dtree{
int classe;
int attribut;
struct dtree* faux;
struct dtree* vrai;
};
typedef struct dtree dtree;
```

Le type précédent sert à représenter les arbres de décisions, avec :

- Un champ `classe` qui vaut le numéro de la classe associée si on est dans une feuille, et $-1$ sinon
- Un champ `attribut` qui vaut le numéro d'attribut servant à la décision pour le noeud si c'est un noeud interne, et $-1$ sinon
- Des pointeurs `faux` et `vrai` vers des sous-arbres correspondant aux sous-ensembles d'éléments pour lesquels l'attribut du noeud vaut faux ou vrai respectivement, si c'est un noeud interne.

1) Définir une fonction `dtree *init_feuille(int classe)` qui prend en entrée un numéro de classe et renvoie l'arbre correspondant à une feuille associée à cette classe.

2) Définir une fonction `dtree *init_feuille(int classe)` qui prend en entrée un numéro de classe et renvoie l'arbre correspondant à une feuille associée à cette classe.

On considère le type `tableau_bool` défini de la façon suivante :

```
struct tableau_bool{
bool *tab;
int *dim;
};
typedef struct tableau_bool tableau_bool;
```

La structure contient un champ `tab` qui contient un tableau de booléens, et un champ `dim` correspondant à la taille de ce tableau. On utilise ce type pour représenter des $n$-uplets de valeurs d'attributs pour un individu dans un ensemble de données.

3) Ecrire une fonction `int classe(dtree *a, tableau_bool *t)` qui prend en entrée un pointeur vers un arbre, un pointeur vers un élément de type `tableau_bool` correspondant aux valeurs des attributs pour un individu et qui renvoie la classe prédite pour cet individu.