1 Comprendre le principe des arbres de décision et performances

Soit un problème de classification binaire sur des données décrites par trois variables (attributs $a_1 \in [1,2,3]$, $a_2 \in$ [V,F], et $a_3 \in [a,b]$, étiquetées selon deux classes $\mathcal{Y} = \{+,-\}$. A partir d'un échantillon S de données étiquetées, on décide d'apprendre un arbre de décision $A(x)$ pour en dériver une séquence de règles de prédiction dans $\mathcal{Y}$ pour toute nouvelle donnée x. On sait qu'il y a 30 exemples de données dans S, 15 de la classe + et autant de la classe −.

1. Au début de la construction de l'arbre, on doit choisir un premier test sur les variables décrivant les données : chaque test t_j est à valeurs dans le domaine de la variable a_j, $j \in [1..3]$. La figure 1 indique comment les données se répartiraient en classes dans les noeuds créés par chacun des trois tests potentiels. On note $|S_{t,v}|$ le nombre d'exemples qui passent le test t avec la valeur v. Sous chaque feuille, les nombres d'exemples positifs et négatifs de S y arrivant sont indiqués (illustration : 9 exemples positifs et un exemple négatif parviennent à la première feuille en bas à gauche).

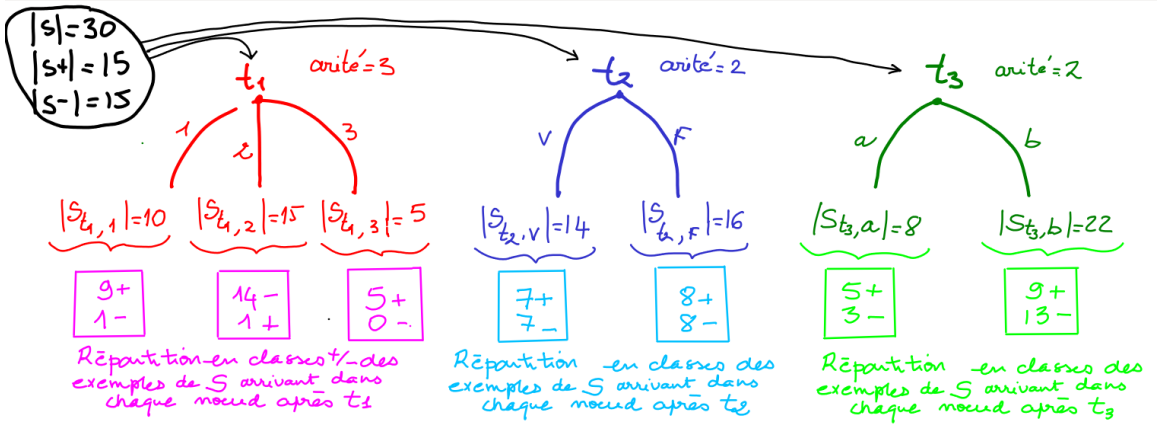

FIGURE 1 – Analyse pour le choix d'un test parmi 3

   (a) Au vu des comptages indiqués sur la figure 1, quel est le meilleur test à choisir à la racine relativement au gain en information ? (justifier rapidement la réponse – sans faire de calcul, juste des comparaisons).

   (b) L'arbre appris ne considère qu'un test à la racine (*un stump*), t_1, t_2 ou t_3 selon réponse précédente. Sur ce stump, quelle est la classe attribuée à chaque feuille ? Quelle est son erreur sur l'échantillon d'apprentissage ?

2. Soit l'échantillon de test T donné dans le tableau ci-après :

	a_1	a_2	a_3	classe
x_1	1	V	a	+
x_2	1	F	a	+
x_3	2	V	b	+
x_4	3	F	b	+
x_5	1	F	b	−
x_6	2	V	a	−
x_7	2	V	b	−
x_8	3	V	a	+

(a) Indiquer la matrice de confusion du classifieur *stump* issu du test t_1 à la racine, sur les exemples de T.

(b) Quelle est l'erreur du choix de t_1 telle que calculée sur T ?

(c) Quel est son rappel des positifs (rappel +) ?

2 Construction d'un arbre de décision

Nous disposons (figure 2) d'un échantillon de données provenant d'une enquête de satisfaction suite à des transactions immobilières opérées par une agence. Nous voulons apprendre un modèle qui permet de prédire la satisfaction d'un acquéreur suite à l'achat d'un bien immobilier via cette agence, à partir de l'emplacement et du type de ce bien, du revenu de l'acquéreur et de l'ancienneté de l'acquéreur en qualité de client de l'agence immobilière.

Emplacement	*Type de maison*	*Revenu*	*Client antérieur?*	*Résultat*
banlieue	Unifamiliale	élevé	non	Insatisfait
banlieue	Unifamiliale	élevé	oui	Insatisfait
rural	Unifamiliale	élevé	non	Satisfait
ville	Jumelée	élevé	non	Satisfait
ville	Jumelée	bas	non	Satisfait
ville	Jumelée	bas	oui	Insatisfait
rural	Jumelée	bas	oui	Satisfait
banlieue	Rangée	élevé	non	Insatisfait
banlieue	Jumelée	bas	non	Satisfait
ville	Rangée	bas	non	Satisfait
banlieue	Rangée	bas	oui	Satisfait
rural	Rangée	élevé	oui	Satisfait
rural	Unifamiliale	bas	non	Satisfait
ville	Rangée	élevé	oui	Insatisfait

FIGURE 2 – Données d'apprentissage

Nous proposons de construire un arbre de décision avec un seul test : l'arbre contient un seul noeud interne (la racine) qui est un test sur un attribut, et chaque fils est un noeud terminal (feuille). Nous appelons un tel arbre un *stump*.
Construire cet arbre en maximisant le gain calculé avec le critère de Gini