Exercice 3

On considère des images en noir et blanc codées sur 4 pixels. Chaque image est donc codée par un élément $(x_1,x_2,x_3,x_4) \in \{0,1\}^4$, où les pixels noirs sont notés 1 et les pixels blancs sont notés 0 et sont numérotés dans l'ordre

x_1	x_2
x_3	x_4

On dispose de la base d'apprentissage ci-dessous pour laquelle les images ont été réparties selon deux classes.

Classe 1 ***Classe 2***

On souhaite construire un arbre de décision sur cet échantillon avec comme variables explicatives (attributs), x_1, x_2, x_3 et x_4.

1) Ecrire la base d'apprentissage sous la forme d'un tableau

Image	X1	X2	X3	X4	Classe
1	1	0	1	1	C1
2	1	0	0	1	C1
3	0	0	0	1	C1
4	1	0	1	0	C1
5	1	0	0	1	C1
6	1	0	0	1	C1
7	0	1	1	0	C2
8	0	0	1	0	C2
9	0	1	1	0	C2
10	1	1	1	0	C2
11	0	1	1	0	C2
12	1	1	0	0	C2

Base d'apprentissage

2) Quelle variable sera choisie à la racine de l'arbre si l'on souhaite maximiser le gain à l'aide de l'indice de Gini ?

On a

$$Gini(E)=(1/2)(1-1/2)+ (1/2)(1-1/2)=1/2$$

Donc,

$$\forall i\in\{1,\dots,4\},\ G(X_i)=(1/2)-P(X_i=1)*Gini(X_i=1)- P(X_i=0)*Gini(X_i=0)$$

où

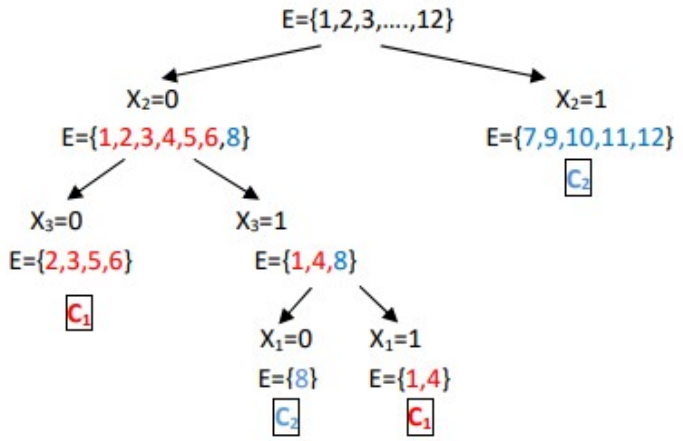
$$Gini(X_i=k)=P(Y=C_1|X_i=k)*[1-P(Y=C_1|X_i=k)]+ P(Y=C_2|X_i=k)*[1-P(Y=C_2|X_i=k)]\ ,\ k\in\{0,1\}$$

On obtient, $G(X_1)=0.129$, $G(X_2)=0.36$, $G(X_3)=0.13$, $G(X_4)=0.36$.

On commence donc l'arbre avec la variable X_2 ou la variable X_4.

3) Représenter ensuite l'arbre T avec un partage des nœuds suivant les variables dans l'ordre suivant : X_2, X_3, ›

On commence avec X_2



4) On considère l'ensemble de test suivant

Parmi tous les élagages possibles de T, quel est celui qui commet le moins d'erreur ?

Image	X1	X2	X3	X4	Classe	Prévision avec 3 nœuds (arbre entier)	Prévision avec 2 nœuds	Prévision avec 1 nœud
1	0	1	1	1	C1	C2	C2	C2
2	1	0	0	1	C1	C1	C1	C1
3	1	0	0	0	C1	C1	C1	C1
4	0	0	1	1	C1	C2	C1	C1
5	0	1	1	0	C2	C2	C2	C2
6	1	1	1	0	C2	C2	C2	C2
					Erreur	2	1	1

Base de test

L'arbre entier et l'arbre avec une seule variable explicative donne le meilleur résultat. On préférera l'arbre avec une seule variable explicative car les feuilles ont un effectif plus important. C'est embêtant d'avoir une feuille avec une seule observation.