Exercice 1 (introduction)

Soit l'échantillon suivant :

no	P1	P2	P3	Classe
1	0	V	N	A
2	1	V	I	A
3	0	F	O	B
4	1	V	N	A
5	1	V	O	B
6	1	F	N	A
7	0	F	O	B
8	0	V	I	A
9	0	F	N	B
10	1	V	I	B
11	1	F	O	A
12	1	F	I	A
13	0	V	O	B

1. Soit L l'ensemble d'apprentissage avec L = {1,...,9}. Construire l'arbre de décision t1 en choisissant les attributs dans l'ordre P3, P2, P1.

2. Même question avec t2 en utilisant l'ordre P1, P2, P3.

3. Peut-on trouver un arbre de décision correct si on ajoute 10 à L ?

4. Soit T l'ensemble de test avec T = {11,12,13}. Soit les arbres t3=A et t4=B si P1=0 A si B=1. Calculer les sorties de t1, t2, t3, t4 sur T et L.

5. Conclure.

Exercice 2 (principe de minimisation de l'entropie)

On garde l'échantillon de l'exercice 1 avec les ensembles L et T.

On veut construire un arbre de décision parfait avec le principe de minimisation l'entropie.

1. Calculer E0 entropie de l'échantillon total.

2. Calculer Eap(P1), Eap(P2), Eap(P3).

   Quel attribut choisir à la racine de l'arbre ?

3. Quels sont les noeuds terminaux à profondeur 1 ?

4. Terminer la construction de l'arbre.

5. Conclure.

Exercice 3 (minimisation de l'entropie, apprentissage et test)

1. a. Dessiner la courbe définie par la fonction E suivante:

```
E(x) = - xlog(x) - (1-x)log(1-x) pour x dans ]0,1[
E(0) = 0
E(1) = 0
```

Soit Set(1-5) l'ensemble suivant :

no	a	b	c	d	e	Classe
1	0	1	0	1	1	+
2	1	1	0	0	1	-
3	0	0	0	1	1	+
4	1	1	1	1	1	-
5	1	0	0	1	0	+

1. b. Calculer E l'entropie de Set(1-5).

Dans les questions 2, 3, 4, on veut construire un arbre de décision Tree(1-5) avec le principe de minimisation de l'entropie.

2. Calculer Eap(a), Eap(b), Eap(c), Eap(d) et Eap(e) les entropies à priori des attributs a, b, c, d, e.

3. En déduire l'attribut minimisant l'entropie et construire la racine de Tree(1-5).

4. Terminer la construction de Tree(1-5).

5. Tester Tree(1-5) sur l'ensemble de test Set(6-10) suivant:

no	a	b	c	d	e	Classe
6	1	0	1	1	1	-
7	0	1	1	0	1	+
8	0	0	1	0	0	-
9	0	1	0	1	0	+
10	1	0	1	0	1	-

Combien Tree(1-5) fait-il d'erreurs ?

6. Sans utiliser a et b, construire un arbre de décision Decide(1-5) sur Set(1-5).

7. Tester Decide(1-5) sur Set(6-10).

   Combien Decide(1-5) fait-il d'erreurs ?

8. Conclure.

Exercice 4 (relativiser le principe de minimisation de l'entropie)

Soit l'ensemble d'apprentissage L = {1,...,5} avec :

no	x	y	z	Classe
1	0	1	0	+
2	1	0	0	-
3	0	1	1	-
4	0	0	1	+
5	1	0	1	-

1. Construire l'arbre de décision parfait t1 suivant le principe de minimisation de l'entropie.

2. Construire un arbre t2 de profondeur 2 utilisant l'attribut 'y' à la racine de l'arbre.

3. Comparer t1 et t2 et conclure.

Exercice 5 (forêt d'arbres de décision)

Soit l'ensemble suivant:

no	x	y	z	Classe
0	0	0	0	+
1	0	0	1	-
2	0	1	0	+
3	0	1	1	+
4	1	0	0	-
5	1	0	1	-
6	1	1	0	-
7	1	1	1	+

1. Soit ADi (1<=i<=5) l'arbre de décision construit sur l'ensemble d'apprentissage Ei en suivant le principe de minimisation de l'entropie avec:

   E1 = {0, 1, 2, 3} E2 = {4, 5, 6, 7} E3 = {0, 2, 4, 6}
   E4 = {1, 3, 5, 7} E5 = {0, 1, 4, 5}

   Construire AD1, AD2, AD3, AD4, AD5.

2. Tester les ADi sur l'ensemble total.

3. On appelle "Foret" un ensemble d'arbres de décision décidant de la classe d'un exemple par un vote majoritaire issu de ses arbres.

   Soit F5 = { ADi | 1<=i<=5 } et F3 = { AD1, AD3, AD5 }

   Tester F5 et F3 sur l'ensemble total.

4. Conclure.

Exercice 6 (données avec valeurs manquantes)

On considère des objets avec les attributs ‘forme’, ‘taille’ et ‘couleur’ prenant respectivement les valeurs ‘rond’ et ‘carré’, ‘petit’ et ‘grand’, ‘bleu’, ‘blanc’ et ‘rouge’. L'attribut ‘classe’ vaut ‘+’ ou ‘-’.

Les valeurs d’attribut manquantes sont représentées avec un ‘?’.

no	forme	taille	couleur	classe
1	rond	petit	bleu	+
2	carré	grand	rouge	-
3	rond	?	blanc	+
4	carré	petit	bleu	+
5	rond	grand	bleu	+
6	carré	grand	blanc	-
7	carré	?	blanc	+
8	carré	grand	bleu	-
9	carré	petit	rouge	+
10	rond	grand	blanc	+

1. Valeur majoritaire de l'attribut:

On remplace les valeurs manquantes par la valeur majoritaire prise par cet attribut sur l'échantillon complet.

Quelle valeur associe-t-on sur notre échantillon ?

Peut-on trouver un arbre de décision parfait ?

2. Valeur majoritaire de l'attribut par classe:

On remplace la valeur manquante d’un attribut d’un objet O par la valeur majoritaire prise par l'attribut pour les objets de la même classe que celle de O.

Quelle valeur associe-t-on sur notre échantillon ?

Peut-on trouver un arbre de décision parfait ?

Quel arbre obtient-on en appliquant l'algorithme basé sur l'entropie ?

3. Méthode de Quinlan:

Il s’agit de la méthode, vue en cours, consistant à attribuer une valeur probabiliste à un attribut avec valeur manquante. Ces probabilités sont estimées avec les fréquences des valeurs de cet attribut pour l'échantillon considéré.

Par exemple, la probabilité que l'attribut taille ait la valeur ‘petit’ est de 3/8 car il y a 3 exemples sur 8 avec la valeur ‘petit’.

Quel arbre obtient-on en appliquant l'algorithme basé sur l'entropie ?