Examen – Apprentissage artificiel

1h. **Aucun document autorisé.**

Prenez soin de lire tous les exercices avant de commencer. La notation est donnée à titre indicatif.

Exercice 1 : (45 mn) 15 pts

RID	age	income	student	credit	C_i: buy
1	youth	high	no	fair	C_2: no
2	youth	high	no	excellent	C_2: no
3	middle-aged	high	no	fair	C_1: yes
4	senior	medium	no	fair	C_1: yes
5	senior	low	yes	fair	C_1: yes
6	senior	low	yes	excellent	C_2: no
7	middle-aged	low	yes	excellent	C_1: yes
8	youth	medium	no	fair	C_2: no
9	youth	low	yes	fair	C_1: yes
10	senior	medium	yes	fair	C_1: yes
11	youth	medium	yes	excellent	C_1: yes
12	middle-aged	medium	no	excellent	C_1: yes
13	middle-aged	high	yes	fair	C_1: yes
14	senior	medium	no	excellent	C_2: no

Soit le fichier ci-dessus possédant une variable de classe BUY, et découpé en 2 sous-ensembles: D_1 et D_2. **D_1 contient les 10 premiers objets**, et D_2 contient les **4 derniers (11 à 14)**.
D_1 sera *l'ensemble d'apprentissage* et D_2 sera *l'ensemble de validation*.

1- Combien y a-t-il d'attributs pertinents permettant de caractériser la classe BUY ; 0,5 pt
2- Pourquoi le classifieur Naïf Bayes est dit Naïf ? 1 pt
3- Construire M_1, le **classifieur bayésien naïf** à partir de D_1. 5 pts
4- D_2 va être utilisé pour tester le modèle M_1. Déterminer la classe des 4 objets de D_2; 2 pts
5- Donner la matrice de confusion de M_1 sur D_2 ; 0,5 pt
6- A partir de cette matrice de confusion, et après avoir **rappelé la formule de calcul** :
   a. Calculer le taux d'erreur apparente de M_1; 1 pt
   b. Calculer la précision de la classe C_1='yes'; 1 pt
   c. Calculer le rappel de la classe C_2='no' ; 1 pt
   d. Calculer la sensibilité de la classe C_1='yes'; 1 pt
   e. Calculer la spécificité de la classe C_2='no' ; 1 pt
   f. Calculer le taux de faux positifs de la classe C_1='yes'; 1 pt

Exercice 2 : (15 mn) 5 pts

Etant donné l'ensemble d'apprentissage suivant :

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	–	T	F
4	+	F	F
5	–	F	T
6	–	F	T

a. Quelle est l'indice de Gini de cette collection d'exemples d'entraînement par rapport à l'attribut classe ? (2 pts)
b. Quel est le gain d'information de a_2 par rapport à ces exemples d'entraînement (considérer l'indice de Gini lors du calcul du gain d'information) ? (3 pts)

ANNEXES

La ***précision*** pour une classe donnée mesure le taux d'exemples corrects parmi les exemples prédits dans cette classe.
Le ***rappel*** mesure le taux d'exemples corrects parmi les exemples de la classe.
Le taux de ***faux positifs*** d'une classe mesure le nombre d'objets positifs parmi ceux n'appartenant pas à la classe.
Le taux de ***vrais positifs*** d'une classe mesure le nombre d'objets positifs parmi les vrais objets de la classe.
Le taux de ***faux négatifs*** d'une classe mesure le nombre d'objets négatifs parmi ceux appartenant à la classe.
Le taux de ***vrais négatifs*** d'une classe mesure le nombre d'objets négatifs parmi ceux n'appartenant pas à la classe.
La ***sensibilité*** est la probabilité qu'un test soit positif si l'objet appartient à la classe.
La ***spécificité*** est la probabilité qu'un test soit négatif si l'objet n'appartient pas à la classe.

Arbres de décision

p(j / t) est la fréquence relative de la classe j au nœud t.

Indice de Gini pour le nœud t : $GINI(t) = 1 - \sum_j [p(j \mid t)]^2$

Indice de Gini pour l'attribut *split :* $GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$

Gain d'information avec l'indice de Gini pour l'attribut *split :* $Gain_{split} = Gini(r) - Gini_{split}$

Le nœud parent **r** a n objets, et est divisé en k partitions. La partition i possède n_i objets.

Entropie du nœud t: $Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$

Gain d'information avec l'entropie pour l'attribut *split*: $GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$

Le nœud parent **p** a n objets, et est partitionné en k partitions. La partition i possède n_i objets.

Indice d'Erreur en classification au nœud t : $Error(t) = 1 - \max_i P(i \mid t)$

Gain d'information avec l'indice d'erreur en classification : $Gain_{split} = Error(r) - Error_{split}$

Le nœud parent **r** a n objets, et est partitionné en k partitions. La partition i possède n_i objets.

Bayésien naïf : Estimation des probabilités conditionnelles

Ai : une valeur de l'attribut A

Nic : Nombre d'objets ayant la valeur Ai dans la classe c

Nc : Nombre d'objets de la classe c

k : nombre de valeurs de l'attribut A

p : probabilité apriori

$$\text{Original: } P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + k}$$

$$\text{m-estimate: } P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$