Examen – Apprentissage

1h30. **Aucun document autorisé.**

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

Exercice : (20 pts)

On dispose du fichier ci-dessous possédant une variable de classe CONTACT-LENSES indiquant le type de lentilles de contact d'une personne. On découpe l'ensemble en 2 : D_1 et D_2. **D_1 contient les 15 premiers objets**, et D_2 contient les **9 derniers**.

1- A quoi correspond le type de fichier ci-dessous ? Quel logiciel l'utilise ? (1 pt)

2- Expliquer ce qu'est l'apprentissage artificiel en vous appuyant sur ce fichier (1 pt)

3- On souhaite construire le modèle M_1 **d'arbre de décision** en utilisant *l'indice d'erreur en classification* (voir annexe)
   a. Construire l'arbre de décision M_1 sur l'ensemble d'apprentissage D_1 ; (5 pts)
   b. Donner sa matrice de confusion sur D_2 ; (2 pts)

4- Construire le modèle **bayésien naïf** M_2 en utilisant D_1 et en appliquant la formule de Laplace (voir annexe) (5 pts)

5- Donner la matrice de confusion sur D_2; (2 pts)

6- Comparer les 2 modèles M_1 et M_2 ; (2 pts)

7- Si l'on souhaite appliquer la technique des réseaux de neurones multicouches, comment doit-on procéder ? (Expliquer, sans chercher à construire un modèle) (2 pts)

```
@relation contact-lenses

@attribute age                  {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip   {myope, hypermetrope}
@attribute astigmatism          {no, yes}
@attribute tear-prod-rate       {reduced, normal}
@attribute contact-lenses       {soft, hard, none}

@data
% 24 instances
%
young,          myope,          no,       reduced,   none
young,          myope,          no,       normal,    soft
young,          myope,          yes,      reduced,   none
young,          hypermetrope,   yes,      reduced,   none
young,          hypermetrope,   yes,      normal,    hard
pre-presbyopic, myope,          no,       reduced,   none
pre-presbyopic, myope,          no,       normal,    soft
pre-presbyopic, myope,          yes,      normal,    hard
pre-presbyopic, hypermetrope,   yes,      reduced,   none
pre-presbyopic, hypermetrope,   yes,      normal,    none
presbyopic,     myope,          no,       reduced,   none
presbyopic,     myope,          no,       normal,    none
presbyopic,     myope,          yes,      normal,    hard
presbyopic,     hypermetrope,   no,       normal,    soft
presbyopic,     hypermetrope,   yes,      reduced,   none

young,          myope,          yes,      normal,    hard
young,          hypermetrope,   no,       reduced,   none
young,          hypermetrope,   no,       normal,    soft
pre-presbyopic, myope,          yes,      reduced,   none
pre-presbyopic, hypermetrope,   no,       reduced,   none
pre-presbyopic, hypermetrope,   no,       normal,    soft
presbyopic,     myope,          yes,      reduced,   none
presbyopic,     hypermetrope,   no,       reduced,   none
presbyopic,     hypermetrope,   yes,      normal,    none
```

ANNEXES

A_i : une valeur de l'attribut A

N_{ic} : Nombre d'objets ayant la valeur A_i dans la classe c

N_c : Nombre d'objets de la classe c

k : nombre de valeurs de l'attribut A

p : probabilité apriori

m : paramètre

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

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

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

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.

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.