

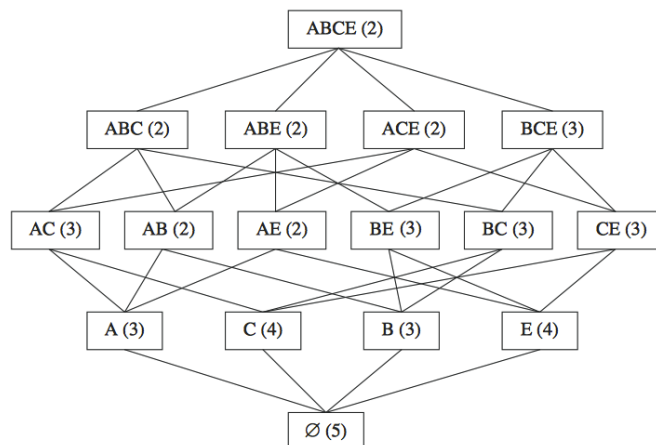
## Examen

### Questions de cours (6 pts : 3 + 3)

1. Compléter le tableau suivant par Vrai ou Faux pour spécifier pour chaque méthode de data mining le type de données qu'elle supporte :

| Méthode               | KPPV | CBA | AD | RB | RNA | SVM |
|-----------------------|------|-----|----|----|-----|-----|
| Données supportées    |      |     |    |    |     |     |
| Classe numérique      |      |     |    |    |     |     |
| Classe symbolique     |      |     |    |    |     |     |
| Sans classe           |      |     |    |    |     |     |
| Attributs numériques  |      |     |    |    |     |     |
| Attributs symboliques |      |     |    |    |     |     |

2. Soit le treillis suivant représentant les fréquences des différents motifs d'un ensemble de données.



Sachant que le support minimum est de 3 exemples, remplir le tableaux suivant par Vrai, Faux.

| Type de motif | Fréquent | Fermé | Fréquent Maximal | Rare | Rare Minimal |
|---------------|----------|-------|------------------|------|--------------|
| B             |          |       |                  |      |              |
| E             |          |       |                  |      |              |
| AB            |          |       |                  |      |              |
| CE            |          |       |                  |      |              |
| AC            |          |       |                  |      |              |
| BCE           |          |       |                  |      |              |
| ABC           |          |       |                  |      |              |

### Exercices (14 pts : 3.5 + 1.5 + 1)

Le département d'informatique désire construire un modèle de décision lui permettant de sélectionner parmi les nouveaux bacheliers ceux susceptibles de réussir une formation en informatique. Pour cela,

on a rassemblé un ensemble de données contenant des informations concernant les anciens étudiants du départements. L'ensemble est composé de quatre attributs représentant un résumé de la formation secondaire de l'étudiant à savoir l'option du Bac (Lettre, Science ou Math), la moyenne du Bac (numérique), Répétition du Bac (Oui ou Non) et le résultat de la formation de l'étudiant au département (Réussi ou Échec). Le tableau suivant présente la base collectée pour l'entraînement :

| N° | Opt | Moy | Rép | Résultats | N° | Opt | Moy | Rép | Résultats |
|----|-----|-----|-----|-----------|----|-----|-----|-----|-----------|
| 1  | L   | 13  | N   | R         | 6  | M   | 10  | N   | R         |
| 2  | L   | 13  | O   | E         | 7  | M   | 14  | O   | R         |
| 3  | L   | 10  | N   | E         | 8  | S   | 12  | N   | E         |
| 4  | M   | 10  | O   | E         | 9  | S   | 15  | N   | R         |
| 5  | M   | 13  | N   | R         | 10 | S   | 15  | O   | E         |

Le modèle construit va être utilisé pour déterminer si les étudiants ayant les résultats suivants vont réussir une formation en informatique ou non :

| N° | Opt | Moy | Rép |
|----|-----|-----|-----|
| 1  | S   | 14  | N   |
| 2  | M   | 12  | N   |
| 3  | M   | 13  | N   |

**Exercice 1 (4.5 pts : 1.5 + 3)**

1. Construire un modèle de décision bayésien naïf en utilisant l'estimateur de Laplace et une distribution normale pour les attributs numériques ( $P(X = x) = \frac{1}{\sqrt{2\pi\rho^2}} e^{-\frac{(x-\bar{x})^2}{2\rho^2}}$ ).
2. Donner les classes des étudiants de la deuxième base.

**Exercice 2 (1.5 pt : 1 + 0.5)**

En supposant la discrétisation suivante pour les attributs numériques :

- Note  $\leq 12$  : F (Faible)
- $12 < \text{Note} \leq 14$  : M (Moyenne)
- Note  $> 14$  : E (Élevée),

Donner les nouvelles bases discrétisées.

**Exercice 3 (4 pts : 3 + 1)**

1. En utilisant la base d'entraînement discrétisée dans l'exercice précédent et la méthode de GiniIndex, construire un arbre de décision permettant de prédire le résultat d'un nouveau bachelier.
2. Utiliser l'arbre construit pour prédire les résultats des étudiants de la deuxième base.

**Exercice 4 (4 pts : 2.5 + 1.5)**

En utilisant la méthode CBA de classification associative avec un support minimum de 40% et une confiance minimale de 50% :

1. Construire un modèle de décision à partir de la base d'entraînement discrétisée dans l'exercice 2.
2. Utiliser ce modèle pour prédire les résultats des étudiants de la deuxième table.

*Bonne Chance*

*Dr A.Djeffal*

## Corrigé type

### Questions de cours (6 pts : 3 + 3)

#### 1. Classes

| Méthode<br>Données supportées | KPPV | CBA | AD | RB | RNA | SVM |
|-------------------------------|------|-----|----|----|-----|-----|
| Classe numérique              | F    | F   | F  | F  | V   | V   |
| Classe symbolique             | V    | V   | V  | V  | V   | V   |
| Sans classe                   | F    | F   | F  | F  | V   | V   |
| Attributs numériques          | V    | F   | F  | V  | V   | V   |
| Attributs symboliques         | F    | V   | V  | V  | F   | V   |

**3 pts**

#### 2. Motifs

| Type de motif<br>Motif | Fréquent | Fermé | Fréquent Maximal | Rare | Rare Minimal |
|------------------------|----------|-------|------------------|------|--------------|
| B                      | V        | F     | F                | F    | F            |
| E                      | V        | V     | F                | F    | F            |
| AB                     | F        | F     | F                | V    | V            |
| CE                     | V        | F     | F                | F    | F            |
| AC                     | V        | V     | V                | F    | F            |
| BCE                    | V        | V     | V                | F    | F            |
| ABC                    | F        | F     | F                | V    | V            |

**3 pts**

### Exercice 1 (4.5 pts : 1.5 + 3)

#### 1. Modèle de décision bayésien naïf :

| Attribut  | Classe   |          |                |
|-----------|----------|----------|----------------|
|           | R (5/10) | E (5/10) |                |
| Opt       |          |          | <b>0.25 pt</b> |
| L         | 1/5      | 2/5      | <b>0.25 pt</b> |
| S         | 1/5      | 2/5      |                |
| M         | 3/5      | 1/5      |                |
| Moy       |          |          | <b>0.75 pt</b> |
| $\bar{X}$ | 13       | 12       |                |
| $\rho$    | 1.67     | 1.89     |                |
| Rép       |          |          | <b>0.25 pt</b> |
| O         | 1/5      | 3/5      |                |
| N         | 4/5      | 2/5      |                |

#### 2. Classes des exemples de la deuxième base :

–  $X1 = \langle S, 14, N \rangle$

$$- P(R/X1) = P(R) \times P(Opt = S/R) \times P(Moy = 14/R) \times P(Rep = N/R)$$

$$= \frac{5}{10} \times \frac{1}{5} \times \frac{1}{\sqrt{2 \times 3.14 \times 1.67^2}} e^{-\frac{(14-13)^2}{2 \times 1.67^2}} \times \frac{4}{5}$$

$$= 0.5 \times 0.2 \times 0.19 \times 0.8 = \mathbf{0.015}$$

$$- P(E/X1) = P(E) \times P(Opt = S/E) \times P(Moy = 14/E) \times P(Rep = N/E)$$

$$= \frac{5}{10} \times \frac{2}{5} \times \frac{1}{\sqrt{2 \times 3.14 \times 1.89^2}} e^{-\frac{(14-12)^2}{2 \times 1.89^2}} \times \frac{2}{5}$$

$$= 0.5 \times 0.4 \times 0.12 \times 0.4 = \mathbf{0.009}$$

- $\Rightarrow$  **Résultat X1 = R** 1 pt
- X2 = <M, 12, N>
- $P(R/X2) = P(R) \times P(Opt = M/R) \times P(Moy = 12/R) \times P(Rep = N/R)$ 

$$= \frac{5}{10} \times \frac{3}{5} \times \frac{1}{\sqrt{2 \times 3.14 \times 1.67^2}} e^{-\frac{(12-13)^2}{2 \times 1.67^2}} \times \frac{4}{5}$$

$$= 0.5 \times 0.6 \times 0.19 \times 0.8 = \mathbf{0.04}$$
- $P(E/X2) = P(E) \times P(Opt = M/E) \times P(Moy = 12/E) \times P(Rep = N/E)$ 

$$= \frac{5}{10} \times \frac{1}{5} \times \frac{1}{\sqrt{2 \times 3.14 \times 1.89^2}} e^{-\frac{(12-12)^2}{2 \times 1.89^2}} \times \frac{2}{5}$$

$$= 0.5 \times 0.2 \times 0.21 \times 0.4 = \mathbf{0.008}$$
- $\Rightarrow$  **Résultat X2 = R** 1 pt
- X3 = <M, 13, N>
- $P(R/X2) = P(R) \times P(Opt = M/R) \times P(Moy = 13/R) \times P(Rep = N/R)$ 

$$= \frac{5}{10} \times \frac{3}{5} \times \frac{1}{\sqrt{2 \times 3.14 \times 1.67^2}} e^{-\frac{(13-13)^2}{2 \times 1.67^2}} \times \frac{4}{5}$$

$$= 0.5 \times 0.6 \times 0.23 \times 0.8 = \mathbf{0.05}$$
- $P(E/X2) = P(E) \times P(Opt = M/E) \times P(Moy = 13/E) \times P(Rep = N/E)$ 

$$= \frac{5}{10} \times \frac{1}{5} \times \frac{1}{\sqrt{2 \times 3.14 \times 1.89^2}} e^{-\frac{(13-12)^2}{2 \times 1.89^2}} \times \frac{2}{5}$$

$$= 0.5 \times 0.2 \times 0.18 \times 0.4 = \mathbf{0.007}$$
- $\Rightarrow$  **Résultat X3 = R** 1 pt

**Exercice 2 (1.5 pt : 1 + 0.5)** Bases discrétisées :

Base d'entraînement :

1 pt

| N° | Opt | Moy | Rép | Résultats | N° | Opt | Moy | Rép | Résultat |
|----|-----|-----|-----|-----------|----|-----|-----|-----|----------|
| 1  | L   | M   | N   | R         | 6  | M   | F   | N   | R        |
| 2  | L   | M   | O   | E         | 7  | M   | M   | O   | R        |
| 3  | L   | F   | N   | E         | 8  | S   | F   | N   | E        |
| 4  | M   | F   | O   | E         | 9  | S   | E   | N   | R        |
| 5  | M   | M   | N   | R         | 10 | S   | E   | O   | E        |

Base 2 :

0.5 pt

| N° | Opt | Moy | Rép |
|----|-----|-----|-----|
| 1  | S   | M   | N   |
| 2  | M   | F   | N   |
| 3  | M   | M   | N   |

**Exercice 3 (4 pts : 3 + 1)**

1. -  $Gini_{split}(Opt) = \frac{3}{10}Gini(Opt = L) + \frac{4}{10}Gini(Opt = M) + \frac{3}{10}Gini(Opt = S)$ 

$$= 0.3 \times [1 - (\frac{1}{3})^2 - (\frac{2}{3})^2] + 0.4 \times [1 - (\frac{1}{4})^2 - (\frac{3}{4})^2] + 0.3 \times [1 - (\frac{1}{3})^2 - (\frac{2}{3})^2]$$

$$= 0.13 + 0.15 + 0.13 = \mathbf{0.41}$$
 0.25 pt
- $Gini_{split}(Moy) = \frac{2}{10}Gini(Moy = E) + \frac{4}{10}Gini(Moy = F) + \frac{4}{10}Gini(Moy = M)$ 

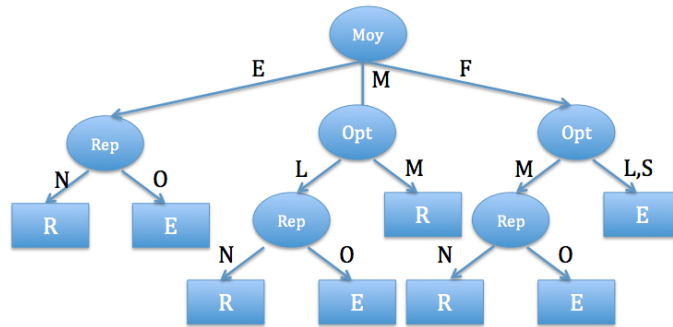
$$= 0.2 \times [1 - (\frac{1}{2})^2 - (\frac{1}{2})^2] + 0.4 \times [1 - (\frac{1}{4})^2 - (\frac{3}{4})^2] + 0.4 \times [1 - (\frac{1}{4})^2 - (\frac{3}{4})^2]$$

$$= 0.1 + 0.15 + 0.15 = \mathbf{0.4}$$
 0.25 pt
- $Gini_{split}(Rep) = \frac{6}{10}Gini(Rep = N) + \frac{4}{10}Gini(Rep = O)$ 

$$= 0.6 \times [1 - (\frac{4}{6})^2 - (\frac{2}{6})^2] + 0.4 \times [1 - (\frac{1}{4})^2 - (\frac{3}{4})^2]$$

$$= 0.26 + 0.15 = \mathbf{0.41}$$
 0.25 pt
- Le premier attribut de segmentation est **Moy** 0.25 pt
- Moy = E
- Rep = O  $\Rightarrow$  Résultat = E

- Rep = N  $\Rightarrow$  Résultat = R
- Moy = F
  - Opt = L ou S  $\Rightarrow$  Résultat = E
  - Opt = M
    - Rep = O  $\Rightarrow$  Résultat = E
    - Rep = N  $\Rightarrow$  Résultat = R
- Moy = M
  - Opt = M  $\Rightarrow$  Résultat = R
  - Opt = L
    - Rep = O  $\Rightarrow$  Résultat = E
    - Rep = N  $\Rightarrow$  Résultat = R
- L'arbre est donc :



3 pts

2. Classes des exemples de la base 2 :

- X1 = <S, 14, N>  $\Rightarrow$  Résultat = R
- X2 = <M, 12, N>  $\Rightarrow$  Résultat = R
- X3 = <M, 13, N>  $\Rightarrow$  Résultat = R

1 pt

**Exercice 4 (4 pts : 2.5 + 1.5)**

– Table formelle

|    | OL | OM | OS | MF | MM | ME | RO | RN | RR | RE |
|----|----|----|----|----|----|----|----|----|----|----|
| 01 | 1  | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 1  | 0  |
| 02 | 1  | 0  | 0  | 0  | 1  | 0  | 1  | 0  | 0  | 1  |
| 03 | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 1  | 0  | 1  |
| 04 | 0  | 1  | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  |
| 05 | 0  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 1  | 0  |
| 06 | 0  | 1  | 0  | 1  | 0  | 0  | 0  | 1  | 1  | 0  |
| 07 | 0  | 1  | 0  | 0  | 1  | 0  | 1  | 0  | 1  | 0  |
| 08 | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 1  | 0  | 1  |
| 09 | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 1  | 1  | 0  |
| 10 | 0  | 0  | 1  | 0  | 0  | 1  | 1  | 0  | 0  | 1  |

**0.5 pt**

– Motifs fréquents : Support minimum = 4/10

\*  $C_1 = \{OL, OM, OS, MF, MM, ME, RO, RN, RR, RE\}$

\* Support =  $\{3/10, 4/10, 3/10, 4/10, 4/10, 2/10, 4/10, 6/10, 5/10, 5/10\}$

\*  $F_1 = \{OM, MF, MM, RO, RN, RR, RE\}$

\* Pour  $C_2$ , On ne prend que les motifs contenant des attributs classe :

\*  $C_2 = \{OMRR, MFRR, MMRR, RORR, RNRR, OMRE, MFRE, MMRE, RORE, RNRE\}$

\* Support =  $\{3/10, 1/10, 3/10, 1/10, 4/10, 1/10, 3/10, 1/10, 3/10, 2/10\}$

\*  $F_2 = \{RNRR\}$

\*  $F = \{OM, MF, MM, RO, RN, RR, RE, RNRR\}$

**1 pt**

– Règles solides :

\* Une seule règle candidate :  $RN \Rightarrow RR$

\* Confiance =  $\frac{4/10}{6/10} = 0.66 \Rightarrow$  règle solide

**0.5 pt**

– Modèle de décision

\* Si Rép = N alors Résultat = R

\* Sinon Résultat = E

**0.5 pt**

– Classes des exemples de la base 2 :

–  $X1 = \langle S, 14, N \rangle \Rightarrow$  Résultat = R

–  $X2 = \langle M, 12, N \rangle \Rightarrow$  Résultat = R

–  $X3 = \langle M, 13, N \rangle \Rightarrow$  Résultat = R

**1.5 pt**