Université Mohamed Khider-Biskra
Faculté des sciences exactes et des sciences de la nature et de la vie
Département d'informatique

Master SIOD — Fouille de Données Avancée
13 Jan 2019 — 10:00-11:30, S5,S6

Examen

Questions de cours (6 pts : 2 + 2 + 2)

1. Comparer les deux méthode de classification, réseaux de neurones et arbres de décision. Quels sont les avantages et les inconvénients de l'une par rapport à l'autre ?
2. Définir la capacité de généralisation d'un modèle, et dire comment peut-on l'améliorer ?
3. Que représentent les vecteurs supports dans une SVM ? Dans quels phase sont découverts : l'entrainement, le test ou l'utilisation ?

Exercices (14 pts : 5 + 4 + 5)

On désir dans cet exercice analyser l'utilisation des téléphones mobiles par un ensemble de personnes en se basant sur leur revenu, age, niveau d'étude et leur situation familiale. On dispose de l'ensemble de données d'entrainement suivant :

Num	Revenu	Age	Niveau Etud	Situation Fam	Utilisation
1	Bas	Vieux	Universitaire	Marié	Basse
2	Moyen	Jeune	Moyen	Célibataire	Moyenne
3	Bas	Vieux	Universitaire	Marié	Basse
4	Elevé	Jeune	Universitaire	Célibataire	Elevée
5	Bas	Vieux	Universitaire	Marié	Basse
6	Elevé	Jeune	Moyen	Célibataire	Moyenne
7	Moyen	Jeune	Moyen	Marié	Moyenne
8	Moyen	Vieux	Lycéen	Célibataire	Basse
9	Elevé	Vieux	Universitaire	Célibataire	Elevée
10	Bas	Vieux	Lycéen	Marié	Basse
11	Moyen	Jeune	Moyen	Marié	Moyenne
12	Moyen	Vieux	Lycéen	Célibataire	Basse
13	Elevé	Vieux	Universitaire	Célibatire	Elevée
14	Bas	Vieux	Lycéen	Marié	Basse
15	Moyen	Jeune	Moyen	Marié	Moyenne

On dispose également de la base de test suivante :

Num	Revenu	Age	Niveau Etud	Situation Fam	Utilisation
1	Moyen	Vieux	Universitaire	Marié	Basse
2	Moyen	Jeune	Moyen	Célibataire	Moyenne
3	Bas	Vieux	Lycéen	Marié	Basse
4	Elevé	Jeune	Universitaire	Célibataire	Elevée
5	Elevé	Vieux	Moyen	Marié	Moyenne

Exercice 1 (5 pts : 4 + 0.5 + 0.5)

1. Construire un modèle de décision à partir de la base d'entrainement en utilisant la méthode CBA avec un support de fréquence minimum de 30% et une confidence de 80%.
2. Donner son taux de reconnaissance sur la base de test.
3. Donner sa matrice de confusion.

Exercice 2 (4 pts : 2.5 + 0.5 + 1)

1. Construire un modèle de décision à partir de la base d'entrainement en utilisant la méthode OneR.
2. Donner son taux de reconnaissance sur la base de test.
3. Donner sa moyenne harmonique.

Exercice 3 (5 pt : 4 + 0.5 + 0.5)

1. Construire un arbre de décision à partir de la base d'entrainement en utilisant le Gini Index.
2. Donner son taux de reconnaissance sur la base de test.
3. Donner sa matrice de confusion.

Bonne Chance

Dr A.Djeffal

Questions de cours (6 pts : 2 + 2 + 2)

1. Les réseaux de neurones est une méthode connexionniste par contre les AD est à base de règles,
   Les avantages des RN : Numériques, plus précis, régression, sensible à l'ordre des exemples, développé en Deep learning
   Les avantages des AD : Lisibles, catégoriels, transformés facilement en règles
2. La capacité de généralisation de système est sa capacité à prédire les exemples non vus lors de l'entrainement et l'extension des conclusion tirées des exemples d'entrainement pour couvrir les nouveaux exemples.
   Elle peut être amélioré en validant le modèle entrainées sur des données autres que ceux utilisés pour son entrainement
3. Les vecteurs supports dans une SVM représentent les exemples les plus ressemblants les deux classes. Ce sont les exemples qui se situent sur les frontières de la marge. Ils sont découverts lors de l'entrainement et utilisés seuls lors de la prédiction.

Exercice 1 (5 pts : 4 + 0.5 + 0.5)

1. Modèle de décision CBA (support minimum = 30%, confidence min = 80%)
   (a) **Motifs fréquents (1.5 pt) :**
   - Support min = 0.3 ← Frequence min = 15 x 0.3 = 4.5 ≈ 5
   - C_1

$C_1 =$	RB	RM	RE	AV	AJ	NU	NM	NL	SM	SC	UB	UM	UE
Support	**5**	**6**	4	**9**	**6**	**6**	**5**	4	**8**	**7**	**7**	**5**	3

   - $F_1 = \{RB, RM, AV, AJ, NU, NM, SM, SC, UB, UM\}$
   - C_1

C_2	RBUB	RBUM	RMUB	RMUM	AVUB	AVUM	AJUB	AJUM
Support	**5**	0	2	**4**	**7**	0	0	**5**
C_2	NUUB	NUUM	NMUB	NMUM	SMUB	SMUM	SCUB	SCUM
Support	3	0	0	**5**	**5**	3	2	3

   - $F_2 = \{RBUB, AVUB, AJUM, NMUM, SMUB\}$
   - C_3

C_3	RBAVUB	RBSMUB	AVSMUB	AJNMUM
Support	**5**	**5**	**5**	**5**

   - $F_3 = \{RBAVUB, RBSMUB, AVSMUB, AJNMUM\}$
   - C_4

C_4	RBAVSMUB
Support	**5**

   - $C_5 = \phi$

   (b) **Règles (1.5 pt) :**
   - à partir de F_2
     - R1 : RB ⇒ UB , conf $=\frac{5}{5} = 1$: Solide
     - R2 : AV ⇒ UB , conf $=\frac{7}{9} = 0.77$: non Solide
     - R3 : AJ ⇒ UM , conf $=\frac{5}{6} = 0.83$: Solide
     - R4 : NM ⇒ UM conf $=\frac{5}{5} = 1$: Solide
     - R5 : SM ⇒ UB , conf $=\frac{5}{8} = 0.62$: non Solide

- à partir de F_3
  - R1 : RBAV ⇒ UB, conf $=\frac{5}{5}=1$: Solide
  - R2 : RBSM ⇒ UB, conf $=\frac{5}{5}=1$: Solide
  - R3 : AVSM ⇒ UB, conf $=\frac{5}{5}=1$: Solide
  - R4 : AJNM ⇒ UM, conf $=\frac{5}{5}=1$: Solide
- à partir de F_4
  - R1 : RBAVSM ⇒ UB, conf $=\frac{5}{5}=1$: Solide

(c) **Règles solides triées par odre de confidence et fréquence (0.25 pt) :**

- R1 : RB ⇒ UB
- R2 : NM ⇒ UM
- R3 : RBAV ⇒ UB
- R4 : RBSM ⇒ UB
- R5 : AVSM ⇒ UB
- R6 : AJNM ⇒ UM
- R7 : RBAVSM ⇒ UB
- R8 : AJ ⇒ UM

(d) **Règles raffinées (0.25 pt) :**

- R1 : RB ⇒ UB
- R2 : NM ⇒ UM
- R3 : RBAV ⇒ UB **X (R1)**
- R4 : RBSM ⇒ UB **X (R1)**
- R5 : AVSM ⇒ UB
- R6 : AJNM ⇒ UM X (R2)
- R7 : RBAVSM ⇒ UB **X (R1)**
- R8 : AJ ⇒ UM

(e) **Modèles (0.5 pt) :**

- R1 : RB ⇒ UB
- R2 : NM ⇒ UM
- R3 : AVSM ⇒ UB
- R4 : AJ ⇒ UM
- R5 : Sinon UB

2. **Taux de reconnaissance sur la base de test (0.5 pt).**

Num	**Revenu**	**Age**	**Niveau Etud**	**Situation Fam**	**Utilisation**	Prédiction du modèle
1	Moyen	Vieux	Universitaire	Marié	Basse	Basse (R3)
2	Moyen	Jeune	Moyen	Célibataire	Moyenne	Moyenne (R2)
3	Bas	Vieux	Lycéen	Marié	Basse	Basse (R1)
4	Elevé	Jeune	Universitaire	Célibataire	Elevée	Moyenne (R4)
5	Elevé	Vieux	Moyen	Marié	Moyenne	Moyenne (R2)

Taux de reconnaissance $=\frac{4}{5}=80\%$

3. **Matrice de confusion (0.5 pt) :**

Observations \ Prédictions	Basse	Moyenne	Elevée
Basse	2	0	0
Moyenne	1	1	0
Elevée	1	0	0

Exercice 2 (4 pts : 2.5 + 0.5 + 1)

1. Modèle OneR.

   (a) **Attribut : Revenu (Erreur = 3)** **(0.5 pt)**

Attribut \ Classe	Basse	Moyenne	Elevée
Bas	5	0	0
Moyen	**2**	4	0
Elevé	0	**1**	3

   (b) **Attribut : Age (Erreur = 3)** **(0.5 pt)**

Attribut \ Classe	Basse	Moyenne	Elevée
Jeune	0	5	**1**
Vieux	7	0	**2**

   (c) **Attribut : Niveau d'étude (Erreur = 3)** **(0.5 pt)**

Attribut \ Classe	Basse	Moyenne	Elevée
Universitaire	3	0	**3**
Moyen	0	5	0
Lycéen	4	0	0

   (d) **Attribut : Situation familiale (Erreur = 7)** **(0.5 pt)**

Attribut \ Classe	Basse	Moyenne	Elevée
Marié	5	**3**	0
Célibataire	**2**	**2**	3

   (e) **Si on choisi l'attibut Revenu le modèle sera :** **(0.5 pt)**

   - Si Revenu= Basse alors Utilisation = Basse
   - Si Revenu= Moyen alors Utilisation = Moyenne
   - Si Revenu= Elevé alors Utilisation = Elevée

2. **Taux de reconnaissance sur la base de test**

Num	**Revenu**	**Age**	**Niveau Etud**	**Situation Fam**	**Utilisation**	Prédiction du modèle
1	Moyen	Vieux	Universitaire	Marié	Basse	Moyenne
2	Moyen	Jeune	Moyen	Célibataire	Moyenne	Moyenne
3	Bas	Vieux	Lycéen	Marié	Basse	Basse
4	Elevé	Jeune	Universitaire	Célibataire	Elevée	Elevée
5	Elevé	Vieux	Moyen	Marié	Moyenne	Elevéé

   Taux de reconnaissance = $\frac{3}{5} = 60\%$ **(0.5 pt)**

3. Moyenne harmonique

   - Classe "Basse"

- CP = 1 ; FP= 0 ; CN = 3 ; FN = 1
- $Sv_1 = \frac{CP}{CP+FN} = \frac{1}{2} = 0.5$
- $Sp_1 = \frac{CN}{CN+FP} = \frac{3}{3+0} = 1$

– Classe "Moyenne"
- CP = 1 ; FP= 1 ; CN = 3 ; FN = 0
- $Sv_2 = \frac{CP}{CP+FN} = \frac{1}{1+0} = 1$
- $Sp_2 = \frac{CN}{CN+FP} = \frac{3}{3+1} = 0.75$

– Classe "Elevée"
- CP = 1 ; FP= 1 ; CN = 2 ; FN = 1
- $Sv_3 = \frac{CP}{CP+FN} = \frac{1}{1+1} = 0.5$
- $Sp_3 = \frac{CN}{CN+FP} = \frac{2}{2+1} = 0.66$

(0.5 pt)

- $Sv = \sum Sv_i = 0.5 + 1 + 0.5 = 2$
- $Sp = \sum Sp_i = 1 + 0.75 + 0.66 = 2.41$ **(0.25 pt)**
- Moyenne harmonique = $\frac{2 \times Sv \times Sp}{Sv+Sp} = \frac{2 \times 2 \times 2.41}{2+2.41} = \frac{9.64}{4.41} = 2.18$ **(0.25 pt)**

Exercice 3 (5 pt : 4 + 0.5 + 0.5)

1. Construire un arbre de décision à partir de la base d'entrainement en utilisant le Gini Index.

   (a) Gini(Revenu) = $\frac{5}{15}$Gini(Bas) + $\frac{6}{15}$Gini(Moyen) + $\frac{4}{15}$Gini(Elevé)
   $= 0.33 \times [0] + 0.4 \times [1 - (\frac{4}{6})^2 - (\frac{2}{6})^2] + 0.26 \times [1 - (\frac{3}{4})^2 - (\frac{1}{4})^2]$
   $= 0 + 0.4 \times [0.45] + 0.26 \times [0.37] = 0.18 + 0.09 = 0.27$

   (0.5 pt)

   (b) Gini(Age) = $\frac{6}{15}$Gini(Jeune) + $\frac{9}{15}$ Gini(View)
   $= 0.4 \times [1 - (\frac{5}{6})^2 - (\frac{1}{6})^2] + 0.6 \times [1 - (\frac{7}{9})^2 - (\frac{2}{9})^2]$
   $= 0.4 \times [0.3] + 0.6 \times [0.36] = 0.12 + 0.21 = 0.33$

   (0.5 pt)

   (c) Gini(Niveau d'étude) = $\frac{6}{15}$Gini(Universitaire) + $\frac{5}{15}$Gini(Moyen) + $\frac{4}{15}$Gini(Lycéen)
   $= 0.4 \times [1 - (\frac{3}{6})^2 - (\frac{3}{6})^2] + 0.33 \times [1 - (\frac{5}{5})^2] + 0.26 \times [1 - (\frac{4}{4})^2]$
   $= 0.4 \times [0.5] + 0.33 \times [0] + 0.26 \times [0] = 0.2$

   (0.5 pt)

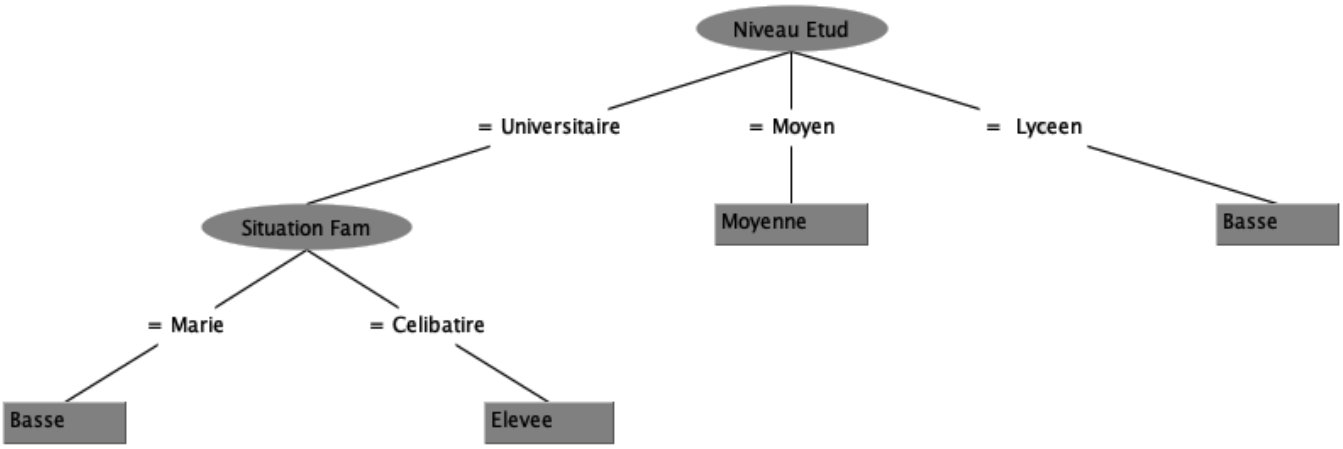
   (d) Gini(Situation familiale) = $\frac{7}{15}$Gini(Célibataire) + $\frac{8}{15}$Gini(Marié)
   $= 0.46 \times [1 - (\frac{3}{7})^2 - (\frac{2}{7})^2 - (\frac{2}{7})^2] + 0.53 \times [1 - (\frac{5}{8})^2 - (\frac{3}{8})^2]$
   $= 0.46 \times [0.65] + 0.53 \times [0.53] = 0.22 + 0.28 = 0.55$

   (0.5 pt)

   (e) L'attribut de Gini min est le "**Niveau d'étude**"
   - Niveau d'étude = Lycéen ⇒ Utilisation = Basse
   - Niveau d'étude = Moyen ⇒ Utilisation = Moyenne
   - Niveau d'étude = Universitaire
     i. Situation familiale = Marié ⇒ Utilisation = Basse
     ii. Situation Célibataire = ⇒ Utilisation = Elevé

   (1 pt)

(f) Arbre de décision (1 pt)



2. Taux de reconnaissance sur la base de test.

Num	**Revenu**	**Age**	**Niveau Etud**	**Situation Fam**	**Utilisation**	Prédiction du modèle
1	Moyen	Vieux	Universitaire	Marié	Basse	Basse
2	Moyen	Jeune	Moyen	Célibataire	Moyenne	Moyenne
3	Bas	Vieux	Lycéen	Marié	Basse	Basse
4	Elevé	Jeune	Universitaire	Célibataire	Elevée	Elevée
5	Elevé	Vieux	Moyen	Marié	Moyenne	Moyenne

Taux de reconnaissance = $\frac{5}{5} = 100\%$ **(0.5 pt)**

3. **Matrice de confusion (0.5 pt) :**

Prédictions / Observations	Basse	Moyenne	Elevée
Basse	2	0	0
Moyenne		2	0
Elevée	0	0	1