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

Master IDM
4 Fév 2015

Fouille de Données Avancée
16:00-17:30, Salles S1, S2

Examen

Questions de cours (4 pts : 2 + 1 + 1)

1. Dans la classification supervisée (l'apprentissage), l'évaluation de la qualité du modèle construit peut être effectuée en calculant son taux de reconnaissance sur les données d'entrainement elles-mêmes ou sur des données écartées dès le départ des données d'entrainement appelées données de test. A votre avis, comment peut-ont évaluer la qualité d'un clustering ?
2. A quoi sert la méthode de Bagging.
3. Justifier l'utilisation des noyaux dans l'apprentissage par machines à vecteurs supports.

Exercice 1 Motifs fréquents (8 pts : 2 + 2 + 2 + 2)

Supposons qu'il exite 6 articles numérotés de 1 à 6 $(a_1, ..., a_6)$, et 12 paniers numérotés de 1 à 12 $(p_1, ..., p_{12})$. L'article a_i est ans le panier p_j si et seulement si i divise j, par conséquent, a_1 est dans tous les paniers et a_2 est dans les paniers de numéro pair, et ainsi de suite. Avec un support minimum de 25% et une confidence minimale de 90%, donner :

1. La base de données formelle.
2. L'ensemble des motifs fréquents.
3. Les motifs fréquents fermés et les motifs fréquents maximaux.
4. Les règles solides.

Exercice 2 Classification **(8 pts : 4 + 2 + 2)**

Nous considérerons l'ensemble d'exemples représentant la nature de différents échantillons de champignons : toxique ou non selon les critères de couleur, taille, forme et le milieu de croissance :

Coleur	**Taille**	**Forme**	**Milieu**	**Toxique**
marron	petit	plat	terre	oui
jaune	petit	sphère	terre	oui
marron	moyen	conique	bois	non
blanc	moyen	sphère	terre	non
blanc	grand	plat	terre	non

1. Construire l'arbre de décision correspondant à cet ensemble en utilisant l'algorithme ID3.

Tournez la page ...

2. Donner la précision de l'arbre et sa moyenne harmonique sur la table de test suivante :

Coleur	**Taille**	**Forme**	**Milieu**	**Toxique**
marron	grand	plat	bois	non
blanc	moyen	conique	terre	oui
jaune	moyen	sphère	terre	oui
marron	moyen	conique	bois	non
jaune	petit	plat	terre	oui

3. Dites si le champignon blanc sphérique qui pousse sur le bois est toxique ou non, en utilisant la classification bayésienne naïve.

★★★ Bonne chance ★★★

Dr A.Djeffal

Corrigé type

Questions de cours (4 pts)

1. En utilisant les distances intra et inter-clusters permettant de mesurer respectivement le rapprochement des exemples de chaque cluster et l'éloignement des clusters les uns des autres. 2 pts
2. La méthode de Bagging se base sur le Bootstrap. Elle subdivise l'ensemble D d'exemples en n sous-ensembles. À partir de chaque sous-ensemble D_i, on apprend un modèle M_i en utilisant la méthode Bootstrap. L'ensemble de ces modèles forme un modèle composé M_*. Pour classiffier un nouvel exemple, il est exposé à chaque modèle M_i pour obtenir une classe c_{M_i}. Chaque décision est considérée comme un vote. La classe de décision est prise comme la classe la plus votée. 1 pt
3. Les noyaux sont utilisés dans les SVMs pour trouver un espace où les données sont linéairement séparables. 1 pt

Motifs fréquents (8 pts : 2 + 2 + 2 + 2)

1. La base formelle

	a_1	a_2	a_3	a_4	a_5	a_6
p_1	1	0	0	0	0	0
p_2	1	1	0	0	0	0
p_3	1	0	1	0	0	0
p_4	1	1	0	1	0	0
p_5	1	0	0	0	1	0
p_6	1	1	1	0	0	1
p_7	1	0	0	0	0	0
p_8	1	1	0	1	0	0
p_9	1	0	1	0	0	0
p_10	1	1	0	0	1	0
p_11	1	0	0	0	0	0
p_12	1	1	1	1	0	1

2 pts

2. Motifs fréquents = $\{F_1 \cup F_2 \cup F_3\}$
   - $F_1 = \{a_1, a_2, a_3, a_4\}$
   - $F_2 = \{a_1a_2, a_1a_3, a_1a_4, a_2a_4\}$
   - $F_3 = \{a_1a_2a_4\}$

   2 pts
3. – Motifs fréquents fermés = $\{a_1, a_1a_2, a_1a_3, a_1a_2a_4\}$ 1 pt
   – Motifs fréquents maximaux = $\{a_1a_3, a_1a_2a_4\}$ 1 pt
4. Les règles solides :
   (a) $a_2 \Rightarrow a_1$
   (b) $a_3 \Rightarrow a_1$
   (c) $a_4 \Rightarrow a_1$

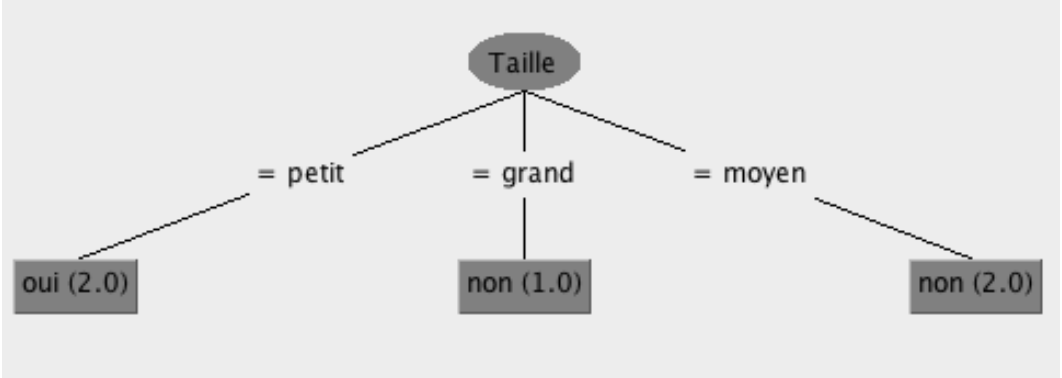
(d) $a_4 \Rightarrow a_2$

(e) $a_2 a_4 \Rightarrow a_1$

(f) $a_1 a_4 \Rightarrow a_2$

Classification (8 pts : 4 + 2 + 2)

1. Arbre construit par l'algorithme ID3



4 pts

2. – Précision $= \frac{3}{5} = 60\%$ 1 pt
   – Moyenne harmonique :

$$\begin{cases} Sv = & \frac{CP}{CP+FN} & = \frac{1}{1+2} = \frac{1}{3} \\ Sp = & \frac{CN}{CN+FP} & = \frac{2}{2+0} = 1 \end{cases}$$

$$Moyenne\ harmonique = \frac{2 \times Sv \times Sp}{Sv + Sp} = \frac{1/3 \times 1}{1/3 + 1} = 0.25$$

1 pt

3. Classification bayésienne
   – $P(Toxique = oui) = \frac{2}{5} = 0.4$
     – $P(Couleur = blac/Toxique = oui) = \frac{0}{2}$
     – $P(Taille = \forall/Toxique = oui) = 1$
     – $P(Forme = Sphere/Toxique = oui) = \frac{1}{2}$
     – $P(Milieu = bois/Toxique = oui) = \frac{0}{2}$
     – **On utilise l'estimateur de Laplace : ajouter 1 aux numérateurs et le nombre de valeurs distincts de l'attribut aux dénominateurs :**
     – $P(Couleur = blac/Toxique = oui) = \frac{0+1}{2+3} = \frac{1}{5}$
     – $P(Forme = Sphere/Toxique = oui) = \frac{1}{2}$
     – $P(Milieu = bois/Toxique = oui) = \frac{0+1}{2+2} = \frac{1}{4}$
     – $P(Toxique = oui/Couleur = blanc \wedge Forme = Sphere \wedge Milieu = bois) = \frac{1}{5} \times \frac{1}{2} \times \frac{1}{4} \times 0.4 = 0.01$
   – $P(Toxique = non) = \frac{3}{5} = 0.6$
     – $P(Couleur = blac/Toxique = non) = \frac{2}{3}$
     – $P(Taille = \forall/Toxique = oui) = 1$
     – $P(Forme = Sphere/Toxique = non) = \frac{1}{3}$

  - $P(Milieu = bois/Toxique = non) = \frac{1}{3}$
  - $P(Toxique = non/Couleur = blanc \wedge Forme = Sphere \wedge Milieu = bois) = \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times 0.6 = 0.13$
- **Donc le champignon blanc sphérique qui pousse sur le bois n'est pas toxique.**

2 pts