Université de Mons

Datawarehousing and datamining

Travaux pratiques avec Weka

Auteur :
Maxime De Wolf

13 mars 2018

Table des matières

1 Weka : Tutoriel

1.1 Questions 17.1.9 et 17.1.10

Ces questions portent sur l'arbre de décision crée à partir du fichier *iris.arff*. Voici donc l'arbre de décision obtenu :

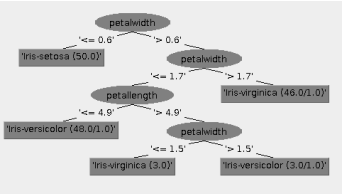

FIGURE 1 – Arbre de décision du *dataset iris.arff*

Question 17.1.9

Cette question consiste à évaluer la qualité de cet arbre (Figure 1) grâce à différentes options de tests. Ici, on effectuera ces tests une première fois avec le *dataset* complet et la 2^e fois avec la technique *10-fold cross-validation*. Nous comparons ensuite les résultats obtenus sur base des 2 *confusion matrix* :

(a) *Dataset* complet

a	b	c	
50	0	0	a = Iris-setosa
0	49	1	b = Iris-versicolor
0	2	48	c = Iris-virginica

(b) *10-fold cross-validation*

a	b	c	
49	1	0	a = Iris-setosa
0	47	3	b = Iris-versicolor
0	2	48	c = Iris-virginica

TABLE 1 – *Confusion matrix* obtenues grâce à deux méthodes de test différentes

Nous remarquons que le test sur le *dataset* complet classifie correctement 98% des instances tandis que ce chiffre descend à 96% avec le test *10-fold cross-validation*. Tester le modèle avec le *dataset* complet est une mauvaise idée car il donne une estimation optimiste de la qualité du modèle. En revanche, *10-fold cross-validation* permet de se faire une bonne idée de la généralisation du modèle et offre donc une meilleure mesure de qualité.

Question 17.1.10

En observant la localisation de ces erreurs, nous remarquons que certaines instances de classe *Iris-Verginica* ont des valeurs d'attributs équivalentes à celles d'instance de classe *Iris-Versicolor*. Le modèle n'a donc aucune chance de les différencier si nous voulons éviter l'*overfitting*. D'autre part, nous remarquons que l'instance de classe *Iris-Setosa* qui a été mal identifier aurait dû être correctement classé selon l'arbre de décision final obtenu.

1.2 Questions 17.2.4 à 17.2.11

Question 17.2.4

Le but de cette question est d'étudier la précision du classificateur *5-nearest neighbor* en fonction des attributs utilisés lors de cette classification. Ici, nous exécutons cette algorithme sur le *dataset glass.arff* et nous le test grâce à la technique *10-fold cross-validation*. Les résultats ainsi obtenus sont résumés dans la table suivante :

TABLE 2 – Précision obtenue en utilisant *IBk* pour différents sous-ensemble d'attributs

Nombre d'attributs	Attribut retiré	Précision de la classification
9	$\emptyset$	67.757
8	Si	71.4953
7	Fe	73.3645
6		
5		
4		
3		
2		
1		
0		

2 CoIL Challenge 2000