

TP Classification et Arbres de décision

Objectif TP: Mise en œuvre de méthodes de classification, Application de C4.5 et ID3 sous Weka, interprétation des résultats.

Jeux de données : nous travaillerons sur les bases de données zoo et titanic

Exercice 1 : KNN

Supposons que l'on a un problème de classification qui consiste à déterminer la classe d'appartenance de nouvelles instances X_i . Le domaine de valeurs des classes possibles est : $\{1, 2, 3\}$.

Selon la base de connaissance suivante, déterminez à la main (ou sous excel) la classe de l'instance X6, dont les valeurs pour les attributs A1 à A5 (numériques) sont $\langle 3, 12, 4, 7, 8 \rangle$, à l'aide de l'algorithme des k-voisins les plus proches (K-NN). Montrez tous les calculs.

Instances	A1	A2	A3	A4	A5	Classe
X1	3	5	4	6	1	1
X2	4	6	10	3	2	2
X3	8	3	4	2	6	3
X4	2	1	4	3	6	3
X5	2	5	1	4	8	2

Préparation du TP et lecture de résultats sur Iris

Rappel

La matrice de confusion

Exemple sous Weka

```
=== Confusion Matrix ===
```

```
  a  b  c  <-- classified as
15  0  0 |  a = Iris-setosa
 0 19  0 |  b = Iris-versicolor
 0  2 15 |  c = Iris-virginica
```

Ici en ligne les classes d'affectation et en colonne les classe a priori

```
=== Evaluation on test split ===
=== Summary ===
```

```
Correctly Classified Instances      49           96.0784 %
Incorrectly Classified Instances     2            3.9216 %
Kappa statistic                     0.9408
Mean absolute error                  0.0396
Root mean squared error              0.1579
Relative absolute error              8.8979 %
Root relative squared error          33.4091 %
Total Number of Instances           51
```

Dans cette partie, il est important de noter :

Le nombre de bien classé : 49 sur 51 soit 96.0784%

Le nombre de mal classé : 2 sur 51 soit 3.9216%

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Iris-setosa
1	0.063	0.905	1	0.95	0.969	Iris-versicolor
0.882	0	1	0.882	0.938	0.967	Iris-virginica

		Classe prédite	
		Oui	Non
Classe réelle	Oui	TP (vrai positif)	FP (faux positif)
	Non	FN (faux négatif)	TN (vrai négatif)

TP rate : taux des « vrais positifs » 15/17 (=0.882)

FP rate : taux des « faux positifs » 0/34 (=0)

Precision : $P = TP / (TP + FP)$ ici 15/15 (=1)

Rappel : « recall » : $R = TP / (TP + FN)$ ici 15/17 (=0.882)

La F-measure proposée par (Van Rijsbergen, 1979) combine les mesures de précision et de rappel

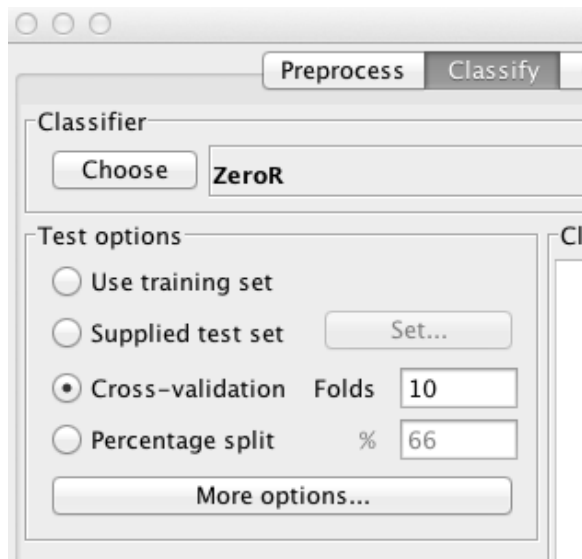
$F_{measure} = 2 * P * r / (P + R)$

Ici

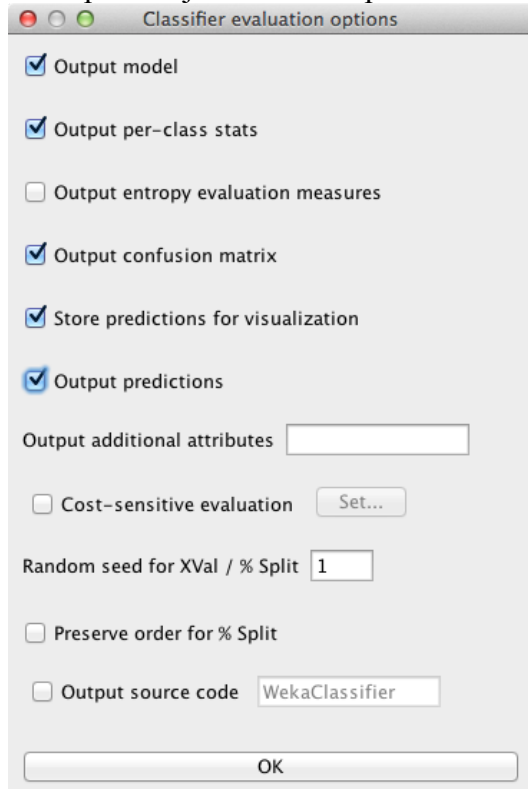
F-Measure : $2 * 1 * 0.882 / (1 + 0.882) = 0.938$

Les paramètres de Weka concernant le découpage des données

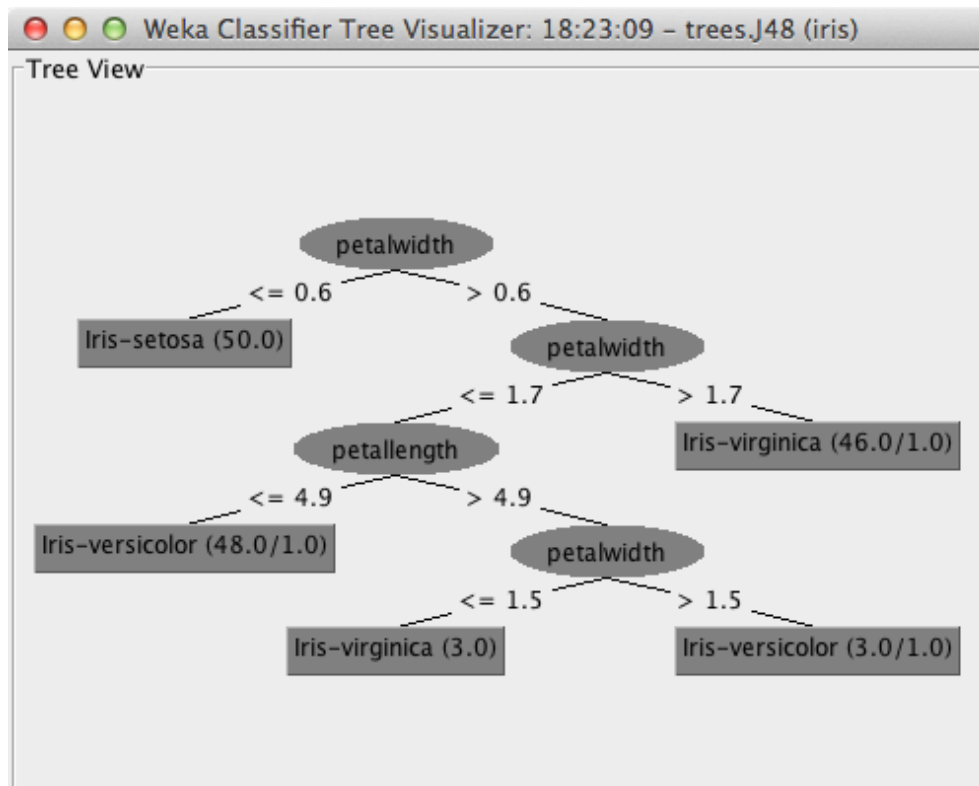
- Situation idéale : grand ensemble de données d'entraînement et ensemble de données de test distinct (arrive peu souvent)
- Découpage des données en deux sous-ensembles (« Percentage split »)
 - Ensemble d'entraînement (e.g. 66%)
 - Ensemble de test (e.g. 33%)
- Validation croisée en n strates (« Cross-validation »)
 - Partitionnement en n sous-ensembles
 - n - 1 sous-ensembles utilisé en entraînement et 1 pour le test
 - Processus répété n fois (un par partitionnement)



Vous pouvez jouer sur les options en cliquant sur « More options... »



Un click droit dans la fenêtre de Result vous permet de visualiser l'arbre



Exercice 2: le Zoo

L'objectif est de décider de la classe d'un animal(mammifère, poisson, oiseau, invertébré, insectes, amphibien, reptile) à partir des caractéristiques suivantes : présence de poils, plumes, ponte d'oeuf, production de lait, capacité de voler, capacité de nager, prédateur, présence de dents, de colonne vertébrale, respiration à l'air, venimosité, présence de palmes ou de nageoires, nombre de pattes, queue, domesticable ou encore sa taille.

Q2.1 : Tentez de créer, à la main et intuitivement, un arbre de décision classant les animaux suivants selon leurs caractéristiques ci-avant présentées : antilope, ours, poisson chat, poule, crabe, abeille

Q2.2 : A partir d'un sous-échantillon des données contenues dans le fichier zoo.csv (les 15 premières lignes), appliquez votre arbre de décision pour classer ces animaux. Quel est votre taux d'erreur ?

Q2.3 Transformez le fichier zoo.csv en un fichier arff en vous aidant de la documentation <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>. Importez alors le fichier dans Weka et exécutez l'algorithme J48 (version WEKA de C4.5) dessus. Modifiez les paramètres d'exécution. Certains attributs posent ils problème ? Pourquoi parle-t-on de classification supervisée ?

Q2.4 : Expérimentez différents algorithmes d'apprentissage, lesquels offrent la meilleure précision ? le meilleur rappel ? Pour quelles valeurs de paramètres ?

Q2.5 : Essayez rapidement de trouver des règles d'association entre les attributs ...

Exercice 3 : Comparaison d'algorithmes de classification sur le jeu de données Titanic

Le base de données présente quatre attributs pour chacun des 2201 passagers du titanic.

- CLASS avec comme valeurs: 1st, 2nd, 3rd, crew (La classe dans laquelle voyageait le passager ou s'il s'agissait d'un membre d'équipage.)
- AGE avec comme valeur : adult, child
- SEX avec comme valeurs : female, male
- SURVIVED avec comme valeurs: no, yes (Est-ce que le passager a survécu?)

La question est de trouver un lien entre la classe, l'âge, le sexe et le fait d'avoir survécu ou non au naufrage du Titanic.

Interprétation des résultats

Q3.1 : Lancez ID3 sous Weka, quelles performances a l'algorithme ? (vous commenterez une sortie de weka).

Q3.2 : Quelle méthode de validation avez vous choisie? Comparer avec d'autres méthodes de validation.

Q3.3 : Quelle particularité a le jeu de données Titanic (ordre des instances) ? Cela a-t-il de l'influence? Vérifiez en les options dans « More Options... »

Q3.4 : Visualisation d'arbres obtenus. Les arbres sont ils différents ?

Q3.5 : Utilisez d'autres algorithmes de classification et comparez les résultats

L'objectif est de remplir le tableau suivant

	Titanic	
	Correctly Classified Instances	F-mesure
Algorithme 1 et paramètres		
Algorithme 2 et paramètres		
Algorithme 3 et paramètres		
....		

Q3.7 : Commentez les résultats (lisibilité, performance ...)