Rapport du TP de weka
Romain RINCÉ

1 Exercice 1

1.1 Question 1

Chaque instance est un mail qui sera déterminé comme étant un spam ou non. Il y a 4601 instances. Pour déterminer si une instance est un spam, on étudie la fréquence d'un ensemble de mots, divers calculs sur la longueur des chaines de caractères n'étant que des lettres capitales et la fréquence de certains caractères.

1.2 Question 2

Reorder réorganise l'ordre d'études des attributs. Standardize va "décaler" les valeurs numériques pour que leurs moyennes soient à 0.

1.3 Question 3

Voici les résultats obtenus pour les différents algorithmes :

- OneR 78.0917% Correct, 477 mails classés spams, 531 spams passent, temps 0.17 secondes
- NaiveBayes 79.2871% Correct, 865 mails classés spams, 88 spams passent, 0.12 secondes
- J48 92.9798% Correct, 156 mails classés spams, 167 spams passent, temps 0.92 secondes
- RandomForest 94.8272% Correct, 78 mails classés spams, 160 spam passent, temps 1.16 secondes
- MultilayerPerceptron 91.4366% Correct, 192 mails classés spams, 202 spams passent, temps 95.83 secondes
- SMO 90.4151% Correct, 134 mails classés spams, 307 spams passent, temps 0.63 secondes

RandomForest est relativement rapide et est le plus efficace sur les vrais négatifs (C'est à dire les mails classés comme spams).

2 Exercice 2

2.1 Question 1

Voir script.py

2.2 Question 3

On peut voir que la matrice de coût n'influence pas l'algorithme appliqué. Cependant elle offre une information supplémentaire qui correspond au coût de la classification (Un coût total et un cout moyen). Ainsi on peut voir les résultats suivants :

- ZeroR Coût total = 1500 et Coût moyen = 1.5
- OneR Coût total = 1284 et Coût moyen = 1.284
- NaiveBayes Coût total = 850 et Coût moyen = 0.850
- J48 Coût total = 1027 et Coût moyen = 1.027
- RandomForest Coût total = 967 et Coût moyen = 0.967
- SMO Coût total = 885 et Coût moyen = 0.885

On a donc une information supplémentaire pour évaluer la performance de l'algorithme pour un contexte donné.