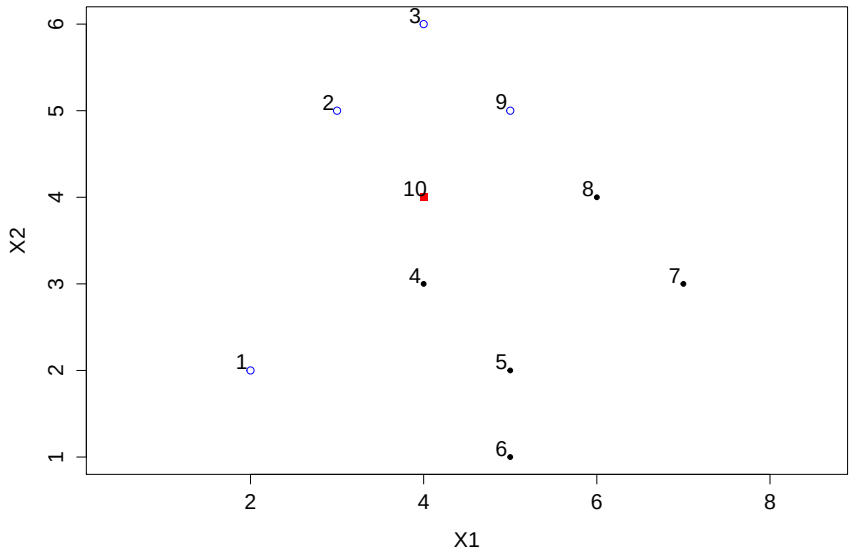
Université Paris Descartes
UFR de Mathématiques et Informatique

Classification
Examen du 6/1/2016 - Correction

Exercice 1

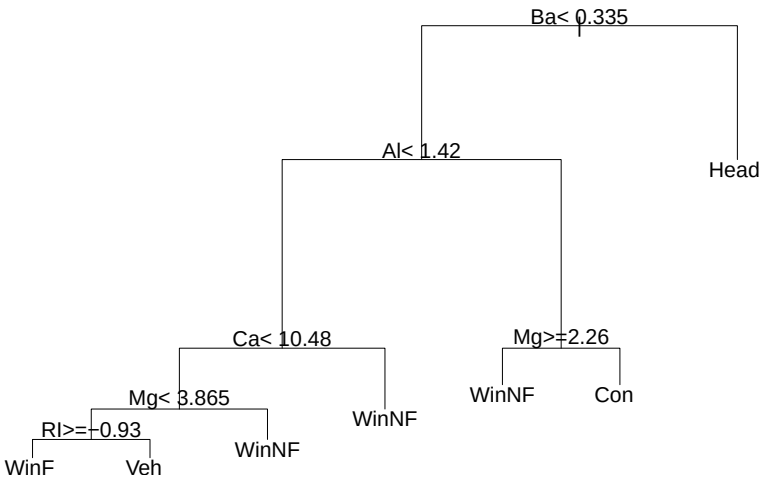
1. .



2. Le plus proche voisin de X_{10} pour la distance euclidienne est X_4 (distance 1), suivi de X_2 et X_9 (distance $\sqrt{2}$), suivis de X_3 et X_8 (distance2), suivis de X_5 (distance $\sqrt{5}$), suivis de X_1 (distance $\sqrt{8}$), suivis de X_6 et X_7 (distance $\sqrt{10}$). Par conséquent,
   - le classement au plus proche voisin de X_{10} est $\hat{Y}_{10} = 1$ (car $Y_4 = 1$),
   - le classement aux 3 plus proches voisins de X_{10} est $\hat{Y}_{10} = 0$ (2 voisins à $Y = 0$ contre 1 seul à $Y = 1$),
   - le classement aux 5 plus proches voisins de X_{10} est $\hat{Y}_{10} = 0$ (3 voisins à $Y = 0$ contre 2 à $Y = 1$),
   - le classement aux 7 plus proches voisins de X_{10} est $\hat{Y}_{10} = 0$ (4 voisins à $Y = 0$ contre 3 à $Y = 1$),
   - le classement aux 9 plus proches voisins de X_{10} est $\hat{Y}_{10} = 1$ (4 voisins à $Y = 0$ contre 5 à $Y = 1$).

Exercice 2

4. .



5. Le premier individu a les données suivantes :

```
    RI    Na   Mg  Al    Si    K   Ca Ba Fe type
1 3.01 13.64 4.49 1.1 71.78 0.06 8.75  0  0 WinF
```

   D'après l'arbre représenté ci-dessus, l'individu va être classé ainsi :
   – $Ba = 0 < 0.335$ donc on va dans la branche de gauche,
   – $Al = 1.1 < 1.42$ donc on descend dans la sous-branche de gauche,
   – $Ca = 8.75 < 10.48$ donc on descend dans la sous-branche de gauche,
   – $Mg = 4.49 > 3.865$ donc on descend dans la sous-branche de droite,
   et par conséquent l'individu est classé dans le groupe `WinNF` (donc mal classé puisqu'il est du groupe `WinF`).

8. On trouve des taux d'erreurs moyens comparables (autour de 0.3) que l'on développe les arbres jusqu'au bout ou non. Donc en tout cas le fait de développer jusqu'au bout n'apporte pas d'amélioration et c'est forcément plus lent puisque les arbres sont plus grands. De plus la tendance est plutôt inverse : les arbres entièrement développés ont plutôt des performances légèrement inférieures : on a tendance à faire du surapprentissage.

10. On trouve un taux d'erreurs moyen inférieur (de l'ordre de 0.25), donc la méthode est légèrement plus performante. Le bagging introduit de l'aléa dans la construction des arbres afin de pouvoir ensuite faire une moyenne des classifications obtenues.