Université Paris Descartes
UFR de Mathématiques et Informatique

Classification Master 1 Ingénierie Mathématique
Examen final du 4 mai 2018

On écrira toutes les commandes demandées dans l'éditeur en prenant soin d'enregistrer le fichier à la fin de la séance (peu importe le nom et l'emplacement du fichier). On pourra aussi enregistrer les résultats graphiques dans des fichiers pdf (Menu "Export" puis "Save as pdf").

Exercice 1

Les commandes R suivantes permettent de construire un jeu de données synthétique contenant 400 observations avec deux variables quantitatives X^1, X^2 et une classe $Y \in 0, 1$ pour chaque observation.

```
N = 100
donnee= data.frame(X1=c(runif(N,0,1),runif(N,2,3),runif(N,2,3),runif(N,4,5)),
                   X2=c(runif(N,2,3),runif(N,0,1),runif(N,2,3),runif(N,0,1)),
                   Y=c(rep(0,2*N),rep(1,2*N)))
plot(donnee$X1,donnee$X2,pch=2+donnee$Y)
```

La représentation graphique de ces données est reproduite sur la feuille annexe en a), où les triangles correspondent aux individus de la classe $Y = 0$ et les croix aux individus de la classe $Y = 1$.

1. Recopier les commandes R puis construire avec R un arbre de classification grâce à la méthode CART, avec les options par défaut. Afficher sous R l'arbre de classification obtenu.
2. L'arbre obtenu est du type b) (cf feuille annexe). Compléter l'arbre b) avec les variables et valeurs obtenues, puis représenter graphiquement la classification sur la figure a) en traçant la frontière entre les deux classes. Que vaut le taux d'erreurs d'apprentissage ?
3. La même règle de classification peut en fait être obtenue avec un arbre de type c). Compléter cet arbre afin d'obtenir la même règle.
4. Expliquer pourquoi la méthode CART ne pouvait pas aboutir à l'arbre c) dans cet exemple.
5. Rappeler le principe de la méthode des forêts aléatoires (sans réexpliquer la méthode CART). Quel étape ou élément de cette méthode permet de dire que dans l'exemple précédent la méthode des forêts aléatoires permettrait de générer aussi bien des arbres du type b) que du type c) ?

Exercice 2

1. Rappeler le principe de la méthode des machines à vecteurs de support (SVM) dans le cas linéaire et lorsque les classes sont séparables par un hyperplan. On précisera la fonctionnelle à maximiser ou minimiser correspondant au problème résolu par la méthode.

Sur la feuille annexe est représentée un exemple de jeu de données avec deux variables quantitatives X^1, X^2, et deux classes $Y \in \{0, 1\}$. Les observation de la classe $Y = 0$ sont représentées avec des triangles, et les individus de la classe $Y = 1$ avec des croix.

2. Trouver graphiquement la solution du problème de SVM linéaire : Tracer l'hyperplan séparateur et entourer les individus correspondant aux vecteurs support, et représenter la marge maximale obtenue. Expliquer votre raisonnement. Quel taux d'erreurs d'apprentissage obtient-on ?

Exercice 3

Les données Glass de la librairie e1071 contiennent différentes analyses chimiques relevées sur des morceaux de verre de différents types. Le but est de parvenir à prédire le type de verre en fonction de ses caractéristiques chimiques.

1. Charger les données puis les observer pour comprendre comment elles sont organisées. Expliquer ce que contiennent les différents éléments du jeu de données.
2. Effectuer une classification aux 3 plus proches voisins sur toutes les observations, en cherchant à classer les données d'apprentissage elles-mêmes. Déterminer le taux d'erreurs obtenu.
3. On fait varier à présent le paramètre k de la méthode des k-plus proches voisins afin de comparer les performances. Écrire des commandes R permettant de représenter sur un même graphique les taux d'erreurs d'apprentissage et les taux d'erreurs obtenus par validation croisée "leave-one-out" en fonction de k. Commenter le graphique obtenu. Quelle valeur de k semble la meilleure ?
4. Utiliser à présent la méthode CART pour classer les observations : construire un arbre de classification avec la méthode CART en développant l'arbre jusqu'au bout, afficher l'arbre obtenu, puis classer les données d'apprentissage et calculer le taux d'erreurs.
5. Effectuer un élagage de l'arbre, puis déduire le sous-arbre qui semble optimal d'après la validation effectuée directement par la fonction rpart.
6. On cherche maintenant à vérifier "à la main" que le sous-arbre ainsi détecté est optimal. Écrire des commandes R permettant de faire de la validation croisée par "leave-one-out" sur le jeu de données et avec la méthode CART. Comme précédemment avec la méthode k-NN, représenter sur un même graphique les taux d'erreurs d'apprentissage et les taux d'erreurs obtenus par validation croisée "leave-one-out" en fonction cette fois des étapes de l'élagage de l'arbre final. Commenter le graphique obtenu. Quel sous-arbre semble optimal ?