SY19 – A18
TP 6: Arbres de décision, méthodes d'ensemble, bootstrap

Exercice 1 : Filtrage de spams

Le fichier `spambase.dat` contient des données relatives à 4601 courriels décrits par 58 variables. La variable correspond à la dernière colonne du tableau est binaire est indique si le courriel est un spam ou non (cf. fichier `spambase.names` pour une description de ces données.

1. Partitionner aléatoirement les données en un ensemble d'apprentissage et un ensemble de test.
2. Construire un arbre de décision sur ces données. Représenter graphiquement cet arbre. Estimer sa probabilité d'erreur de test.
3. Appliquer à l'arbre précédent la procédure d'élagage, en choisissant le meilleur arbre par validation croisée. Représenter graphiquement l'arbre obtenu. Calculer la matrice de confusion et le taux d'erreur de test. Peut-on interpréter l'arbre obtenu ?
4. Appliquer le bagging et les forêts aléatoires sur ces données. Pour les forêts aléatoires, optimiser le paramètre m par validation croisée.

Exercice 2 : Bootstrap

On considère à nouveau les données `prostate` du TP2, en prenant la variable `lpsa` comme variable à expliquer.

1. En utilisant le bootstrap, estimez les erreurs standards et des intervalles de confiance à 95% sur les coefficients de la régression linéaire. Comparez les résultats obtenus avec les estimateurs basés sur le modèle gaussien (résultats renvoyés par la `reg`).
2. Estimer les coefficients uniquement sur les données d'apprentissage (train=TRUE), et prédire la valeur de `lpsa` sur les données de test. Représenter les valeurs prédites et les intervalles de confiance à 95% sur l'espérance conditionnelle de la variable `lpsa` en fonction des valeurs observées, calculés sous hypothèse gaussienne (par la fonction `predict`), et estimés par boostrap.

3. On utilise cette fois la régression ridge, avec le coefficient de régularisation λ déterminé par validation croisée. Peut-on encore utiliser les formules du cours pour estimer les erreurs standards des coefficients dans ce cas ? Utilisez le bootstrap. Que constatez-vous ?
4. Calculez les intervalles de confiance à 95% sur l'espérance conditionnelle de la variable `lpsa` en fonction des valeurs observées pour les données de test, en utilisant la régression ridge et le boostrap. Comparez ces intervalles avec ceux obtenus en réponse à la question 2.