

ALGORITHMES D'APPRENTISSAGE STATISTIQUE ET MÉTHODES ENSEMBLISTES: L'EXEMPLE DES FORÊTS ALÉATOIRES

Vous aurez probablement besoin pour ce TP des librairies R suivantes :

randomForest

Partie 1 : Forêts aléatoires de classification.

- (1) Par opposition aux arbres de décision, nous disposerons ici d'un estimateur agrégé (estimateur ensembliste). Quel en est l'intérêt ?
- (2) Quels sont les paramètres de tuning dans cet algorithme ?
- (3) Reprenons les données d'iris, sur lesquelles on cherchera toujours à prévoir l'espèce. La construction d'une forêt se fait avec **randomForest**. Pour cette fonction :
 - i) Comprendre son utilisation : consulter l'aide de la fonction et...
 - a) : ...comprendre chacun de ses arguments ;
 - b) : ...comprendre ce qu'elle retourne (attributs) ;
 - ii) Implémenter l'algorithme :
 - a) construire une forêt de 500 arbres, en tirant aléatoirement 2 des variables explicatives pour chaque étape de division de l'espace des covariables ;
 - b) afficher les résultats numériques de la forêt : notamment, combien vaut l'erreur out-of-bag (OOB) ? Matrice de confusion de la forêt ?
 - c) pouvez-vous afficher des résultats graphiques de la forêt ?
 - d) accéder aux attributs de l'objet. Combien y en a-t-il ? Les comprendre un par un. Quels sont les taux d'erreur de classification par espèce et par arbre ?
 - e) Quelle est l'importance de chacune des variables explicatives ? Après avoir consulté les résultats de cette importance, afficher la graphiquement avec **varImpPlot**. Quelle est la variable la plus importante ? Les résultats de la forêt sont-ils cohérents avec ceux des arbres CART du précédent TP ?
 - f) Grâce à la commande **plot** sur l'objet de classe **randomForest**, afficher l'évolution des erreurs OOB et de classification en fonction du nombre d'arbres de la forêt. Qu'observez-vous ? Combien d'arbres vous semble-t-il raisonnable de prendre pour construire la forêt ?
 - g) Avec la méthode **treeSize**, faire un histogramme de la taille des arbres de votre forêt. Dispersion ? Que cela traduit-il ?
 - h) Vous n'avez à ce stade qu'optimisé le nombre d'arbres à construire. Sur quels autre(s) paramètre(s) avez-vous la main ? Vous pouvez optimiser le paramétrage de votre modèle grâce aux fonctions **tuneRF** ou **tune**.
 - i) Quel est donc le meilleur couple (*ntree*, *mtry*) à vos yeux ?

Partie 2 : Forêts aléatoires de régression.

- (1) Reprenez les mêmes étapes que précédemment avec les données **cu.summary**.
- (2) Comment intégreriez-vous des pondérations de vos observations (Exposure) ?
- (3) D'autres librairies existent : **caret**, **ranger**, **randomForestSRC**. Les consulter.