INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN NORMANDIE

Mastère Spécialisé "Expert en Sciences des Données" Normandie Univ - INSA Rouen

Cours Arbres CART et Forêts aléatoires - Importance et sélection de variables

Jean-Michel POGGI

TP Partie 2 : Forêts aléatoires

Guide pour le TP avec le scénario compagnon, la documentation et les deux articles :

`cran.r-project.org/web/packages/VSURF/index.html`
`journal.r-project.org/archive/2015-2/genuer-poggi-tuleaumalot.pdf`
`hal-descartes.archives-ouvertes.fr/hal-01387654v2`

Un compte-rendu de TP unique et global portant sur les deux parties, validera les deux journées de formation (à rédiger seul ou en binôme).
Veuillez envoyer un pdf (et uniquement un pdf) à l'adresse suivante :
`Jean-Michel.Poggi@parisdescartes.fr`

1. Les données (voir TP1)
   1. Charger la librairie `kernlab`
   2. Charger le jeu de données `spam` dans R construire les *dataframes* d'apprentissage et de test (qui servira à évaluer les erreurs)

2 et 3. Arbres CART (voir TP1)

4. Les forêts aléatoires

1. Charger la librairie `randomForest`
2. Construire une RF pour *mtry=p* (bagging non élagué) et calculer le gain en termes d'erreur par rapport à un arbre seul
3. Construire une RF par défaut
4. Calculer l'erreur et comparer au bagging
5. Etudier l'évolution de l'erreur OOB en fonction *ntree* en utilisant `do.trace`

5. L'importance des variables

1. Calculer l'importance des variables de `spam` pour la RF par défaut
2. Quelles sont les variables les plus importantes ?
3. Calculer l'importance des variables de spam pour la RF de stumps
4. Illustrer de l'influence du paramètre *mtry* sur l'erreur OOB ainsi que sur la VI

6. Sélection de variables à l'aide des forêts aléatoires

1. Charger la librairie `VSURF`
2. Appliquer `VSURF` sur un sous-ensemble de 500 observations du tableau de données `spam.app`
3. Commenter les résultats des différentes étapes
4. Expérimenter la version parallèle en vous basant sur l'article consacré à `VSURF` (hors CR, question subsidiaire)