HPC-Big Data 2021 : Apprentissage Statistique TP 2

Modèles additifs généralisés (GAM)

Description des données :

Les méthodes testées dans ce TP seront les modèles additifs généralisés en régression puis discrimination.

Un producteur éolien utilise des prévisions de force de vent pour en déduire, par exploitation de la courbe de réponse des éoliennes, sa production électrique du lendemain. Afin d'assurer l'équilibre du réseau national de transport d'électricité, cette prévision de production est en effet exigée par le gestionnaire RTE du réseau français.
Non satisfait des prévisions de vent calculées par un modèle météorologique, ce producteur vous demande de lui fournir un modèle statistique capable d'améliorer les scores de prévisions de force de vent sur son parc éolien. Il aurait également besoin, pour son activité, de prévisions d'occurrence de dépassement du seuil de **13 m/s**.

Vous disposez d'une archive de prévisions de différentes variables, issues du modèle météorologique exploité jusqu'alors par votre client, ces prévisions constituant de potentiels prédicteurs pour vos modèles statistiques, ainsi que de l'archive correspondante des mesures de force de vent effectuées sur le parc éolien du producteur.

Le fichier ***DataTP.txt*** contient les 11 variables suivantes :

HU : humidité relative prévue en %

N : nébulosité (= couverture nuageuse) prévue en octas (entiers de 0 à 8)

P : pression prévue en hPa

u : composante zonale du vent prévue en m/s

v : composante méridienne du vent prévue en m/s

hel : hélicité prévue en m^2/s^2 (indice de vorticité)

DD : direction du vent prévue en rad

mois : mois de validité de la prévision

heure : heure de validité de la prévision

FFp : force du vent prévue en m/s

FFo : force du vent **observée** en m/s.

1. **Chargement des librairies et des données :**

Après installation, charger les packages : ***gam, akima***

Charger les données dans une data.frame :
data=read.table("DataTP.txt",header=TRUE)

2. **Modèles GAM :**

- La librairie ***gam*** propose via sa fonction ***gam*** d'estimer des modèles additifs généralisés par application d'un algorithme de backfitting. Les méthodes d'estimation non-paramétrique proposées par la fonction ***gam*** sont les smoothing splines ***s()*** et les régressions locales LOESS ***lo()*** : ***?gam ; ?s ; ?lo***.

- Estimer différents modèles GAM en régression dans le but de prévoir la variable **FFo**.
  Analyser les fonctions lisses estimées (***plot***).
  Analyser les résumés des modèles (***summary***).
  Comment piloter la flexibilité des fonctions lisses ?
  Comment sélectionner les prédicteurs ?

  Des interactions entre prédicteurs peuvent être introduites en exploitant les régressions locales, tester par exemple ***lo(u,v)***. Les fonctions lisses 2D estimées seront visualisables par la fonction ***plot*** après chargement du package ***akima***.

  Confronter, en exploitant votre procédure d'évaluation, les performances des modèles GAM avec celles des modèles linéaires gaussiens en terme de **RMSE**. Conclure.

- De même, estimer des modèles GAM en discrimination pour prévoir le prédictand **OCC** défini au TP précédent, puis comparer avec la régression logistique et l'analyse discriminante en terme de **BS**, **PSS** ou **ROC AREA**. Conclure.