HPC-Big Data 2021 : Apprentissage Statistique TP 1

Modèle linéaire gaussien - Régression logistique - Analyse discriminante

Description des données :

Le but de ce TP est la mise en place d'un protocole de sélection du meilleur modèle statistique dans un but de prévoir un prédictand quantitatif (force de vent) puis qualitatif (dépassement d'un seuil de force de vent). Les méthodes testées dans ce premier TP seront le modèle linéaire gaussien, la régression logistique et l'analyse discriminante.

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 :**

Charger le package '**MASS**' :
library(MASS)
Charger les données dans une data.frame :
data=read.table("DataTP.txt",header=TRUE)

2. **Etude préliminaire :**

- Faire une analyse des données dont vous disposez.

- Représenter la série de force de vent observée (**FFo**) et ajouter sur ce graphe des croix bleues pour les prévisions **FFp** du modèle météorologique et des croix rouges pour les estimations issues de la régression de **FFo** par **FFp**. Calculer les **Biais** et **RMSE**.

3. **Régression - Modèle linéaire gaussien :**

- Estimer le modèle linéaire gaussien exploitant l'ensemble des prédicteurs potentiels, et afficher le bilan de l'estimation (***summary***), ainsi que les graphiques des diagnostics. Quels prédicteurs conserveriez-vous ?

- Estimer le modèle exploitant l'ensemble des prédicteurs mais également toutes les interactions d'ordre 2 possibles. Afficher le bilan de l'estimation, ainsi que les graphiques des diagnostics. Comparez au précédent modèle.

- **Sélection automatique des prédicteurs** : fonction ***stepAIC***

  *Le critère d'Akaike (**AIC** - Akaike Information Criterion)* :

  On cherche le modèle qui minimise l'indice **AIC**, q étant la dimension du modèle :

$$\mathbf{AIC} = \ln\left(\frac{\left\|\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\beta}}\right\|^2}{\mathbf{n}}\right) + \frac{\mathbf{2q}}{\mathbf{n}}$$

  Effectuer une sélection AIC descendante à partir des 2 modèles précédents.

  *Le critère de Schwartz (**BIC** – Bayesian Information Criterion)* :

  On cherche le modèle qui minimise l'indice **BIC**, q étant la dimension du modèle :

$$\mathbf{BIC} = \ln\left(\frac{\left\|\mathbf{Y} - \mathbf{T}\hat{\boldsymbol{\beta}}\right\|^2}{\mathbf{n}}\right) + \frac{\mathbf{q}\ln(\mathbf{n})}{\mathbf{n}}$$

  Effectuer une sélection **BIC** descendante à partir des 2 modèles précédents.
  Comparer aux modèles issus de la sélection automatique basée sur l'**AIC**.

4. **Evaluation des modèles :** surapprentissage, robustesse, validation croisée

- Coder une fonction calculant les **Biais** et **RMSE** des prévisions. Calculer alors ces scores pour les 4 modèles précédents. Que concluez-vous ?

- A l'aide des fonctions R ***sample*** et ***setdiff***, créer un fichier d'apprentissage datapp contenant 80% des données et un fichier de test datatest avec les données restantes.

- Réestimer les 4 modèles précédents sur les données d'apprentissage puis calculer les scores sur apprentissage puis sur test (fct ***predict***). Quelle analyse pouvez-vous faire ?

- Coder une procédure de validation croisée permettant l'analyse de la robustesse des modèles testés ainsi que de la significativité des résultats (sensibilité des estimations à l'échantillonnage) pour pouvoir définir le meilleur modèle. Votre code devra au final générer un graphe contenant des boîtes à moustaches du score **RMSE** pour les différents modèles, sur apprentissage et test.

- Quel est finalement le meilleur modèle ? Illustrer le phénomène de sur-apprentissage.

- Retracer la série relative à la force de vent observée (**FFo**) en ajoutant des croix de couleurs différentes pour les prévisions de force de vent (**FFp**) et les estimations de votre meilleur modèle de régression.

5. **Discrimination - Régression logistique et Analyse discriminante** :
Prévision du dépassement du seuil de vent de 13 m/s.

Après installation si nécessaire, charger les packages '**MASS**' et '**verification**' :
library(MASS)
library(verification)

- Ajouter à la data.frame **data** deux nouvelles variables, **OCC** et **OCCp**, de type factor, pour respectivement l'occurrence observée de dépassement du seuil et l'occurrence prévue par le modèle météorologique. Les occurrences seront codées 1 et les non-occurrences 0.

- Exécuter le script **scores.R** qui permet d'utiliser la fonction **scores** (cf. Annexe).

- **Régression logistique** : fonction ***glm***

  Estimer plusieurs modèles de régression logistique, avec et sans interactions, en utilisant les procédures automatiques de sélection de prédicteurs.
  Exploiter la courbe ROC pour choisir le seuil de probabilité permettant de définir la prévision déterministe de dépassement du seuil de 13 m/s à partir des prévisions probabilistes issues du modèle logistique (fct ***roc.plot***).
  Calculer les scores issus de la table de contingence ainsi que le **Brier Score** (fct ***brier***) sur apprentissage puis test.

- Au final, effectuer une validation croisée pour pouvoir comparer rigoureusement sur apprentissage comme sur test les différents modèles de régression logistique en terme de **BS** (Brier Score) puis de **PSS** (Score de Peirce). Conclure.

- **Analyse discriminante linéaire et quadratique** : fonctions ***lda*** et ***qda***

  Les analyses discriminantes exploitent uniquement des **prédicteurs quantitatifs**.
  La bibliothèque standard **MASS** de R pour l'analyse discriminante ne propose pas de procédure automatique de choix de prédicteurs. Une telle procédure exploiterait la maximisation de la distance de Mahalanobis avec un test approprié.

  Estimer plusieurs modèles d'analyse discriminante, linéaire puis quadratique.
  Quels prédicteurs garderiez-vous ? Justifiez.
  Exploiter la courbe ROC pour choisir le seuil de probabilité permettant de définir la prévision du dépassement du seuil de 13 m/s.
  Calculer les scores issus de la table de contingence ainsi que le Brier Score sur apprentissage puis test.

- Effectuer une validation croisée pour pouvoir comparer rigoureusement sur apprentissage comme sur test vos modèles d'analyse discriminante. Au final, confronter les méthodes statistiques : régression logistique et analyse discriminante. Conclure.

Annexe - Rappels sur les scores :

SCORES ELABORES A PARTIR D'UNE TABLE DE CONTINGENCE :

On note A l'occurrence du phénomène A, et NA sa non-occurrence.

	OBSERVE	
PREVU	A	NA
A	a	b
NA	c	d

Taux de succès global : (a+d)/(a+b+c+d). Pas forcément un bon indicateur (si A rare).

H taux de bonnes prévisions, F taux de fausse alerte :

$$H = \frac{a}{a+c} \qquad F = \frac{b}{b+d}$$

Score de Peirce PSS=H-F $-1 \leq PSS \leq 1$.

SCORE PROBABILISTE : LE BRIER SCORE

Mesure la performance de la prévision probabiliste d'occurrence d'un événement binaire : BS = erreur quadratique moyenne de la probabilité prévue

$$BS = \frac{1}{M}\sum_{i=1}^{M}(p_i - o_i)^2$$

o_i = 1 si occurrence, 0 sinon
p_i est la probabilité d'occurrence prévue
M= nb de prévisions

COURBE ROC :