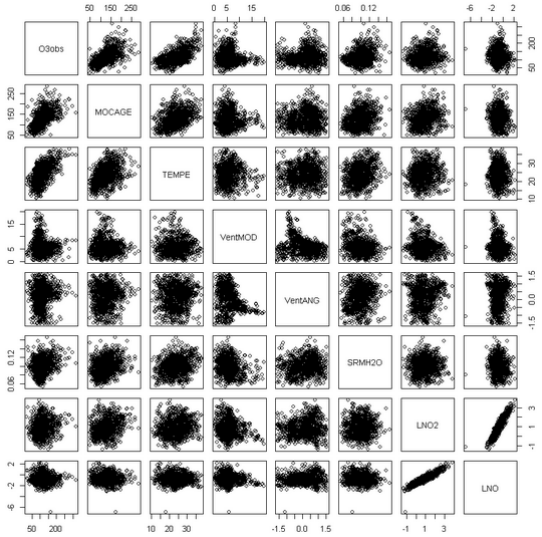
Compte-rendu TP "Prévision du Pic d'Ozone" avec R pour le cours "Intelligence et apprentissage artificielle"

Michel Nguyen
Etudiant en master 2 cybersécurité & sciences des données

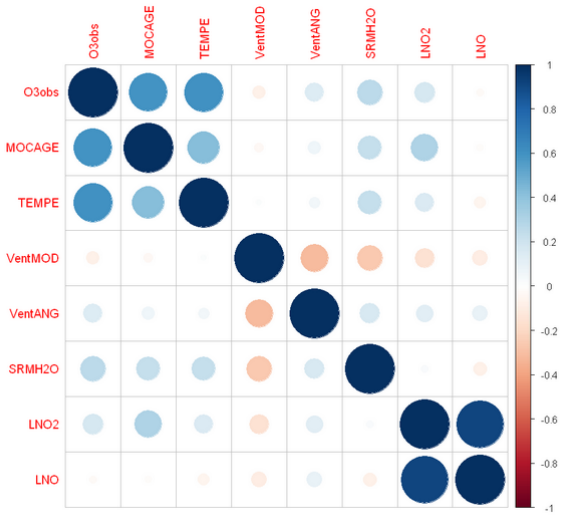
Novembre 2021

1 Introduction

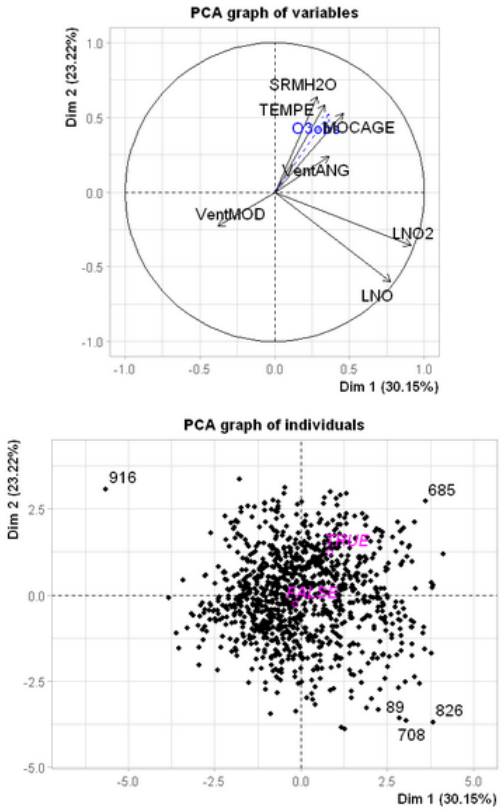
Que dire sur les relations des variables 2 à 2 ?
Pour la plupart des graphes, on peut remarquer une certaine corrélation car certains forment un amas de points et d'autres forment une droite linéaire.



Compléter en visualisant les corrélations avec la fonction 'corrplot' (package corrplot). Quelle est la limite de ce type de diagnostic numérique : quel type de corrélation est mesuré ?
On remarque 3 couples fortement corrélés (>0.6) qui sont : MOCAGE & O3obs, MOCAGE & TEMPE, LNO & LNO2. La limite est que ce diagnostic ne prend pas en compte l'unité de mesure des variables.



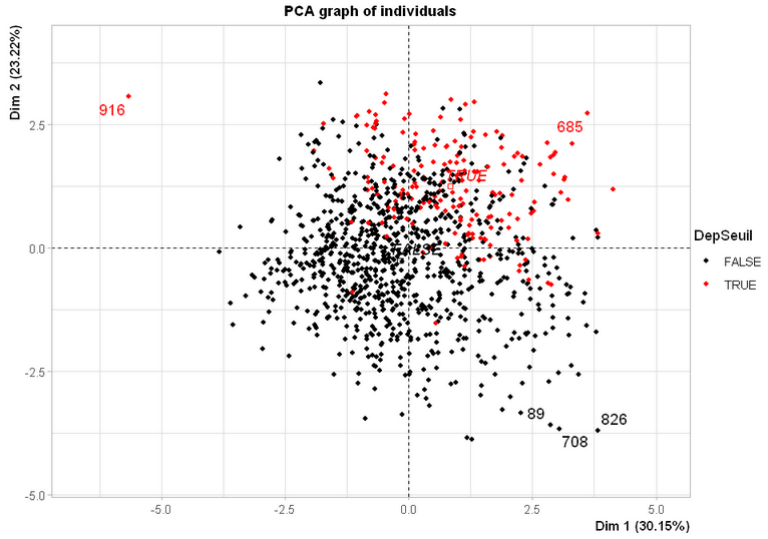
Que sont ces graphiques ?
Ce sont des graphes obtenus par ACP. On peut identifer des clusters c'est-à-dire des caractéristiques corrélés en étant proches l'un de l'autre. Le premier graphe est en forme de cercle, l'autre est sur un plan en deux dimensions.



Que dire du choix de la dimension, des valeurs atypiques ?
On remarque que l'addition des deux dimensions vaut environ 50%. Ceci signifie que les deux dimensions représentent 50% des variations des données.

Que dire de la structure de corrélation des variables ? Est-elle intuitive ?
Pour le premier graphe, il est lisible et clair car on peut observer les caractéristiques qui sont corrélés. Il est donc intuitive. Trois groupes se distinctes : SRMH2O/TEMPE/MOCAGE/Q3obs/VentANG ; VENTMOD ; LNO2/LNO.
Concernant le deuxième graphe, on observe un regroupement autour de l'origine. Elle manque d'intuitivité car il y a beaucoup de points qui sont représentés. Donc, le premier graphe semble plus intuitive que le deuxième.

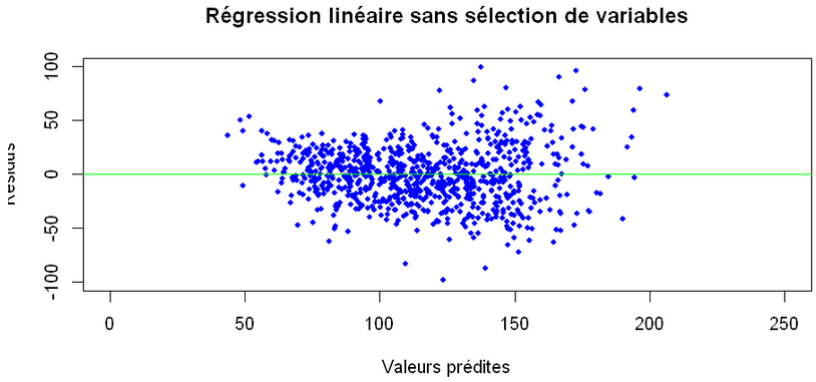
Une discrimination linéaire (hyperplan) semble-t-elle possible ?
Oui, car les points TRUE sont majoritairement dans le sous-plan $\{x > 0, y > 0\}$. Donc on peut tracer une droite pour séparer les points TRUE et FALSE.



Commenta appelle-t-on cette procédure spécifique de validation croisée ?
Il s'agit de la validation croisée à k blocs.

2 Prévision par modèle gaussien

Que dire de la distribution de ces résidus ?
La distribution des résidus se concentre autour de $[0, \pm 50]$.



La forme du nuage renseigne sur les hypothèses de linéarité du modèle et d'homoscédasticité. Que dire de la validité de ce modèle ?
Le modèle possède de l'homoscédasticité car pour $x \in [50, 150]$, la dispersion des résidus est semblable quelque soit x. Puisque la plupart des résidus se situe entre $[0, \pm 50]$ et que $x \in [50, 150]$, la valeur des résidus semblent trop importantes, donc ce modèle n'est pas performant.

Ce premier modèle est comparé avec celui de la seule prévision déterministe MOCAGE. Qu'en conclure ?
On observe que plusieurs caractéristiques sont significatives : c'est le cas pour TEMPE, VentANG et LNO. Par ailleurs, leurs coefficients sont plus élevés que celui de MOCAGE (0.38).

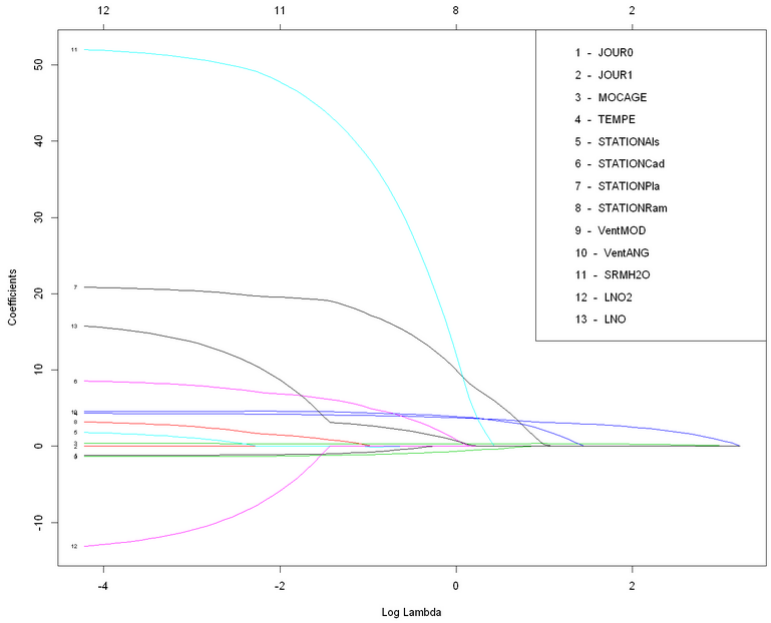
```
             Df Sum Sq Mean Sq F value   Pr(>F)
JOUR          1    106     106   0.134 0.714791
MOCAGE        1 470173  470173 590.680  < 2e-16 ***
TEMPE         1 225427  225427 283.204  < 2e-16 ***
STATION       4  10163    2541   3.192 0.012926 *
VentMOD       1  13846   13846  17.395 3.36e-05 ***
VentANG       1  10088   10088  12.673 0.000392 ***
SRMH2O        1    273     273   0.343 0.558101
LNO2          1   3337    3337   4.193 0.040918 *
LNO           1   9006    9006  11.314 0.000805 ***
Residuals   819 651913     796
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Entrée [153]: `coef(reg.lm)`

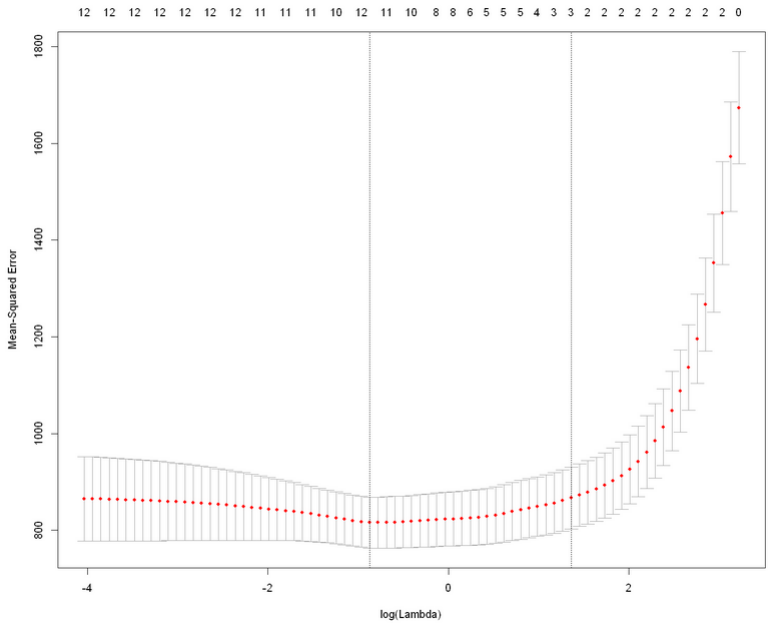
(Intercept)	-16.0273450980118
JOUR1	1.21444628791877
MOCAGE	0.384981448178631
TEMPE	4.32704535988958
STATIONAls	2.14259575252116
STATIONCad	8.81157992073352
STATIONPla	21.0287163608663
STATIONRam	3.46281508807015
VentMOD	-1.375691793879
VentANG	4.5070676797891
SRMH2O	52.5682991119587
LNO2	-14.2991137044367
LNO	16.9322029309928

Que fait la commande model.matrix ? Comment sont gérées les variables catégorielles ?
La commande model.matrix permet de créer une matrice à partir d'un jeu de données. Les variables catégorielles vont se transformer en variables numériques où chaque valeur correspond à une classe (dummy variable).

Que représentent les courbes ci-dessus, appelées "chemins de régularisation" ?
Il s'agit d'un modèle de régression LASSO. Les courbes représentent les valeurs des coefficients selon le paramètre lambda de LASSO.



Que représente la courbe rouge ? Et la bande qui est autour ? Comment sont obtenues les valeurs de log(lambda) correspondant aux lignes verticales en pointillé ?
Chaque point rouge de la courbe rouge est le centre du *min* et le *max* de la bande blanche qui est le calcul de l'erreur quadratique moyenne en faisant varier les paramètres.



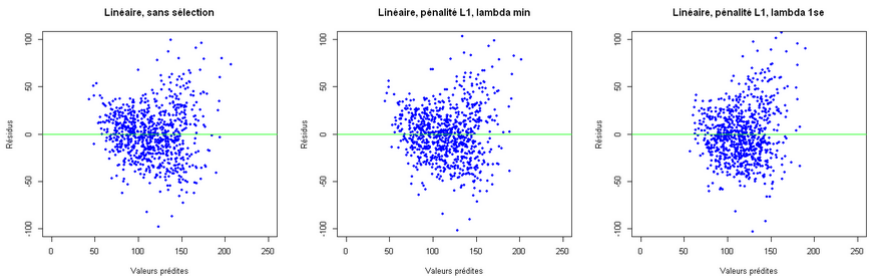
Combien restent-ils de coefficients non nuls. Vérifier sur les chemins de régularisation.
Il reste trois coefficients non nuls : MOCAGE, TEMPE et VentANG.

```
'CV estimate of lambda : 3.896'

14 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept) 1.1658431
JOUR0       .
JOUR1       .
MOCAGE      0.3370194
TEMPE       2.9855716
STATIONAls  .
STATIONCad  .
STATIONPla  .
STATIONRam  .
VentMOD     .
VentANG     0.4939696
SRMH2O      .
LNO2        .
LNO         .
```

Même question en choisissant l'autre valeur de lambda retenue par glmnet, i.e. "reg.lasso.cv$lambda.min"
Il reste 11 coefficients non nuls. Les coefficients nuls sont : STATIONRam et LNO2.

```
'CV estimate of lambda : 0.418'

14 x 1 sparse Matrix of class "dgCMatrix"
                          1
(Intercept) -1.802987e+01
JOUR0       -6.710120e-01
JOUR1        5.446190e-14
MOCAGE       2.991228e-01
TEMPE        4.005317e+00
STATIONAls  -9.333231e-02
STATIONCad   4.713484e+00
STATIONPla   1.673098e+01
STATIONRam   .
VentMOD     -1.065688e+00
VentANG      4.327967e+00
SRMH2O       3.578514e+01
LNO2         .
LNO          2.550344e+00
```

Commenter.
On observe que les trois graphes sont similaires. Donc l'application de LASSO et selon lambda ne modifie pas réellement le modèle.



Calculer le critère MSE (moyenne des carrés des résidus) pour les deux modèles. Pourquoi celui obtenu par LASSO est-il moins bon ? Quel critère LASSO minimise t-il ?

```
'Modèle linéaire sans séletion: 783.549239590938'
'LASSO avec lambda.min: 793.58476403388'
'LASSO avec lambda.1se: 859.648321090238'
```

Estimer l'erreur de généralisation du modèle de régression linéaire simple sans sélection de variables par validation croisée. Comparer avec celle du LASSO. Qu'observez-vous ?

Quel autre critère, équivalent à AIC dans le cas gaussien et de variance résiduelle connue, est utilisée en régression linéaire?
L'autre critère est la p-value.

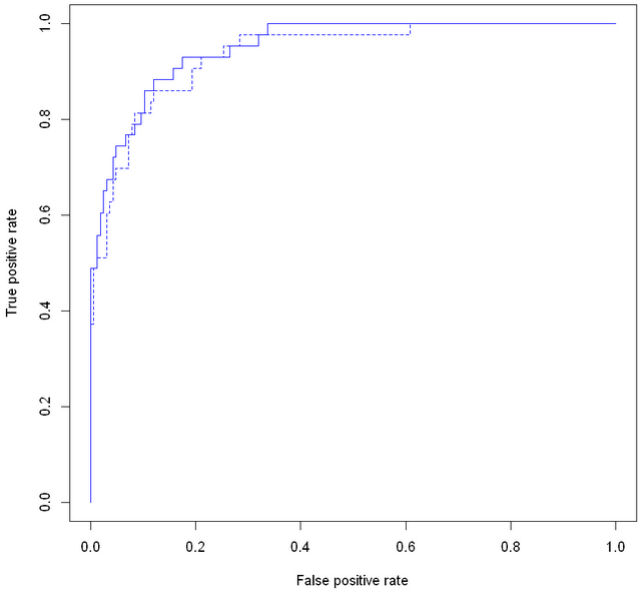
3 Prévision par modèle binomial

Comparer avec l'approche précédente. Mémoriser les résultats obtenus pour comparer avec les autres méthodes.
Cette méthode est un peu mieux car le taux de précision est plus grand que celui de la première méthode c'est-à-dire qu'on obtient 0.89 contre 0.88 pour la première méthode.

Que sont sensibilité et spécificité d'une courbe ROC ?
La sensibilité est le taux de vrais positifs et la spécificité est le taux de faux positifs.

Les performances des deux approches gaussiennes et binomiales sont-elles très différentes ?

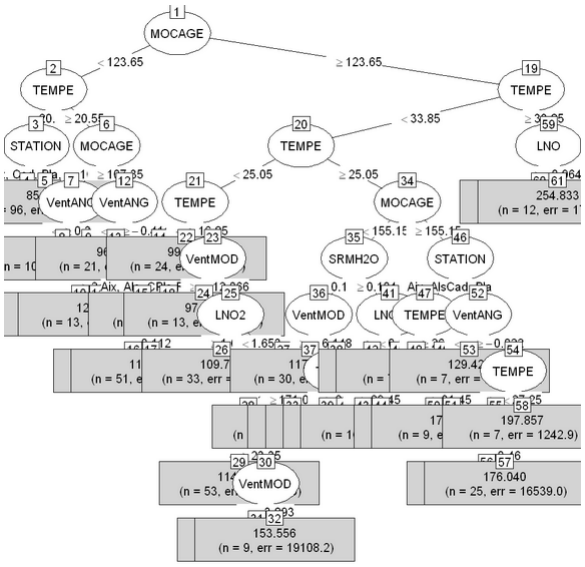
Sur le graphe ci-dessus, ajouter la courbe ROC pour le modèle déterministe MOCAGE. Qu'observez-vous?
On observe que le modèle (courbe en bleu discontinu) a une bonne performance car le taux de vrais positifs est de 0.9 pour un taux de faux positifs de 0.3.



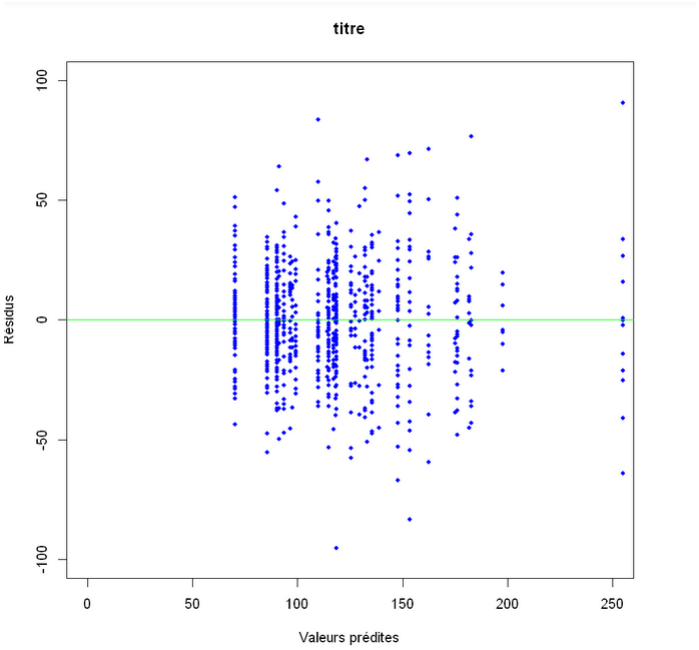
4 Arbre de décision binaire

Quel critère est optimisé lors de la création d'un noeud ? de l'arbre ?
Le critère optimisé lors de la création d'un noeud est le paramètre cp, alors que pour l'arbre, il s'agit du paramètre maxdepth.

Quelle est la variable qui contribue le plus à l'interprétation ?
La variable qui contribue le plus est la variable de la racine : MOCAGE.



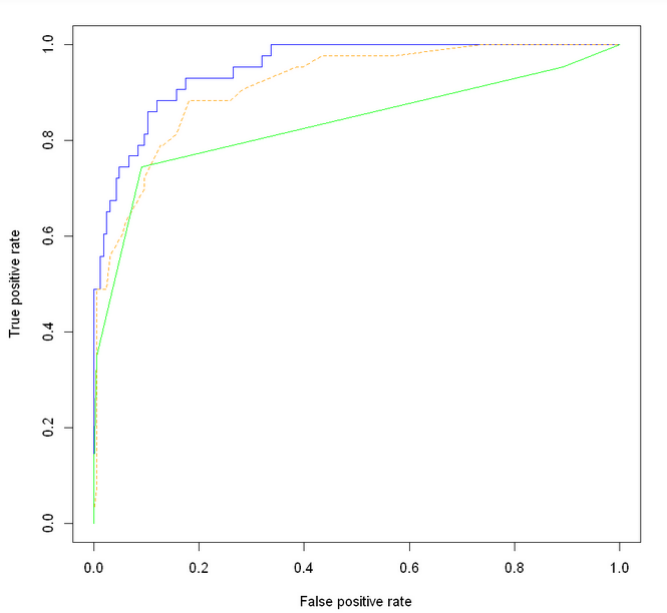
A quoi est due la structure particulière de ce graphe ?
Elle est due à la dispersion des résidus.



Quel autre critère d'hétérogénéité est utilisé ?
L'autre critère est le paramètre cp=0.001.

Quelle stratégie semble meilleure à ce niveau ?
Sur mes résultats obtenus, j'ai un taux de précision de 0.88 pour la matrice de confusion qui concerne la régression, l'autre est de 0.86. Donc la meilleur stratégie à retenir est celle de la régression.

Une meilleure méthode se dégage-t-elle?
La méthode de régression (courbe en jaune) se distingue lorsque $x > 0.1$ car le taux de vrais positifs est supérieur à celui de la méthode de discrimination (courbe en vert).



5 Réseau de neurones

Quelle fonction de transfert pour le dernier neurone en régression ?
La fonction à utiliser est la fonction Unité Linéaire Rectifiée (ReLU).

Quelle fonction de transfert pour le dernier neurone en discrimination binaire ?
La fonction à utiliser est la fonction de seuil.

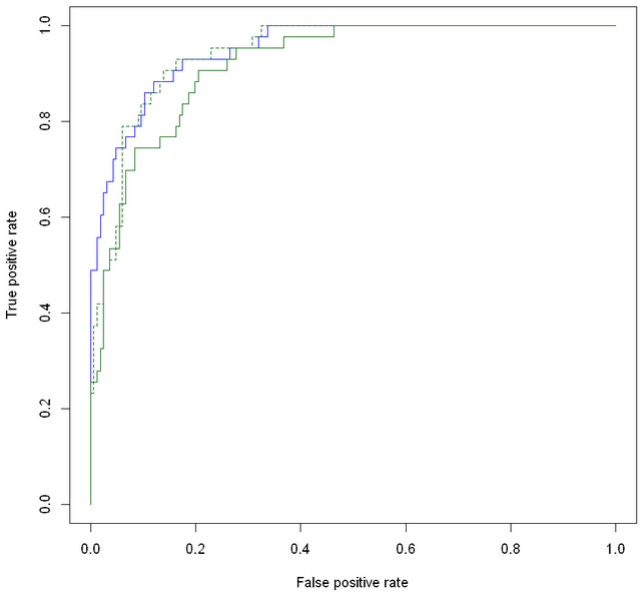
Quid de la discrimination avec plusieurs classes ?
La fonction à utiliser est la fonction softmax.

Quel est le choix par défaut pour les neurones de la couche cachée ?
Le choix par défaut est la fonction Unité Linéaire Rectifiée (ReLU).

Quel est le paramètre decay de la fonction nnet ?
Le paramètre decay est un paramètre qui influe sur l'ajustement des poids. Sa valeur par défaut est 0.

Indiquer une autre façon d'éviter le sur-apprentissage.
Une autre façon d'éviter le sur-apprentissage est d'exploiter le paramètre decay en cherchant la meilleure valeur. En effet, la valeur par défaut qui est 0 tend le modèle à sur-apprendre le jeu de données d'apprentissage.

Une méthode semble-t-elle significativement meilleure?
Une méthode semble se distinguer. Il s'agit de la courbe de régression (courbe en vert discontinue) car le taux de vrais positifs est supérieur pour presque tout $x \in \{0, 1\}$ à celui de la courbe de discrimination (courbe en vert).



6 Bilan

Tout d'abord, je souhaite rappeler l'objectif du cours : comprendre le raisonnement mathématique derrière un concept de machine learning, ainsi que chercher des mesures pour quantifier l'observation.

Grâce à ce TP écrit en langage R, j'ai pu m'immerger dans la peau d'un véritable ingénieur, qui raisonne et cherche la meilleur solution parmi les différentes techniques de machine learning : l'analyse en composantes principales, la régression logistique, l'arbre de décision binaire et les réseaux de neurones. Aussi, j'ai appris plusieurs notions : la courbe ROC et la validation croisée. De plus, grâce au support de cours, j'ai pu approfondir mes connaissances en découvrant les raisonnements mathématiques qui constituent ces techniques.
Ainsi, ayant dans l'optique de me tourner vers le métier de data scientist, ce cours m'a apporté des connaissances nécessaires en apprentissage supervisé pour la suite de mon parcours. Ainsi que la méthodologie d'un data scientist doit avoir en suivant le TP de pic d'ozone.