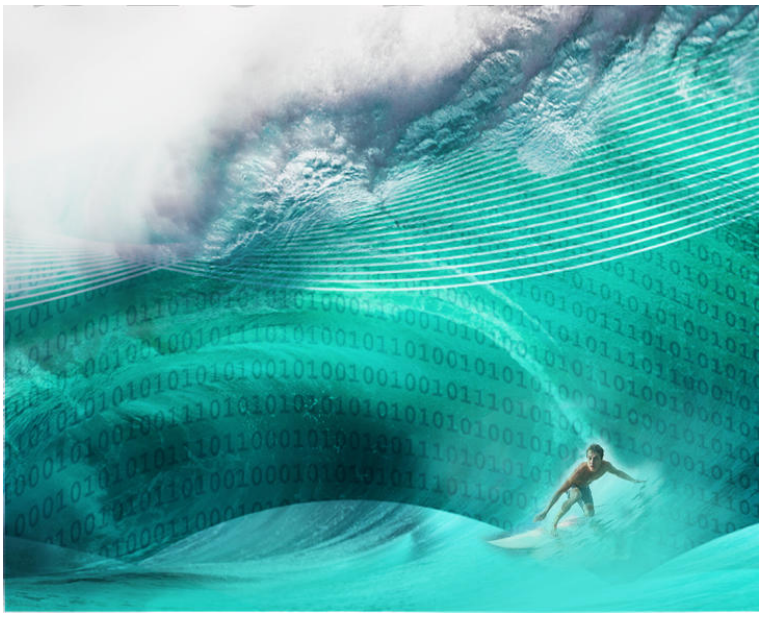
Certificat Science des Données – Module de Sensibilisation

Introduction à l'Apprentissage Statistique

Philippe Besse & Béatrice Laurent

Université de Toulouse - INSA



1 Introduction

La partie précédente aborde les méthodes de visualisation et d'exploration des données afin de comprendre leurs structures et donc leurs propriétés ou caractéristiques. Concernant les données d'identification d'une activité humaine, cette étape montre les bonnes propriétés des données transformées selon des méthodes et approches classiques en traitement du signal alors que les signaux bruts semblent très confus. L'objectif d'identification semble donc bien atteignable sur les données transformées. Pour ce faire il est nécessaire d'introduire quelques méthodes de modélisation ou apprentissage automatique. Cette partie met l'accent sur celles issues des modèles de régression classiques en Statistique. Ils ne sont pas utilisés dans un objectif explicatif visant à montrer l'impact d'un facteur mais à seule fin de prévision d'une variable Y quantitative, dans un objectif dit de régression, ou Y qualitative dans un objectif de classification supervisée. Après avoir introduit les grands principes de l'apprentissage statistique, notamment pour la sélection de modèle, ceux-ci sont illustrés par les exemples de régression linéaire multiple et de régression logistique pour la classification binaire. Ils sont appliqués à la prévision de la concentration en ozone, problème de régression, à celle du dépassement du seuil d'ozone, problème de classification binaire et enfin à la reconnaissance de la classe d'activité à partir des transformations des signaux d'un smartphone. Les bons résultats obtenus dans ce dernier cas viennent confirmer l'intuition de la phase exploratoire.

2 Aborder les Principes de l'Apprentissage Statistique

2.1 Modélisation et prévision

Dans les années 30 et notamment à la suite des travaux de Ronald Fisher, la Statistique a été développée avec une finalité principalement explicative pour un objectif d'aide à la décision. Par exemple : tester l'efficacité d'une molécule et donc d'un médicament, comparer le rendement de semences ou optimiser le choix d'un engrais, montrer l'influence d'un facteur (consommation de tabac, de sucre) sur des objectifs de santé publique. La prise de décision est alors soumise à un *test statistique* permettant de contrôler le risque d'erreur encouru. Mais il se trouve que les mêmes modèles statistiques peuvent aussi être utilisés avec une finalité seulement *prédictive* : prévoir la concentration en ozone du lendemain, le risque de défaut de paiement d'une entreprise... Ils sont les premiers algorithmes d'apprentissage statistique, encore très utilisés et les rares transparents, c'est-à-dire suffisamment simples et explicites pour conduire à une interprétation sur la façon dont la prévision fonctionne.

Deux modèles statistiques sont considérés :

1. La régression linéaire multiple qui vise à modéliser, prévoir, une variable quantitative Y (cible ou dépendante) à l'aide d'un ensemble de variables (*features* ou caractéristiques) quantitatives et/ou qualitatives $X^j_{j=1\ldots}p$.

2. La régression logistique qui est une adaptation de la précédente afin de prévoir l'occurrence (défaut, succès, maladie...) d'une variable qualitative binaire Y ou plutôt, en premier lieu, la probabilité de cette occurrence.

Commençons par introduire le modèle de régression linéaire. Ayant observé les valeurs $(y_i, \boldsymbol{x}_i = x_i^j; j = 1, \ldots, p)$ des variables sur un ensemble d'apprentissage $i = 1 \ldots, n$ d'objets individus ou instances, la modélisation par régression linéaire consiste à estimer au mieux les paramètres d'un modèle :

$$y_i = b_0 + b_1 x_1^i + \cdots + b_p x_i^p + \varepsilon_i, \quad \text{ou matriciellement} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon},$$

où les ε_i sont des termes d'erreur.

L ?estimation des paramètres par minimisation des moindres revint à minimiser la moyenne des carrés (MSE ou *mean square error*). La prévision de Y pour un nouvel individu $\boldsymbol{x}_0$ est alors obtenue en appliquant le modèle :

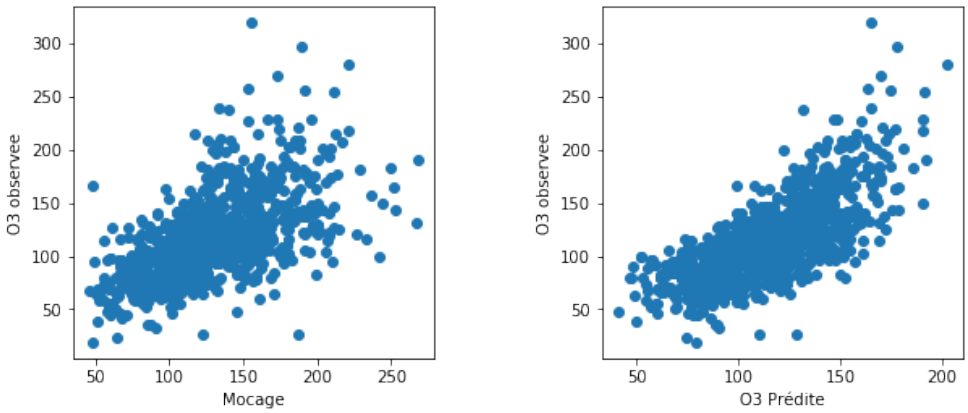
$$\hat{y}_0 = \hat{f}(\boldsymbol{x}_0).$$

Moyennant un ensemble d'hypothèses sur la loi des résidus : normalité, homoscédasticité, l'estimation du modèle peut être définie par un principe de maximisation de la log-vraisemblance. C'est en fait la même solution que celle issue des moindres carrés mais cette approche confère de meilleures propriétés au modèle. Ces hypothèses rendent possibles des tests et prévisions par *intervalle de confiance* afin de contrôler le risque d'erreur. Ce point n'est pas développé ; se reporter à l'abondante littérature sur ce sujet.

Avant d'aborder le problème plus complexe de reconnaissance d'une activité humaine à partir des signaux enregistrés par un smartphone, nous nous intéressons aux données de prévision de la concentration en ozone vues dans la partie 2 précédente et détaillées dans le tutoriel `ML4IoT-Tutorial-Ozone`. L'objectif est d'illustrer les modèles statistiques de régression. Cet exemple, à la fois de régression et de discrimination binaire, présente des vertus pédagogiques certaines qui permettent de l'utiliser comme fil rouge de comparaison entre toutes méthodes.

Le graphique ci-dessous utilise ces données pour comparer deux ajustements

de modèle.



Celui obtenu par le modèle déterministe MOCAGE à gauche et celui à droite obtenu par adaptation statistique en introduisant les autres variables explicatives. Le modèle (choix de variables) a été optimisé comme cela est expliqué par la suite. Les graphes représentent les valeurs observées y_i en fonction des valeurs prédites $\hat{y}_i$. Plus celles-ci sont proches de la diagonale et meilleur est l'ajustement du modèle (plus petits résidus). Le carré R^2 du coefficient de corrélation entre $\boldsymbol{y}$ et $\hat{\boldsymbol{y}}_i$ est appelé coefficient de détermination. Il vaut $0,10$ pour Mocage et $0,52$ pour le modèle statistique intégrant un choix optimal de variables.

2.2 Erreur de prévision

Il est fondamental de ne pas confondre l'*erreur d'estimation* (*MSE*) minimisée lors de l'ajustement du modèle et l'*erreur de prévision* du modèle ou *risque*, *erreur de généralisation* qui doit nécessairement être estimée sur un échantillon indépendant. En effet, l'erreur d'ajustement, encore appelée erreur apparente, est estimée sur des observations qui participent à l'ajustement du modèle. Par principe, elle est une version *optimiste* de l'erreur de prévision qui concerne des observations nouvelles qui n'ont pas participé à l'ajustement du modèle. Les écarts ou résidus de la prévision sont en effet, sauf cas particuliers, de nature à être plus grands que les résidus minimisés lors de l'estimation.

Une façon élémentaire d'estimer une erreur de prévision consiste à opérer en deux phases après avoir séparé aléatoirement l'échantillon D_n en deux parties, celle $D_{n_a}^{\text{Appr}}$ d'apprentissage pour estimer un modèle ou entraîner un algorithme et celle $D_{n_t}^{\text{Test}}$ de test avec $n = n_a + n_t$.

Une estimation de l'erreur de prévision ou risque : *RMSE* ou *risk mean square error*, est alors fournie par :

$$RMSE = \frac{1}{n_t} \sum_{\boldsymbol{x}_i \in D_{n_t}^{\text{Test}}} (\hat{f}(\boldsymbol{x_i}) - y_i)^2.$$

L'*objectif majeur* d'une étape d'*apprentissage statistique* est de déterminer le modèle ou l'algorithme, parmi tous ceux disponibles ou parmi une classe réduite de ceux interprétables, qui conduit à la plus *petite erreur de prévision* ou plus petit risque.

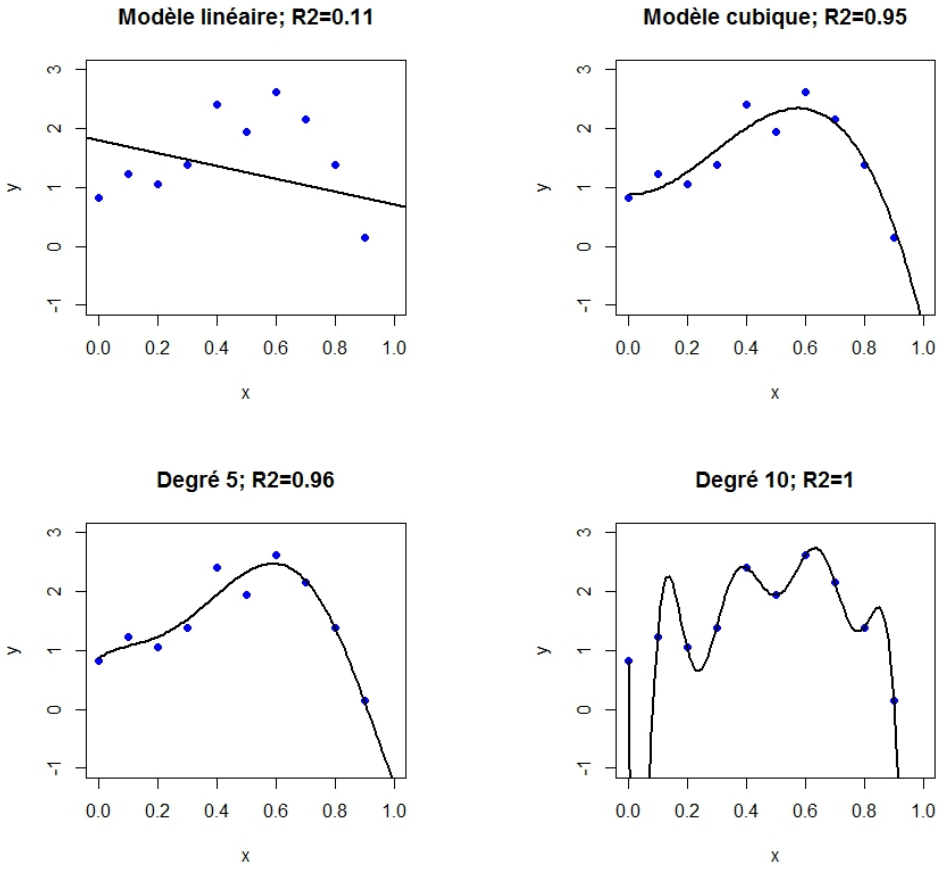
2.3 Sélection de Modèle

La recherche du modèle statistique ou de l'algorithme d'apprentissage optimal nécessite de prendre conscience d'une réalité importante :

> Le modèle ou l'algorithme qui *ajuste le mieux* les données d'apprentissage n'est pas nécessairement celui qui conduit à la *meilleure prévision*.

Rechercher le meilleur modèle parmi une famille donnée consiste, d'un point de vue statistique, à réaliser un *meilleur compromis entre biais et variance*. Ceci peut être illustré simplement dans le cas de la régression polynomiale.

Le graphique ci-dessous compare les ajustements d'un nuage de points par régression polynomiale. Les points sont successivement ajustés par des polynômes de degré 1, 3, 5 et 10.



Les coefficients de détermination qui mesure la qualité de l'ajustement sont successivement : $R^2 = 0.11$ (degré 1) , $R^2 = 0.95$ (degré 3), $R^2 = 0.95$ (degré 5), $R^2 = 1$: degré 10 ou interpolation exacte des points.

Clairement, la qualité de l'ajustement du modèle croît avec le R^2 et donc avec la complexité, nombre de paramètres ou degré du polynôme. Intuitivement, le nuage de points a la forme d'un modèle relativement simple mais dont les observations sont entachées d'erreurs. Vouloir ajuster au mieux ces observations conduit à interpoler la composante d'erreur ou de bruit au détriment de la régularité du modèle. La conséquence directe en est la construction de prévisions qui prendront des valeurs absolues beaucoup trop importantes par rapport à celles généralement observées. La *variance des erreurs* de prévision prend alors une très grande valeur qui conduit à l'explosion du *RMSE*. En revanche, accepter un modèle, éventuellement plus simple ou *biaisé*, que le supposé vrai modèle, évite d'ajuster le modèle aux erreurs de mesure et réduit la *variance*. C'est le principe de la recherche d'un meilleur compromis biais / variance ; compromis qui dépend de la variance du bruit par rapport à celle des observations (rapport

signal / bruit).

La stratégie de choix de modèle consiste donc à la recherche d'un modèle *parcimonieux* (*sparse*) au sens où sa complexité optimise le compromis entre biais et variance. C'est facilement illustré dans le cas de la régression mais ce principe se retrouve dans presque tous les algorithmes d'apprentissage dont il faut contrôler la complexité et donc la flexibilité de l'ajustement aux données. En fonction du type d'algorithme ce peut être le nombre de variables, de feuilles, de voisins, de neurones ou un coefficient de pénalisation : *ridge*, Lasso (*least absolute shrinkage and selection operator*)... qui règle la flexibilité ou complexité du modèle.

2.4 Sélection par pénalisation

Il existe de très nombreuses stratégies (descendante, ascendante, pas-à-pas, par pénalisation *ridge* ou Lasso...) de sélection de modèle en régression, stratégies basées sur une grande variété de critères : test de Fisher, critère d'Akaïke, critère bayésien (BIC), C_p de Mallows. La stratégie actuellement la plus utilisée est celle proposée par Tibshirani (1996) basée sur une pénalisation Lasso dite encore en norme L_1 de la somme des valeurs absolues des paramètres. Ce qui s'écrit encore :

$$\boldsymbol{b}_{\text{Lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left(\sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} x_i^j \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right).$$

qui est équivalent à minimiser les moindres carrés (*MSE*) sous une contrainte :

$$\boldsymbol{b}_{\text{Lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left(\sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} x_i^j \beta_j)^2 \right) \quad \text{avec} \quad \sum_{j=1}^{p} |\beta_j| < r.$$

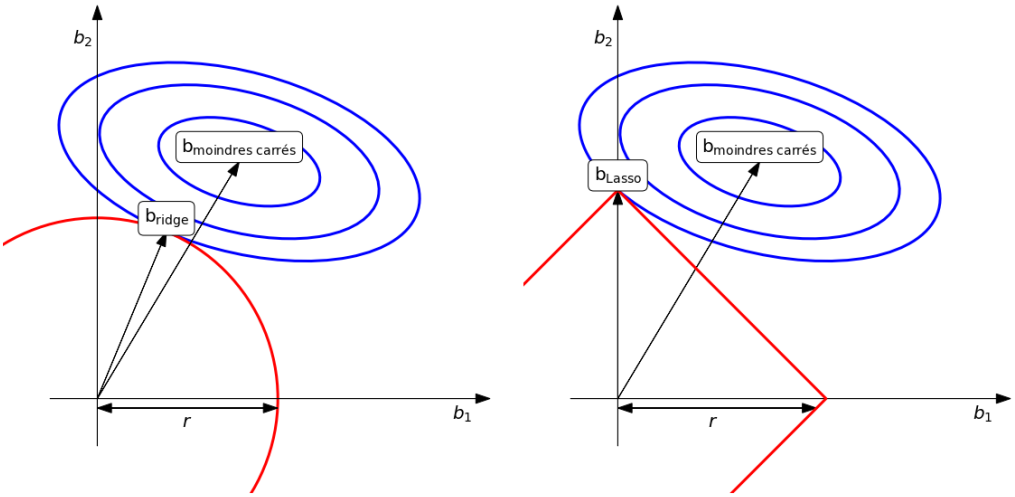
La régression *ridge* satisfait à la même formule :

$$\boldsymbol{b}_{\text{Ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left(\sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} x_i^j \beta_j)^2 \right) \quad \text{avec} \quad \sum_{j=1}^{p} \beta_j^2 < r^2.$$

en remplaçant la norme L_1 par la norme L_2, c'est-à-dire la somme des valeurs absolues par la somme des carrés des coefficients. La régression *ridge* suit le même principe que la régularisation de Tikhonov pour résoudre un problème inverse indéterminé.

Le graphique ci-dessous propose une interprétation géométrique des pénalisations *ridge* et Lasso afin d'expliquer en quoi cette dernière opère automatiquement une sélection de variables par annulation des paramètres b_j les moins

significatifs.

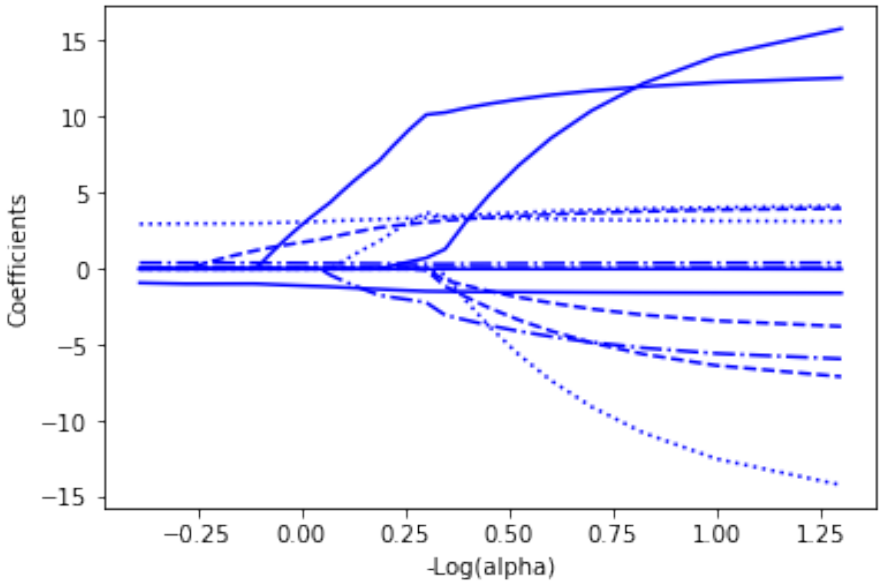


Dans les deux graphiques, les ellipses représentent les valeurs du critère des moindres carrés. Lorsque la pénalisation est nulle ($\lambda = 0$), les paramètres b_j prennent les valeurs des moindres carrés minimaux. En norme L_2, la contrainte est visualisée par une hypersphère, un cercle en dimension 2, de rayon r. La régression *ridge* conduit alors à un *rétrécissement* (*shrinkage*) à la rencontre de l'ellipse des moindres carrés et de la sphère de la contrainte. La pénalisation Lasso remplace l'hypersphère par un hypercube, un carré en dimension 2. Dans ce cas l'ellipse à toutes les chances de rencontrer un coin en dimension 2 ou un hyperplan de dimension $n - 1$ correspondant à l'annulation de certains paramètres, ici $b_1 = 0$.

Cette stratégie de sélection de modèles peut paraître complexe mais elle est largement utilisée, notamment dans la librairie *Scikit-learn* de Python. Seule insuffisance de cette librairie, la difficulté, voire l'impossibilité, de pouvoir prendre en compte simplement des interactions entre les variables alors que la syntaxe de R le permet aisément.

Le graphe ci-dessous représente le *chemin de régularisation* en pénalisation

Lasso lors de la recherche d'un modèle optimal.



Il montre comment les paramètres décroissent et certains s'annulent lorsque la pénalisation croît de droite à gauche des abscisses associées aux valeurs d'un paramètre α qui joue un rôle similaire à λ.

Un troisième type de sélection de modèle dit *elastic net* consiste à associer pénalisation *ridge* et Lasso mais nécessite le réglage de deux paramètres.

2.5 Optimisation par Validation Croisée V-*fold*

La question alors soulevée est de savoir comment *optimiser* la valeur de ce paramètre de pénalisation. La stratégie unanimement employée consiste à minimiser un risque estimé par une procédure de *validation croisée*. L'idée est d'itérer l'estimation sans biais de l'erreur sur plusieurs échantillons dits de *validation* n'ayant pas été utilisés pour l'estimation du modèle ou l'entraînement de l'algorithme puis d'en calculer la moyenne. L'objectif est ainsi d'éviter un biais tout en réduisant la variance (par moyennage) de l'estimation pour améliorer la précision lorsque la taille de l'échantillon initial est réduite. Il existe plusieurs versions de cet algorithme, celle décrite succinctement ci-dessous est la validation croisée V-*fold*, la plus habituelle.

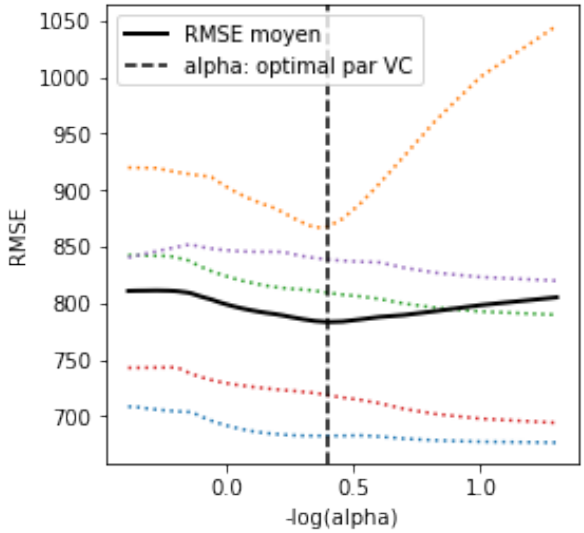
Plus précisément, l'échantillon d'apprentissage est découpé aléatoirement en V groupes (par défaut 5 ou 10) de tailles similaires qui jouent à tour de rôle celui de validation. Le i-ème groupe étant mis de côté, les $(V-1)$ autres groupes constituent l'échantillon d'apprentissage servant à estimer le modèle. L'erreur de prévision ou risque $RMSE_i$ est estimée sur le i-ème groupe. Les V erreurs

$RMSE_i; i = 1, \ldots, V$ ainsi obtenues sont moyennées pour calculer $RMSE_{\text{CV}}$, l'estimation par validation croisée *V fold* du risque ou erreur de prévision.

Minimiser le risque estimé par validation croisée est une approche largement utilisée pour optimiser le choix d'un modèle au sein d'une famille paramétrée par θ où θ est le nombre de variables, de neurones, de feuilles, une pénalisation... Dans ce cas, $\widehat{f}(\theta_{\text{Opt}})$ est défini par

$$\theta_{\text{Opt}} = \arg\min_{\theta} RMSE_{\text{CV}}(\theta).$$

Le graphe ci-dessous illustre l'optimisation du paramètre de pénalisation Lasso par validation croisée 5-*fold*.



Chaque ligne colorée représente l'évolution de l'erreur de prévision calculée sur l'un des échantillons de validation en fonction d'une pénalisation α, le paramètre de la librairie `scikit-learn`. La ligne pleine est la moyenne qui admet un minimum auquel est associé la valeur optimale du paramètre de pénalisation.

3 Comprendre la Classification Supervisée

La section précédente développe la prévision d'une variable cible quantitative Y appliquée par exemple à la prévision de la concentration en ozone. En revanche, le problème de reconnaissance de l'activité humaine vise à la prévision d'une classe d'activité ou encore de la classe d'une variable Y cette fois qualitative. Il s'agit d'un problème de *classification supervisée* ou reconnaissance de forme. *Attention*, ne pas confondre avec l'objectif de classification *non* supervisée ou en anglais *clustering* qui vise à trouver des groupes homogènes dans des

individus ou instances caractérisées par p variables. Contrairement à la classification supervisée, il n'y a pas de variable qualitative cible a priori connue et observée sur l'échantillon.

Historiquement, les deux méthodes partageant l'objectif de discrimination ou de prévision d'une variables quantitative Y et apparues en premier sont

— l'analyse discriminante décisionnelle,

— la régression logistique ou binomiale.

L'analyse discriminante développée par Fisher propose la prévision d'une classe parmi les m modalités de Y alors qu'en principe la régression logistique est adaptée à une variable Y à deux classes ou binaire : succès ou échec, présence d'une maladie, défaut de paiement, occurrence d'un événement... Néanmoins, cette même méthode, à condition que la taille de l'échantillon le permette, est largement utilisée pour de la discrimination en $m > 2$ classes en construisant m modèles d'une classe contre les autres. Pour la prévision de la classe d'un nouvel individu ou d'une nouvelle instance, les m modèles fournissent m probabilités d'occurrence de chaque classe et c'est la classe de probabilité maximale qui l'emporte. Cette stratégie est prise par défaut dans la librairie `scikit-learn` lors de l'utilisation de la régression logistique avec $m > 2$.

Après ces méthodes historiques, bien d'autres modèles ou algorithmes on été proposés avec le même objectif de prévision d'une classe ou modalité d'une variable qualitative binaire ou à m classes : k plus proches voisins, arbre binaire de décision, réseau de neurones (perceptron), machine à vecteur support (SVM) ainsi que les algorithmes d'agrégation d'arbres : *boosting, random forest...* Se reporter aux tutoriels du dépôt `github.com/wikistat` pour les expérimenter.

3.1 Classification binaire par régression logistique

La section précédente explique comment ajuster une variable Y quantitative, à valeurs dans $\mathbb{R}$, par une combinaison linéaire des p variables explicative X^j. Ce que la régression logistique vise à modéliser est la probabilité d'occurrence (succès, maladie...) d'une classe de Y qui est à valeur dans l'intervalle $[0, 1]$.

Plus précisément, le choix d'une *fonction lien* permet de faire correspondre les deux domaines de variation $[0, 1]$ et $\mathbb{R}$ afin de relier une probabilité avec un prédicteur linéaire classique $\boldsymbol{Xb}$. Si π_i désigne la probabilité de la classe 1 de Y ou $P(y_i = 1|\boldsymbol{x}_i)$, la fonction lien dite *canonique* couramment utilisée est la fonction *logistique* et le modèle s'écrit :

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} = b_0 + b_1 x_i^1 + \cdots + b_p x_i^p \quad \text{ou} \quad \hat{\pi}_i = \frac{\exp \boldsymbol{x}_i' \boldsymbol{b}}{1 + \exp \boldsymbol{x}_i' \boldsymbol{b}}.$$

L'estimation des paramètres b_j de ce modèle est obtenue par l'exécution d'un algorithme de maximisation (*e.g.* Newton Raphson) de la log-vraisemblance du modèle.

La prévision d'une probabilité $\hat{\pi}_0$ de $\boldsymbol{x}_0$ est fournie par :

$$\hat{\pi}_0 = \frac{\exp \boldsymbol{x}_0' \boldsymbol{b}}{1 + \exp \boldsymbol{x}_0' \boldsymbol{b}}.$$

La prévision de la classe de $\boldsymbol{x}_0$ est obtenue en comparant cette probabilité avec une valeur seuil ou *cut-off*, par défaut $0,5$; $\hat{y}_i = 1$ si $\hat{\pi}_0 > \frac{1}{2}$ et $\hat{y}_i = 0$ sinon.

Moyennant des hypothèses sur la loi de Y (binomiale), la planification de l'expérience et la répartition de l'échantillon, des procédures de test et d'estimation par intervalle de confiance des prévisions sont accessibles. Consulter la bibliographie à ce sujet. Nous nous limitons ici au seul objectif de prévision.

3.2 Courbe ROC

Matrice de confusion

Une erreur quadratique moyenne (*RMSE*) est généralement utilisée pour évaluer une erreur de prévision ou risque en régression. Ce critère n'est pas adapté au cas de la classification supervisée. Il est souvent remplacé par un simple taux d'erreur calculé à partir de la *matrice de confusion*. Cette matrice est simplement une table de contingence ou tableau obtenu par le croisement des deux variables : classe observée *vs.* classe prédite.

Dans le cas fréquent de la discrimination de deux classes, la plupart des méthodes (*e.g* régression logistique) estiment, pour chaque individu i, un *score* ou une probabilité $\widehat{\pi}_i$ que cet individu prenne la modalité $Y = 1$. Cette probabilité comprise entre 0 et 1 est comparée avec une valeur seuil s fixée *a priori*, par défaut $0,5$:

$$\text{Si} \quad \widehat{\pi}_i > s, \ \widehat{y}_i = 1 \quad \text{sinon} \quad \widehat{y}_i = 0.$$

Pour un échantillon de taille n dont l'observation de Y est connue ainsi que les scores $\widehat{\pi}_i$ fournis par un modèle ; la matrice de *confusion* associée à cette valeur de seuil s est :

Prévision	Observation		Total
	$Y = 1$	$Y = 0$	
$\widehat{y}_i = 1$	$n_{11}(s)$	$n_{10}(s)$	$n_{1+}(s)$
$\widehat{y}_i = 0$	$n_{01}(s)$	$n_{00}(s)$	$n_{0+}(s)$
Total	n_{+1}	n_{+0}	n

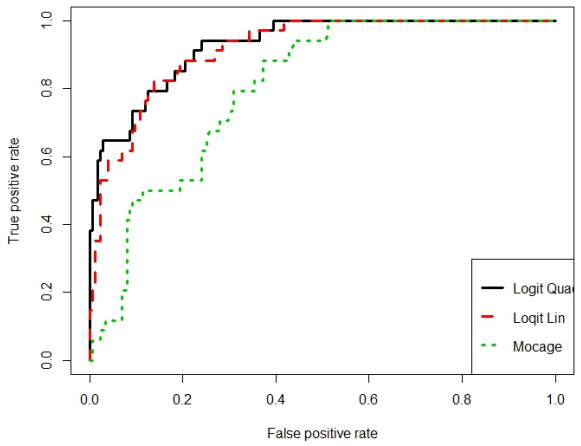
Les quantités suivantes sont considérées :

- Vrais positifs : les $n_{11}(s)$ observations bien classées ($\widehat{y}_i = 1$ et $Y = 1$),
- Vrais négatifs : les $n_{00}(s)$ observations bien classées ($\widehat{y}_i = 0$ et $Y = 0$),
- Faux négatifs : les $n_{01}(s)$ observations mal classées ($\widehat{y}_i = 0$ et $Y = 1$),
- Faux positifs : les $n_{10}(s)$ observations mal classées ($\widehat{y}_i = 1$ et $Y = 0$),
- Le taux d'erreur : $t(s) = \frac{n_{01}(s)+n_{10}(s)}{n}$,
- Le taux de vrais positifs ou *sensibilité* $= \frac{n_{11}(s)}{n_{+1}}$ ou taux de positifs pour les individus qui le sont effectivement,
- Le taux de vrais négatifs ou *spécificité* $= \frac{n_{00}(s)}{n_{+0}}$ ou taux de négatifs pour les individus qui le sont effectivement,
- Le taux de faux positifs $= 1 -$ Spécificité $= 1 - \frac{n_{00}(s)}{n_{+0}} = \frac{n_{10}(s)}{n_{+0}}$.

Courbe ROC et AUC

Les notions de *spécificité* et de *sensibilité* proviennent de la théorie du signal ; leurs valeurs dépendent directement de celle du seuil s. En augmentant s, la sensibilité diminue tandis que la spécificité augmente car la règle de décision devient plus exigeante. Un bon modèle associe grande sensibilité et grande spécificité pour la détection d'un signal. Ce lien est représenté graphiquement par la courbe ROC (*Receiver Operating Caracteristic*) de la sensibilité (probabilité de détecter un vrai signal) en fonction de un moins la spécificité (probabilité de détecter un signal à tort) pour chaque valeur s du seuil.

On montre qu'une courbe ROC est croissante monotone. Plus une courbe de la figure ci-dessous se rapproche du carré, meilleure est la discrimination, correspondant à la fois à une forte sensibilité et une grande spécificité. L'aire sous la courbe : AUC (*area under curve*) mesure la qualité de discrimination du modèle tandis qu'une analyse de la courbe aide au choix du seuil.



Ce graphique compare trois courbes ROC. Celle issue du modèle MOCAGE en vert avec celles issues de deux modèles de régression logistique, l'un linéaire, l'autre quadratique car faisant intervenir des interactions. Pour comparer des modèles ou méthodes de complexités différentes, ces courbes doivent être estimées sur un échantillon test. Elles sont bien évidemment optimistes sur l'échantillon d'apprentissage. De plus, l'AUC ne définit pas un ordre total entre modèles car les courbes ROC peuvent se croiser.

Ces résultats montrent encore plus clairement l'intérêt de l'adaptation statistique de la prévision MOCAGE mais aussi la difficulté de la décision qui découle de la courbe ROC. Le choix du seuil, et donc de la méthode à utiliser si les courbes se croisent, dépend d'un choix dans ce cas politique : quel est le taux de faux positifs acceptable d'un point de vue économique ou le taux de vrais positifs à atteindre pour des raisons de santé publique ? Le problème majeur est de pouvoir quantifier les coûts afférents, par la définition d'une matrice

dissymétrique de ces coûts de mauvais classement en vue d'optimiser le choix de s.

Autre critère pour la discrimination à deux classes

Une autre difficulté concerne les cas où les classes sont déséquilibrées ; ainsi, les jours de dépassement du seuil critique de concentration en ozone sont relativement rares. Un modèle qui ne prédit pas ou presque pas de dépassement à un taux d'erreur au plus égal au ratio du nombre de jours de dépassement. En ce sens, le taux d'erreur de classement n'est pas toujours adapté à une situation de classes très déséquilibrées.

D'autres critères ont été proposés pour intégrer cette difficulté dont le *Score de Pierce* basés sur le taux de bonnes prévisions : $H = \frac{n_{11}(s)}{n_{+1}(s)}$ et le taux de fausses alertes : $F = \frac{n_{10}(s)}{n_{+0}}$. Le score de Pierce est alors défini par PSS$= H - F$ et est compris entre -1 et 1. Il évalue la qualité de la prévision. Si ce score est supérieur à 0, le taux de bonnes prévisions est supérieur à celui des fausses alertes et plus il est proche de 1, meilleur est le modèle.

Le score de Pierce a été conçu pour la prévision d'événements climatiques rares afin de pénaliser les modèles ne prévoyant jamais ces événements ($H = 0$) ou encore générant trop de fausses alertes ($F = 1$). Le modèle idéal prévoyant tous les événements critiques ($H = 1$) sans fausse alerte ($F = 0$). Une autre stratégie consiste à introduire des coûts de mauvais classement pour pondérer un score.

4 Algorithme des k plus poches voisins

De très nombreuses méthodes et variantes d'apprentissage supervisé sont proposés dans la littérature. Le site `github/wikistat` en propose un tour d'horizon tandis que le livre de Hastie et al. 2009 reste une référence bibliographique. En complément des méthodes statistiques linéaires classiques et toujours très utilisées, introduisons un algorithme non linéaire lui aussi très utilisés mais qui souffre, contrairement aux méthodes linéaires, d'un défaut rédhibitoire ; il ne passe pas à l'échelle des de données volumineuses distribuées.

L'algorithme des k plus proches voisins peut être présentés de différentes façons notamment comme un cas particulier d'analyse discriminante décisionnelle (ADD).

4.1 Analyse discriminante décisionnelle linéaire

Plaçons nous dans le cadre de l'analyse factorielle discriminante.

Une variable qualitative Y à m modalités et p variables quantitatives $X^j, j = 1, \ldots, p$ sont observées sur un même échantillon de taille n. L'objectif de l'analyse discriminante décisionnelle déborde le simple cadre descriptif de l'analyse factorielle discriminante (AFD). Disposant d'individus sur lesquels les X^j sont observées mais pas Y, il s'agit de *décider* de la modalité $\mathcal{T}_\ell$ de Y (ou de la

classe correspondante) de ces individus. L'ADD s'applique donc également à la situation précédente de la régression logistique ($m = 2$) mais aussi lorsque le nombre de classes est plus grand que 2. Les variables explicatives devant être quantitatives, celles qualitatives sont remplacées par des indicatrices.

L'objectif est de définir des *règles de décision* (ou d'affectation) ; $\boldsymbol{x} = (x^1, \ldots, x^p)$ désigne les observations des variables explicatives sur un individu, $\{\boldsymbol{g}_\ell; \ell = 1, \ldots, m\}$ les barycentres des classes calculés sur l'échantillon et $\bar{\boldsymbol{x}}$ le barycentre global.

La matrice de covariance empirique se décompose en

$$\boldsymbol{S} = \boldsymbol{S}_e + \boldsymbol{S}_r.$$

où $\boldsymbol{S}_r$ est appelée variance intraclasse (within) ou résiduelle :

$$\boldsymbol{S}_r = \bar{\boldsymbol{X}}_r{}' \boldsymbol{D} \bar{\boldsymbol{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\boldsymbol{x}_i - \boldsymbol{g}_\ell)(\boldsymbol{x}_i - \boldsymbol{g}_\ell)',$$

et $\boldsymbol{S}_e$ la variance interclasse (between) ou expliquée :

$$\boldsymbol{S}_e = \overline{\boldsymbol{G}}' \boldsymbol{D} \overline{\boldsymbol{G}} = \bar{X}_e' \boldsymbol{D} \bar{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\boldsymbol{g}_\ell - \overline{\boldsymbol{x}})(\boldsymbol{g}_\ell - \overline{\boldsymbol{x}})'.$$

Règle de décision issue de l'AFD

L'individu x est affectée à la modalité de Y minimisant :

$$d^2_{\boldsymbol{S}_r^{-1}}(\boldsymbol{x}, \boldsymbol{g}_\ell), \ell = 1, \ldots, m,$$

c'est-à-dire à la classe dont le barycentre est le plus proche au sens de la métrique de Mahalabobis.

Cette distance se décompose en

$$d^2_{\boldsymbol{S}_r^{-1}}(\boldsymbol{x}, \boldsymbol{g}_\ell) = ||\boldsymbol{x} - \boldsymbol{g}_\ell||^2_{\boldsymbol{S}_r^{-1}} = (\boldsymbol{x} - \boldsymbol{g}_\ell)' \boldsymbol{S}_r^{-1} (\boldsymbol{x} - \boldsymbol{g}_\ell)$$

et le problème revient donc à maximiser

$$\boldsymbol{g}_\ell' \boldsymbol{S}_r^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{g}_\ell' \boldsymbol{S}_r^{-1} \boldsymbol{g}_\ell.$$

Il s'agit d'une règle linéaire en $\boldsymbol{x}$ car elle peut s'écrire : $\boldsymbol{A}_\ell \boldsymbol{x} + \boldsymbol{b}_\ell$. Ainsi, cette méthode construit des *hyperplans* dans $\mathbb{R}^p$ qui définissent les frontières des classes.

Généralisation

Il n'est pas très difficile de montrer que la situation précédente est le cas particulier d'une approche plus générale visant à chercher la classe d'un nouvel individu en minimisant un risque bayésien, c'est-à-dire encore la classe qui maximise une probabilité *a postériori*. Les probabilités sont calculées en estimant les densités de la loi multidimensionnelle des variables X^j conditionnellement aux différentes classes de Y.

On montre alors, que sous les hypothèses :

— la loi des X^j est gaussienne multidimensionnelle,
— les lois conditionnelles aux classes partagent la même matrice de variance covariance (intra-classe),
— les coûts de mauvais classement sont identiques,
— les probabilités *a priori* d'appartenance aux classes sont identiques ;

minimiser le risque bayésien ou maximiser la probabilité *a postériori* revient à appliquer la règle linéaire précédente et issue de l'AFD.

Remarque En relaxant l'hypothèse d'homoscédasticité d'égalité des matrices de covariarance des lois conditionnelles, une autre analyse discriminante est définie en estimant, pour chaque classe, la matrice de covariance. Beaucoup plus de paramètres sont alors estimés, nécessitant des données plus volumineuses. On montre que la règle devient quadratique, les frontières classes sont définies par des surfaces coniques dans $\mathbb{R}^2$.

4.2 Analyse discriminante décisionnelle non paramétrique

Si l'hypothèse gaussienne n'est pas raisonnable, il est nécessaire de mettre en œuvre des estimations non paramétriques des lois conditionnelles. Plusieurs pistes sont possibles mais la plus simples et la plus utilisées consiste à opérer des estimations locales des densités conditionnelles. C'est l'objectif de l'algorithme des k plus proches voisins.

Cette méthode d'affectation d'un vecteur $\boldsymbol{x}$ consiste à enchaîner les étapes décrites dans l'algorithme ci-dessous.

Algorithme des k plus proches voisins (k-nn)

1. Choix d'un entier $k : 1 \geq k \geq n$.
2. Calculer les distances $d_{\boldsymbol{M}}(\boldsymbol{x}, \boldsymbol{x}_i)\,, \quad i = 1, \ldots, n$ où $\boldsymbol{M}$ est la métrique de Mahalanobis c'est-à-dire la matrice inverse de la matrice de variance (ou de variance intraclasse).
3. Retenir les k observations $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(k)}$ pour lesquelles ces distances sont les plus petites.
4. Compter les nombres de fois $k_1, \ldots, k_m$ que ces k observations apparaissent dans chacune des classes.
5. Estimer localement les densités conditionnelles par

$$\widehat{h_\ell}(\boldsymbol{x}) = \frac{k_\ell}{k}.$$

Pour $k = 1$, $\boldsymbol{x}$ est affecté à la classe du plus proche élément. Comme toute technique d'apprentissage, l'algorithme des k plus proches voisins nécessitent le réglage d'un paramètre de complexité pris en charge par k. Ce choix s'apparente à un choix de modèle et nécessite le même type d'approche à savoir l'optimisation d'un critère comme l'erreur de prévision estimée par validation croisée. La flexibilité ou la complexité des frontières entre les classes est réglée par la valeurs de k.

Représenter graphiquement le rôle de k en exécutant le tutoriel `CSdD-Intro-Apprent-Python`

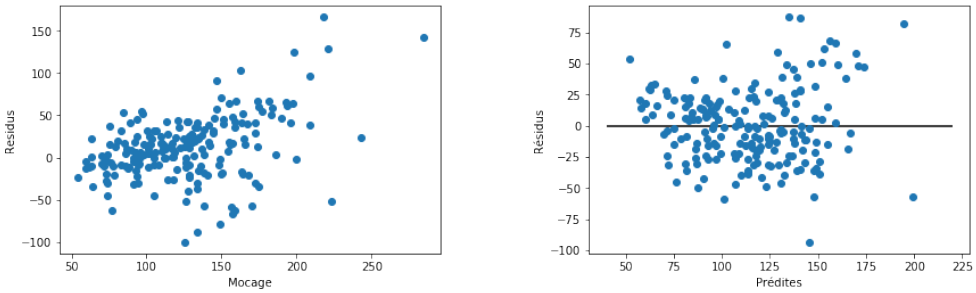
Si y est quantitative, l'algorithme des k plus proches voisins est aussi un algorithme de régression en prenant la valeur moyenne des k voisins comme prévision de $\boldsymbol{x}$.

5 Modéliser des données complexes

Les différentes méthodes sont appliquées à la prévision de la concentration en ozone ou encore à la prévision de dépassement du pic puis à la reconnaissance de l'activité humaine.

5.1 Prévision de la concentration en ozone

Le modèle de régression ou plutôt la sélection des variables a été optimisée par pénalisation Lasso et optimisation du paramètre de pénalisation par validation croisée comme cela est expliqué dans une section précédente et opéré dans le tutoriel `CSdD-Pic-Ozone-Python`. La comparaison des modèles est obtenue en traçant le graphe des résidus de la prévision de l'échantillon test en fonction des valeurs prédites. Comme prévu, les résidus du modèle Mocage se dispersent nettement plus que ceux du modèle obtenu par adaptation statistique.

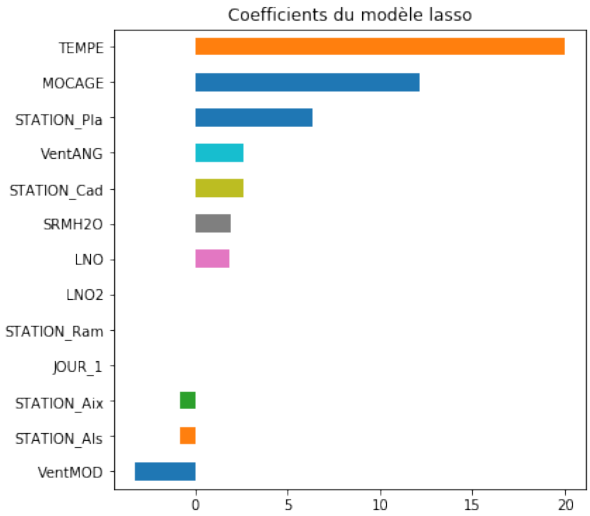


Il faut alors comparer les erreurs de prévisions ou risques estimés sur l'échantillon test : 1565 (Mocage) est à comparer avec 859, de même que les R^2 valant respectivement $0,10$ et $0,52$. Remarquons que l'optimisation du modèle par sélection Lasso de variable a amélioré le $RMSE$ et le R^2 correspondant. Ils valent respectivement 871 et 0,50 pour le modèle linéaire intégrant toutes les variables sans sélection.

Attention à la forme du nuage des résidus de la régression. La variance des résidus est plus importante pour les grandes valeurs de Y que pour les petites valeurs. Il n'y a pas homoscédasticité. En conséquence des prévisions par intervalle de confiance ne seraient pas fiables. De plus, les résidus ne se répartissent pas de façon symétrique de part et d'autre de l'axe $y = 0$. Cette forme de demi-lune ou "banane" révèle une insuffisance du modèle qui ne prend pas en compte une possible composante quadratique ou interaction entre les variables. Difficile

à estimer avec les possibilités offertes par la librairie `scikit-learn`, un modèle avec interactions optimisé dans R (calepin disponible) conduit à des résultats un peu meilleurs en terme de qualité de prévision mais au détriment de la simplicité de l'interprétation et du temps de calcul pour l'optimisation du modèle.

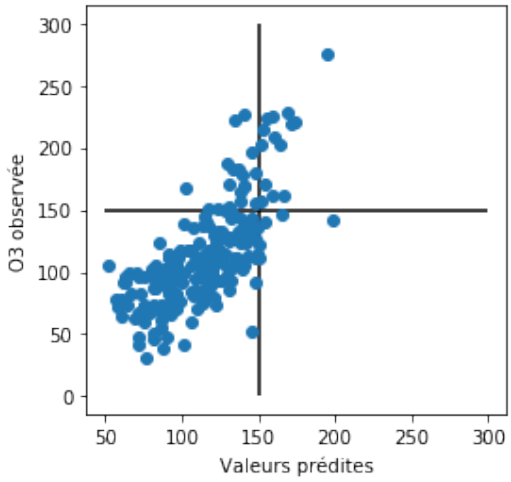
Il est en effet intéressant de se préoccuper des valeurs des paramètres du modèle afin d'évaluer l'importance des variables et de comprendre leur influence sur la concentration en ozone.



Ces paramètres montrent des différences géographiques entre les stations. La situation est plus critique à Plan-de-Cuques (banlieue nord de Marseille) qu'à Aix en Provence. Ils soulignent l'importance de la température dont l'influence locale est sans doute sous-estimée dans le modèle déterministe Mocage qui joue un rôle évidemment important dans la prévision. Un vent fort tend naturellement à réduire la concentration en ozone. *Attention*, un paramètre nul comme pour la concentration en dioxyde d'azote ne signifie pas que cette variable n'a pas d'influence sur celle de l'ozone. Cela signifie que l'influence, si elle existe, est déjà prise en compte par les autres variables du modèle et que l'ajouter dans le modèle accroîtrait la variance des prévisions et donc augmenterait le risque.

Une fois que la concentration en ozone est prévue, il est facile de voir si celle-ci dépasse le seuil légal dans le graphique ci-dessous associée à la matrice

de confusion ;



	Pas de dépassement observé	Dépassement observé
Pas de dépassement prédit	162	20
Dépassement prédit	5	13

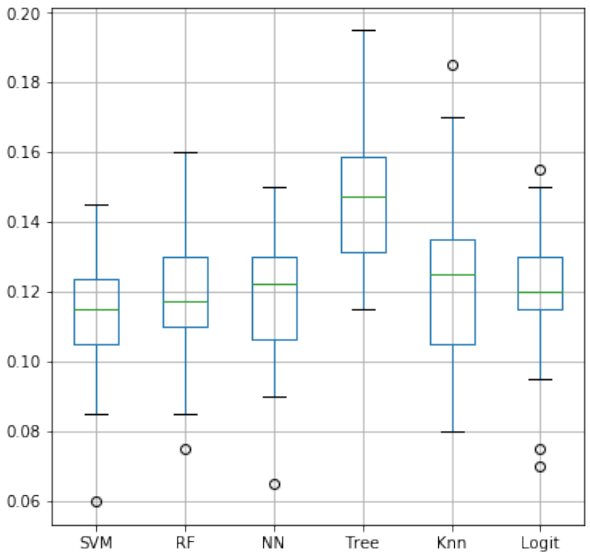
Remarquer la dissymétrie de la matrice de confusion à rapprocher de la forme du nuage des résidus commentée ci-dessus. Remarquer également le nombre relativement élevé de faux négatifs : pas de dépassement prédit alors qu'il a été observé au regard des vrais positifs. Calculer le score de Pierce à titre illustratif. Un modèle prenant en compte les interactions apporte une légère amélioration mais la taille de l'échantillon, notamment le nombre de jours de dépassements de seuil, est trop faible pour espérer améliorer les qualités du modèle sur ces données.

5.2 Prévision de dépassement du seuil

Les mêmes données sont utilisées pour modéliser directement le dépassement de seuil sans passer par l'étape de modélisation de la concentration. Une fois optimisé par validation croisée, le modèle conduit à une prévision similaire de l'échantillon test avec la matrice de confusion ci-dessous.

	Pas de dépassement observé	Dépassement observé
Pas de dépassement prédit	162	18
Dépassement prédit	5	15

La qualité de prévision semble un peu meilleure mais, compte tenu de la faible taille de l'échantillon test, l'estimation du risque est peu fiable et les différences peu significatives. C'est la raison pour laquelle, la procédure d'estimation du risque est itérée B fois en considérant différentes séparations aléatoires des échantillons d'apprentissage et de test. Cette procédure spécifique, dite de validation croisée *Monte Carlo*, conduit à l'estimation de B erreurs de prévision ou risque pour comparer plusieurs algorithmes ou méthodes de prévision. Il est possible de calculer la moyenne de ces B erreurs, une moyenne pour chaque méthode ou encore d'afficher les diagrammes boîtes des distributions de ces erreurs.



Le graphique ci-dessus compare donc plusieurs méthodes de discrimination binaire : machine à vecteurs supports, forêt aléatoire, réseau de neurones, arbre de décision, k plus proches voisins et régression logistique. Même si la taille de l'échantillon test est modeste, le résultat permet de conclure que les méthodes ne conduisent pas à des résultats significativement très différents. Nous laisserons néanmoins de côté, les arbres de décision moins performants et réseaux de neurones, k plus proches voisins avec des erreurs plus dispersées. Finalement, entre les SVM, un peu meilleurs mais opaques et une régression logistique interprétable, il peut être préférable de choisir la régression logistique.

Il faudrait ajouter à ces résultats une comparaison des courbes ROC comme tracées dans la section précédente afin de faire intervenir le choix politique du seuil de décision dans la discussion. Tous ces traitements sont réalisés dans le

calepin `CSdD-Pic-Ozone-Python`.

5.3 Reconnaissance de l'activité humaine

Un modèle ou plutôt 6 modèles de régression logistique sont estimés sur les données issues des transformations des signaux enregistrés par des smartphones. En effet, par défaut, la librairie `scikit-learn` estime autant de modèles que de classes lorsque la variable Y est qualitative avec plus de 2 classes. Comme précédemment, une pénalisation Lasso est utilisée pour opérer une sélection de variables. Chaque modèle bénéficie d'une sélection de variables spécifique mais dirigées par la même valeur du coefficient C de pénalisation Lasso.

La sélection de variables ne conduit pas à un modèle simplifié, l'interprétation des coefficients des 6 modèles n'est pas raisonnablement possible. Les résultats se résument finalement à une matrice de confusion et un taux global d'erreur de moins de 4%, ce qui est tout à fait raisonnable et en accord avec les résultats de la phase exploratoire des données.

	Walking	Walking up.	Walking down.	Sitting	Standing	Laying
Walking	491	3	2	0	0	0
Walking upstairs	18	453	0	0	0	0
Walking downstairs	4	5	411	0	0	0
Sitting	0	4	0	430	56	1
Standing	2	0	0	12	518	0
Laying	0	0	0	0	0	537

Deux activités : assis *vs.* debout restent difficiles à discriminer.

Le calepin `CSdD-ML4IoT-Har-Python` compare les résultats de différentes méthodes d'apprentissage que ces mêmes données, notamment ceux de l'analyse discriminante linéaire et de l'algorithme des k plus proches voisins.

Références

Hastie T., Tibshirani R. et Friedman J. (2009). *The elements of statistical learning : data mining, inference, and prediction*, Springer, second edition.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pages 267-288).