Dauphine | PSL
UNIVERSITÉ PARIS

APPRENTISSAGE STATISTIQUE

Sujet 6

Construction d'un modèle de prédiction du cancer du sein à partir de biomarqueurs

Nathan BOYER
Sacha BITOUN

Guillaume PETIT
Thomas FAURE

7 Juin 2019

Table des matières

Table des figures

Introduction

L'étude réalisée dans ce projet concerne des patientes qui pourraient être atteintes du cancer du sein. Pour cela, nous allons définir des modèles de prédiction autour des biomarqueurs dont nous disposons.

Environnement d'étude

Les patientes observées sont agées de 29 ans à 90 ans. Plusieurs facteurs sont pris en compte dans le jeu de données pour la prédiction du cancer. Le BMI (qui est l'IMC), le taux de glucose dans le sang, le taux d'insuline de la patiente, le HOMA ainsi que 4 autres indices sur la santé de la patiente nous serviront de variables explicatives pour la classification.

Objectif

Nous sommes dans un cadre d'apprentissage supervisé puisque nous connaissons l'état des patientes étudiées grâce à la variable "Classification". Le but de l'étude est de définir un modèle qui permettra de prédire de manière aussi précise que possible si une patiente est atteinte du cancer du sein ou non. Nous sommes donc dans un domaine de classification. Ainsi, nous allons implémenter des méthodes de classification en apprentissage supervisé pour prédire la présence d'une tumeur cancéreuse.

1 Étude préalable des données

La base de données comporte initialement 116 observations de 9 variables explicatives. Nous avons d'abord modifié la variable d'intérêt classification qui renvoie 1 ou 2. Nous la rendons binaire avec 0 qui indique la présence d'un cancer du sein et 1 si la patiente n'est pas atteinte de cancer.

Nous avons voulu étudier les différentes corrélations entre les variables afin d'avoir une intuition concernant la variable d'intérêt et l'importance de ses liens avec les variables explicatives. Néanmoins, aucun lien n'est réellement ressorti de nos études. Qui plus est, nous possédons très peu de variables explicatives. Ainsi nous décidons de garder tous les paramètres.

1.1 Centrage et réduction

Nous notons l'échantillon entier X. Afin d'améliorer les performances des différentes méthodes, nous centrons et réduisons la matrice. Centrer les données n'a pas toujours d'influence. Cependant, lorsqu'on réduit les données, cela veut dire que toutes les variables varient selon la même échelle (les variables réduites sont toutes de variance 1). Donc quand on étudie les distances entre les points (par exemple pour déterminer les plus proches voisins), ces dernières

changent si les données sont réduites et donc les résultats également. Centrer-réduire les données est souvent une bonne idée, c'est même très conseillé lorsque les données ne sont pas dans la même unité. L'échantillon X à l'intervalle devient, avec la formule :

$$\widehat{X}_{ij} = \frac{X_{ij} - E[X_j]}{\sigma_j}$$

1.2 Organisation des données, matrice de confusion et fonction d'erreur

A partir de cette base de données, nous créons un échantillon d'apprentissage de taille n_{app} correspondant à 70% de l'échantillon initial et un échantillon de test de taille n_{test} qui contient les 30% restants.

On se basera sur une matrice de confusion, qui est un résumé des résultats de prédictions sur un problème de classification. Les prédictions correctes et incorrectes seront mises en lumière et réparties par classe (vrais positifs, vrais négatifs, faux positifs, faux négatifs). Les résultats sont ainsi comparés avec les valeurs réelles. Les valeurs réelles sont celles à la verticale et les valeurs prédites sont celles à l'horizontale.

	0	1
0	15	2
1	5	10

FIGURE 1 – Matrice de confusion

Dans cet exemple, 2 personnes sont supposées saines alors qu'elles ont le cancer (en haut à droite) et 5 personnes sont supposées cancéreuses alors qu'elles sont saines (en bas à gauche). Les valeurs dans la diagonale représentent les bonnes classifications.

Nous noterons fonction d'erreur la fonction suivante (proportion de variables d'intérêt non correctement prédites par l'estimateur) :

$$R(h) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} 1_{Y_i \neq h(\widehat{X}_i)}$$

où note h notre fonction de prédiction dépendant de X.

2 Methodes de classification

2.1 Méthodes simples

2.1.1 k-PPV

La première méthode testée est celle des k-plus proches voisins vue en cours. Dans un premier temps, on va déterminer le nombre de voisins k qui minimisera le risque. Bien sûr le nombre maximal de voisin ne doit pas excéder le cardinal de notre échantillon test.
Pour chaque valeur de k, nous utilisons notre échantillonnage apprentissage pour créer le modèle et ensuite calculer l'erreur sur l'échantillonnage test. D'après le graphique, nous observons que l'erreur est minimale pour $k_{opt} = 3$. Nous obtenons un taux d'erreur très bas d'environ 12%

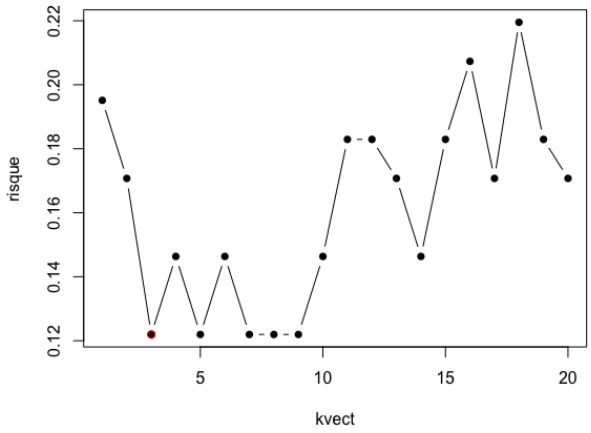

FIGURE 2 – Erreur des k-PPV pour différentes valeurs de k

	0	1
0	16	2
1	2	14

FIGURE 3 – Matrice de confusion pour la méthode des k-PPV

2.1.2 Forêts aléatoires

Notre seconde méthode est celle des fôrets aléatoires. Ici, il s'agit d'estimer le nombre d'arbres optimal pour la réduction de l'erreur. Le raisonnement est similaire au k-PPV pour trouver le nombre optimal d'arbre. Le nombre d'arbres optimal est 49, pour un taux d'erreur d'environ 18%.

Nous obtenons le graphique suivant :

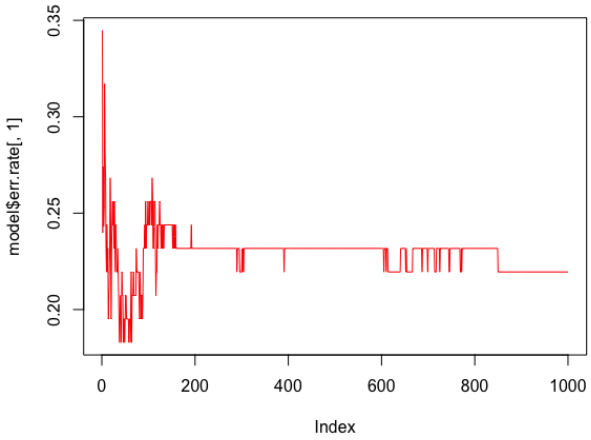

FIGURE 4 – Erreur de la méthode des forêts aléatoires en fonction du nombre d'arbres

La matrice de confusion est :

	0	1
0	18	2
1	4	10

FIGURE 5 – Matrice de confusion de la méthode des forêts aléatoires

On peut noter qu'il y a un plus fort de taux de faux malades avec cette méthode.

2.1.3 Régressions

Dans cette méthode, nous allons étudier le cas de deux régressions grâce aux régressions linéaires généralisées. En effet, notre objectif est de modéliser $\eta(\widehat{X}) = \mathbf{P}(Y = 1|\widehat{X})$ comme une fonction dépendant linéairement des observations. Nous avons peu de paramètres, donc nous ne cherchons pas à annuler certains coefficients. Nous allons donc faire une regression normale puis une régression Ridge.
C'est-à-dire qu'il existe $(\beta_0, ..., \beta_d)$ tel que :

$$\eta(x) = g(\beta_0 + \beta_1 x_1 + ... + \beta_d x_d)$$

avec $g(t) = \dfrac{e^t}{1 + e^t}$.

Après la régression simple, on obtient la matrice de confusion suivante :

	0	1
0	16	7
1	3	8

FIGURE 6 – Matrice de confusion de la régression

On observe un taux d'erreur de 29 %. On notera dans cette méthode le problème inverse aux forêts aléatoires, qui est une plus forte chance d'être faussement prédit sain. Ce qui est en soit plus problèmatique.

Essayons d'améliorer ces résultats avec une régression Ridge. On veut résoudre le problème :

$$\widehat{\beta} \in \underset{\beta \in \mathbf{R}^d}{argmin} \sum_{i=1}^{n} (Y_i - \beta X_i) + \lambda \|\beta\|_2^2$$

On cherche par validation croisée le paramètre qui nous donnera le meilleur taux de réussite. D'après le graphique, on voit que $\lambda_{opt} = 0.1010101$ pour un taux de réussite d'environ 77 %. Au vu des résultats de la matrice de confusion, les résultats se sont sensiblement améliorés mais le même problème persiste.

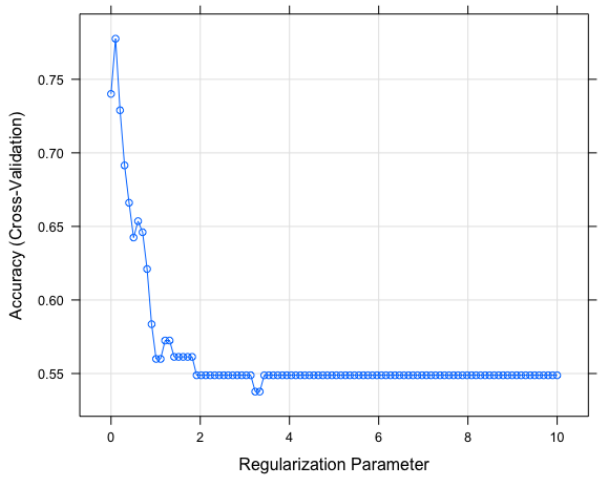

FIGURE 7 – le taux de réussite par rapport au paramètre lambda

On obtient la matrice de confusion suivante :

	0	1
0	17	6
1	2	9

FIGURE 8 – Matrice de confusion de la méthode de Ridge

2.1.4 SVM

On a décidé de faire un algorithme de vecteur de support. On a pris le noyau Gaussien Radial, ie
$\forall (u, v) \in \mathbf{R}^2$,

$$K(u, v) = e^{-\gamma \|x-y\|^2}$$

et on cherche à résoudre :

$$\min_{\beta_0, \beta, \xi_i} \quad \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{n} \xi_i \tag{1}$$

$$s.c. \qquad y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \forall i = 1, ...n \tag{2}$$

$$\xi_i \geq 0, \forall i = 1, ..., n \tag{3}$$

Le paramètre C, appelé fonction de coût, accentue plus ou moins la tolérance aux erreurs. On va donc effectuer une validation croisée pour obtenir la meilleure fonction de coût C pour avoir le meilleur taux de réussite. D'après les graphique, on voit qu'on optimise le taux de réussite à 80 % à peu près pour $C = 8$.

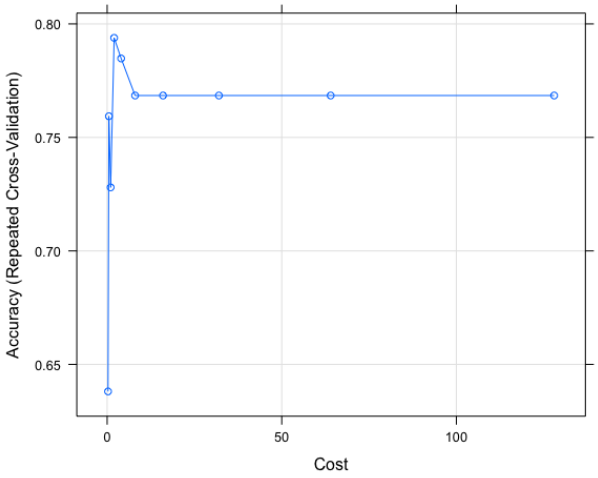

FIGURE 9 – Taux de réussite en fonction de la fonction C

On obtient la matrice de confusion suivante :

	0	1
0	15	3
1	4	12

FIGURE 10 – Matrice de confusion de la méthode SVM

Il est intéressant de noter qu'il y a autant de chance d'être faussement sain que d'être faussement malade avec cette méthode. Ce qui diverge des régression ou des forêts aléatoires.

2.2 Optimisation

Malheureusement, nous faisons face à un problème qui peut fausser les résultats. C'est le peu de données dont nous disposons. Nous allons donc essayer d'optimiser nos résultats.

Premièrement, le schéma apprentissage-test n'est pas adapté lorsque nous traitons des bases de taille réduite. Nous conservons notre échantillon d'apprentissage tel quel et nous procédons par validation croisée pour en estimer les performances. Nous verrons si le taux obtenu est conforme à celui mesuré sur l'échantillon test que nous avons mis à part. Cette validation croisée s'effectue de la manière suivante :

1. On subdivise les données en K blocs (on prendra K=10) ;
2. On répète K fois le processus suivant
   — apprentissage sur K-1 blocs,
   — test sur le K ième bloc,
   — on mesure l'erreur e_k ;
3. on calcule l'erreur de la validation croisée, notée e_{vc} où

$$e_{vc} = \frac{1}{K} \sum_{K} e_k$$

Deuxièmement, nous allons étudier l'indice ROC pour évaluer la performance de notre modèle. La mesure ROC permet de mesurer la performance d'un estimateur binaire. C'est à dire un estimateur qui classe les éléments en deux groupes comme c'est le cas dans notre étude (cancer ou non). Généralement, on représente cette mesure par une courbe dite courbe ROC qui représente le taux de vrais positifs (c'est la sensibilité) en fonction du taux de faux positifs (c'est l'antispécificité) pour différents seuils.

En effet on est classé comme cancereux à partir d'un certain seuil et le but est de choisir le seuil optimal. La courbe ROC représente la sensibilité en fonction de l'antispécificité lorsque le seuil varie. On peut choisir ensuite le seuil optimal en prenant le point le plus proche de (1,1), et le plus éloigné de la diagonale x=y.
À (0, 0) le classificateur déclare toujours 'négatif'.
À (1, 1) le classificateur déclare toujours 'positif'.
À (0, 1) le classificateur n'a aucun faux positif ni aucun faux négatif, et est par conséquent parfaitement exact, ne se trompant jamais.

La courbe ROC permet également de sélectionner le meilleur estimateur grâce à l'aire sous la courbe de (0,0) à (1,1) on parle de AUC. L'AUC fournit une mesure agrégée des performances pour tous les seuils de classification possibles. Un AUC de 0,60 signifie que l'estimateur répond correctement (vrais positifs et vrais négatifs) 60 % du temps.

On considère habituellement que le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7. Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9. Un modèle ayant une AUC supérieure à 0.9 est excellent.

Nous avons ainsi utilisé cette méthode afin de comparer les méthodes des k plus proches voisins, de la random forest, de la régression linéaire classique et de la régression linéaire avec l'estimateur de Ridge.

2.2.1 k-PPV

On va donc garder le même nombre de voisins et faire une validation croisée de notre méthode et obtenir sa courbe ROC. Après traitmeent, on obtient la matrice de confusion suivante :

	0	1
0	15	4
1	6	9

FIGURE 11 – Matrice de confusion de la méthode k-PPV après validation croisée

Le taux d'erreur est ici de 30 %. On voit que le modèle n'est pas du tout robuste et ses performances varient beaucoup.

Nous trouvons une AUC de 0,69 pour les k plus proches voisins. On peut donc douter de la fiabilité du modèle.

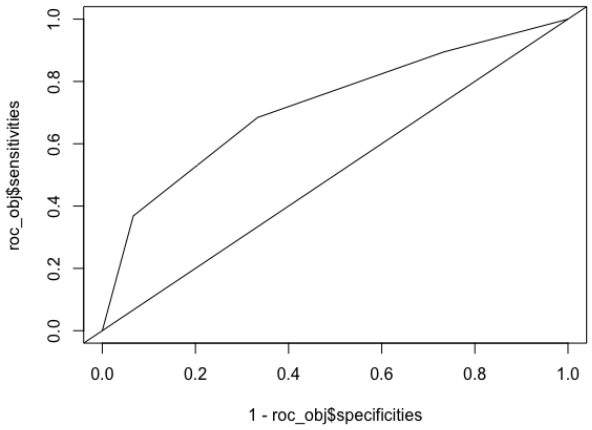

FIGURE 12 – Courbe ROC pour l'estimateur des k PPV

2.2.2 Forêts aléatoires

On réitère la même méthode pour étudier la robustesse de notre modèle. On obtient la matrice de confusion suivante :

	0	1
0	17	3
1	4	10

FIGURE 13 – Matrice de confusion de la méthode des forêts aléatoires après validation croisée

Le taux d'erreur est ici de 21 %. Avant validation croisée, le taux d'erreur était 18 %. Cela semble assez logique. On aurait pu se passer de la subdivision des données en apprentissage test.

Nous trouvons une AUC de 0,80 pour la Random forest. C'est donc un modèle adapté pour notre problématique.

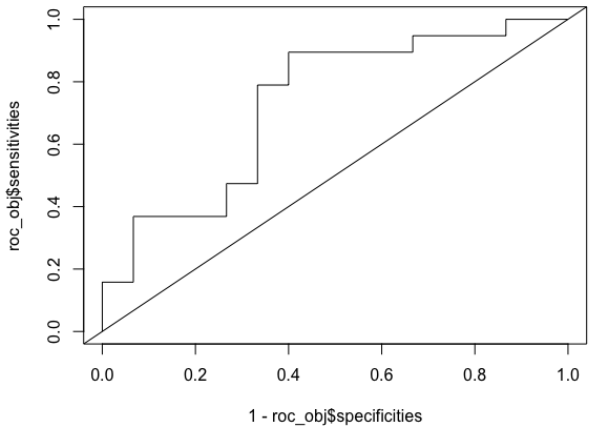

FIGURE 14 – Courbe Roc pour l'estimateur des forêts aléatoires

2.2.3 Régressions

Nous avons effectué la validation croisée que pour le regréssion simple. On obtient la matrice de confusion suivante :

	0	1
0	17	3
1	4	10

FIGURE 15 – Matrice de confusion d'une régression après validation croisée

Le taux d'erreur est ici de 41 %. La méthode n'est dans ce cas pas très robuste.

Nous trouvons une AUC de 0,62 pour la régréssion ordinaire et de 0,758 pour la régression Ridge. On peut donc douter de la fiabilité des régressions.

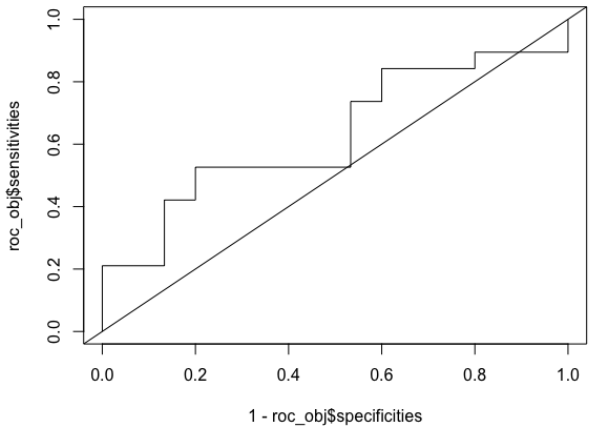

FIGURE 16 – Courbe Roc pour l'estimateur de la régression linéaire

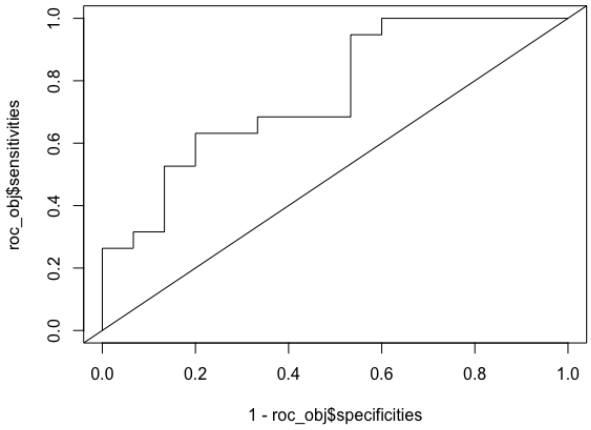

FIGURE 17 – Courbe Roc pour la régression Ridge

2.2.4 SVM

On a décidé d'effectuer une validation croisée plus optimale pour la méthode SVM en essayant d'optimiser la fonction de coût et la paramètre γ. Il faut prendre en compte que cela augmente considérablement le temps de calcul. Elle est donc à proscrire quand on a une quantité de données trop conséquente.

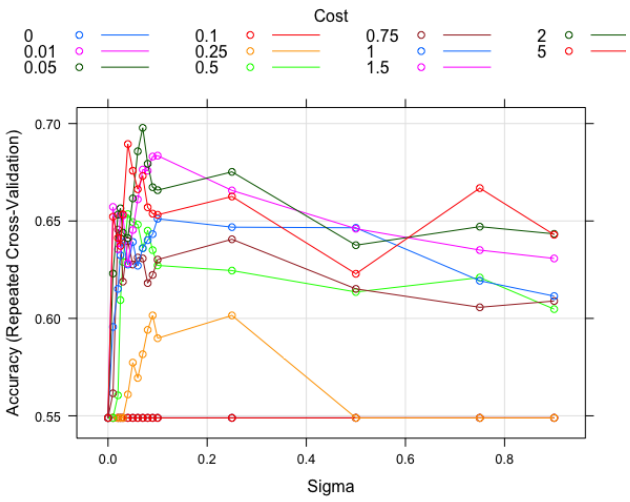

FIGURE 18 – Taux de réussite en fonction de la fonction C et du pramaètre gamma

On a alors un taux de réussite plus élevé à 80 % avec la subdivison training pour $C = 2$ et $\gamma = 0,07$. On obtient la matrice de confusion :

	0	1
0	17	2
1	2	13

FIGURE 19 – Matrice de confusion de SVM radial après validation croisée

On a alors un taux d'erreur de 12 %. On n'a pas pu avoir la courbe ROC de la méthode SVM.

2.3 Importance des paramètres

Dû au peu de données qu'on possède, nos tests possèdent des taux d'erreur tout de même assez élevés. Il est donc assez intéressant d'étudier quel paramètre est le plus important dans l'analyse des algorithmes que nous utilisons. On peut espérer y trouver une similitude parmi les méthodes, donnant une piste d'étude pour les variables. Sous R, grâce au package utilisé, on obtient ces résultats sur une échelle de 100, où la plus significative possède un score de 100 et la moins sgnificative possède une score de 0. On teste l'importance des variables pour les méthodes de régression et des forêts aléatoires. On prend les 4 premiers de chaque test choisi et on obtient le tableau suivant :

Forêt aléatoire		Regression		Ridge	
Glucose	100	Glucose	100	Resistin	100
Age	87	Resistin	97	Glucose	84
Resistin	66	BMI	78	Insuline	62
BMI	44	Age	50	HOMA	50

3 Conclusion

On peut donc conclure que sur les plusieurs méthodes testées, rares sont celles qui sont à la fois performantes et stables (indépendante du jeu de données sans trop de variations). Cependant, on a vu que certains modèles sont plus suceptibles d'être crédibles que d'autres. Notamment le modèle des forêts aléatoires est de bonne qualité - meilleur indice ROC, ce qui signifie que pour une valeur seuil quelconque l'estimateur de la random forest sera plus fiable en général.
De plus, grâce à ces différents tests, on a donc une idée des variables les plus importantes, les plus influentes, pour la classification. Le glucose, l'âge, l'IMC et l'insuline sont des données à prendre en compte.