Modélisation de la faillite d'exploitations agricoles : comparaison de méthodes de régression non linéaire avec R

Thibault LAURENT

Toulouse School of Economics (GREMAQ),
thibault.laurent@univ-tlse1.fr

25 Mai 2009

Plan

Contexte de l'étude

Dominique Desbois (2008), "Introduction to Scoring Methods : Financial Problems of Farm Holdings", CS-BIGS 2(1) : 56-76.

- **Détecter** et **prévenir** les exploitations agricoles qui présentent un risque de faillite.
- **Collecte** et **exploration** des données (ACP, analyse discriminante, classification, etc).
- **Méthodes de prédiction utilisées :** analyse discriminante décisionnelle et régression logistique

Objectifs

- Utiliser d'autres méthodes de régression non linéaires comme les **modèles additifs généralisés** introduits par Hastie et Tibshirani (1986), le **bagging** et les **forêts aléatoires** sur arbre de classification proposés par Breiman (1996, 2001) et les **support vecteur machines** proposées par Vapnik (1999).
- Faire le tour des programmes disponibles sous R pour traiter ces méthodes.

Les données

- **1260** exploitations agricoles spécialisées dans les grandes cultures issues de 4 départements français (Eure-27, Nord-59, Orne-61 et Seine-Maritime-76). Période d'observation : 1988 à 1994.

Les données

- **1260** exploitations agricoles spécialisées dans les grandes cultures issues de 4 départements français (Eure-27, Nord-59, Orne-61 et Seine-Maritime-76). Période d'observation : 1988 à 1994.
- Variable **Y** à expliquer : incident de paiement (1 si défaillant et 0 si sain)

Les données

- **1260** exploitations agricoles spécialisées dans les grandes cultures issues de 4 départements français (Eure-27, Nord-59, Orne-61 et Seine-Maritime-76). Période d'observation : 1988 à 1994.
- Variable **Y** à expliquer : incident de paiement (1 si défaillant et 0 si sain)
- Variables explicatives : structure de l'exploitation (statut juridique, surface agricole utilisée, âge de l'exploitant, etc.) + 22 ratios r_i sélectionnés par thème (structure financière, poids de la dette, liquidité, service de la dette, rentabilité du capital, résultat, activité productive)

Protocole expérimental (1)

Pour **k allant de 1 à 100**

1. Découpage de l'échantillon en échantillon d'apprentissage et test (80% et 20%).

Protocole expérimental (1)

Pour **k allant de 1 à 100**

1. Découpage de l'échantillon en échantillon d'apprentissage et test (80% et 20%).
2. Echantillon d'apprentissage : ajustement des méthodes statistiques, construction de scores, prédiction $\hat{Y}_{App} = 1 \Longleftrightarrow score > c$, calcul du taux de mal classées pour différentes valeurs de c, choix du c qui minimise le taux d'erreur.

Protocole expérimental (1)

Pour **k allant de 1 à 100**

1. Découpage de l'échantillon en échantillon d'apprentissage et test (80% et 20%).
2. Echantillon d'apprentissage : ajustement des méthodes statistiques, construction de scores, prédiction $\hat{Y}_{App} = 1 \Longleftrightarrow score > c$, calcul du taux de mal classées pour différentes valeurs de c, choix du c qui minimise le taux d'erreur.
3. Echantillon test : calcul de scores, prédiction $\hat{Y}_{test} = 1 \Longleftrightarrow score > c$ et calcul du taux de mal classées

Protocole expérimental (2)

1. Méthodes appliquées sur les variables choisies par Desbois

Protocole expérimental (2)

1. Méthodes appliquées sur les variables choisies par Desbois

2. Choix d'autres variables en utilisant la spécificité de chaque méthode statistique.

Méthodes statistiques

Régression Logistique
Modèles additifs généralisés
Bagging/Forêts aléatoires sur arbre de classification
Support Vecteur Machines

Régression Logistique (1)

Desbois (2008) modélise la probabilité de faillite par :

$$P[Y=1] = \frac{e^{(-6.17+5.95r_1+0.952r_{12}+3.36r_{14}+24.23r_{17}-7.3r_{32}+0.61r_{36})}}{1+e^{(-6.17+5.95r_1+0.952r_{12}+3.36r_{14}+24.23r_{17}-7.3r_{32}+0.61r_{36})}},$$

où r_1, r_{12}, r_{14}, r_{17}, r_{32} et r_{36} sont choisies après une sélection pas-à-pas ascendante et prédit :

$$\hat{Y} = 1 \text{ si } P[Y=1] > c,$$

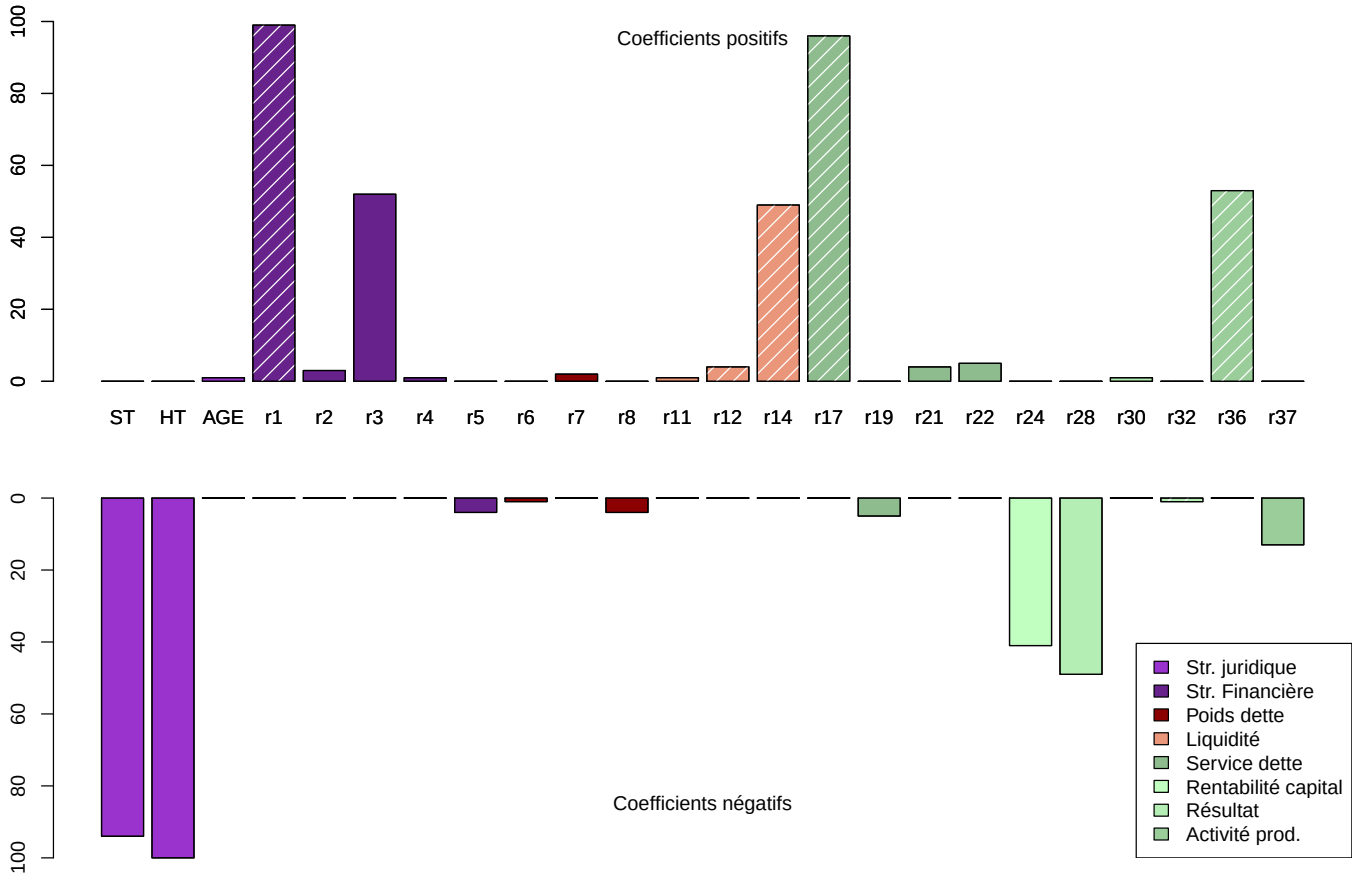
où c est choisi tel que le taux d'erreur ou taux de mal classées (dans cette étude) soit minimum sur l'échantillon d'apprentissage.

Régression Logistique (2)

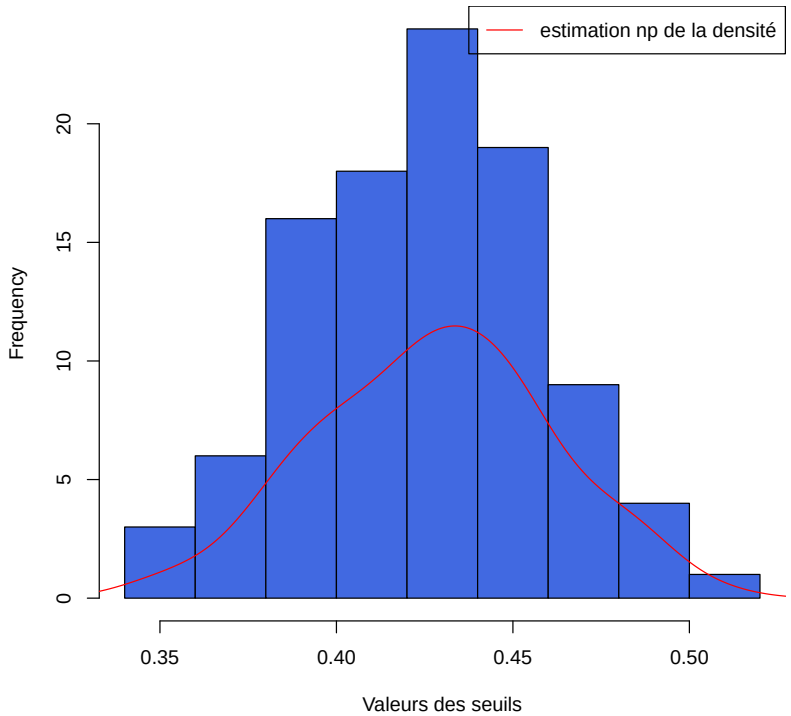
Dans cette étude, nous choisissons pour chaque k allant de 1 à 100 :

- Le modèle qui minimise le critère BIC ($-2\mathcal{L} + p \log n$) en partant des p variables : (fonction *bic.glm()* de la librairie `BMA`)
- Le seuil c qui minimise le taux d'erreur sur l'échantillon d'apprentissage

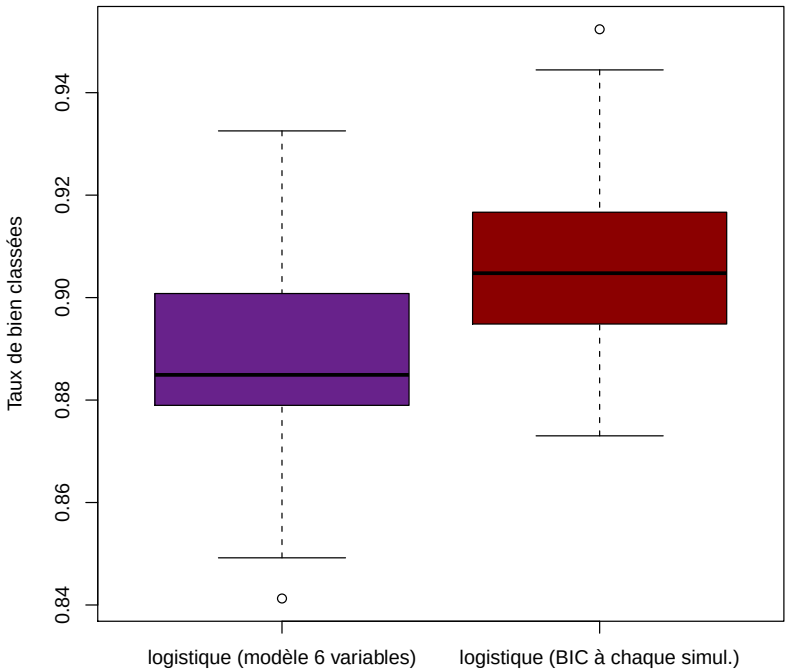
Représentativité des variables sur 100 modèles



Représentation des seuils choisis



Comparaison des méthodes



$$F - statistic = 58.43$$
$$p - value < 0.0001$$

Avantage de la régression logistique

Facilité d'interprétation des coefficients !

exemple : une augmentation de 1% de la variable r_1 (toutes autres choses restant égales par ailleurs) augmente la probabilité de faillite d'une exploitation de ...

Méthodes statistiques

Modèles additifs généralisés

Prise en compte de l'aspect non linéaire dans les régresseurs de Desbois en utilisant un modèle additif généralisé (packages `mgcv` ou `gam`) :

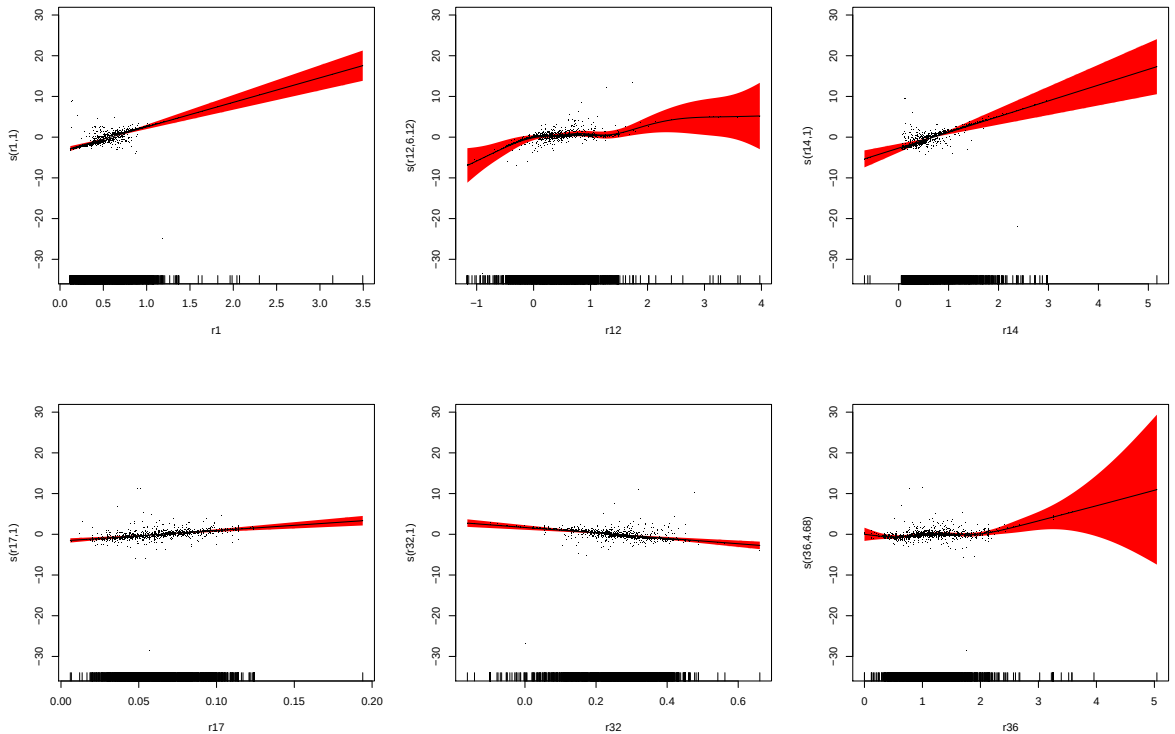
$$P[Y=1] = \frac{e^{(\beta_0+f_1(r_1)+f_2(r_{12})+f_3(r_{14})+f_4(r_{17})+f_5(r_{32})+f_6(r_{36}))}}{1+e^{(\beta_0+f_1(r_1)+f_2(r_{12})+f_3(r_{14})+f_4(r_{17})+f_5(r_{32})+f_6(r_{36}))}},$$

où f_1, f_2, f_3, f_4, f_5 et f_6 sont des splines de régression pénalisées dont les degrés de liberté sont déterminés de façon à minimiser le critère *GCV/UBRE/AIC*.

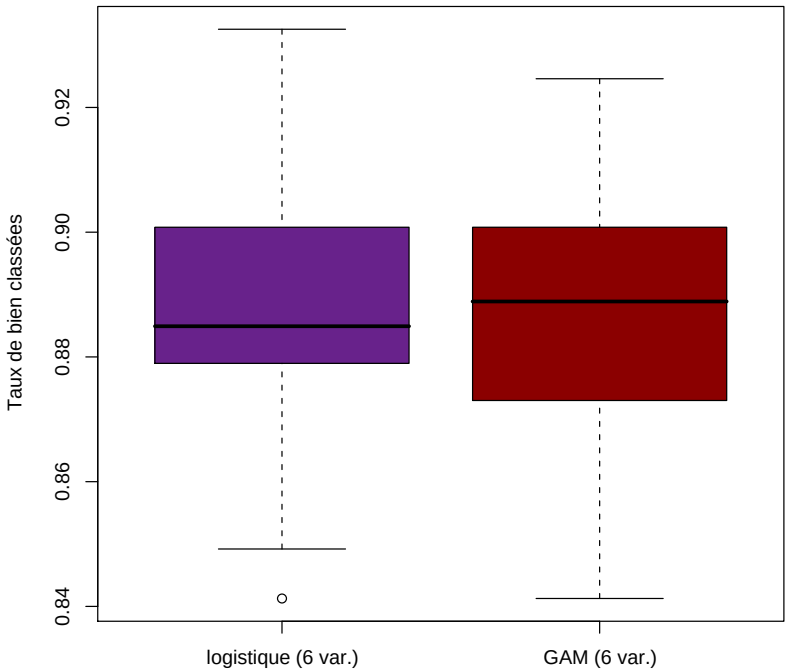
- Paramètres initiaux (choix des noeuds, etc.) par défaut de la fonction *gam()*.

Représentation des résidus partiels

Représentation des résidus partiels (fonction *plot.gam()*) après un ajustement d'un modèle GAM sur les 6 variables sélectionnées par Desbois.



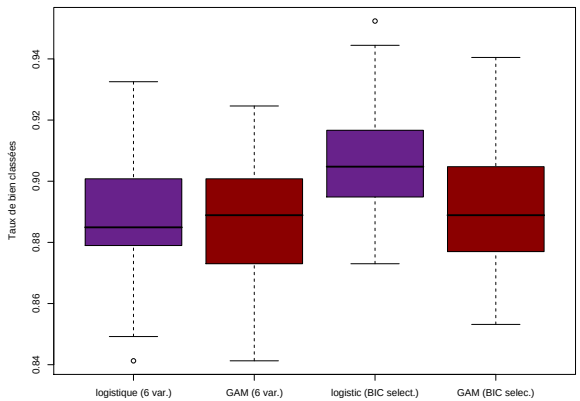
Comparaison du modèle Logistique vs. GAM à 6 variables



$$F - statistic = 0.22$$
$$p - value = 0.64$$

GAM sur variables sélectionnées après la régression logistique

Pour k allant de 1 à 100, on ajuste un modèle GAM à partir des variables sélectionnées par minimisation du critère BIC dans le modèle logistique.

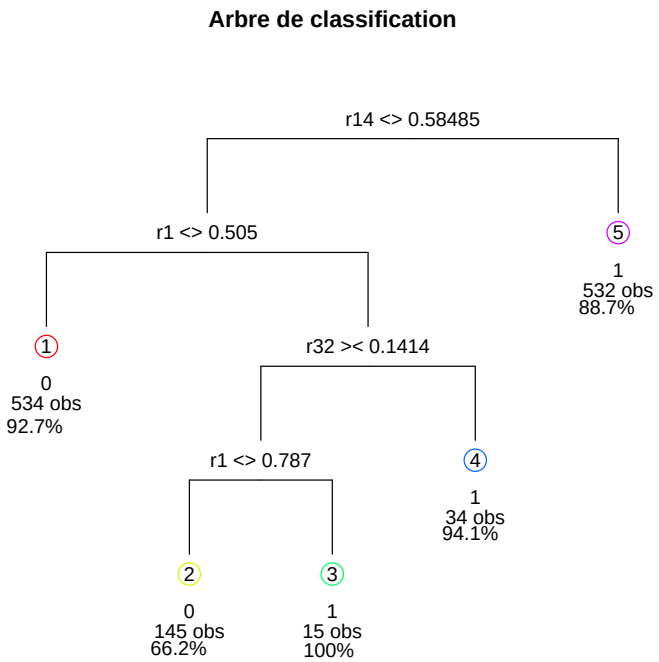


$F - statistic = 25.66$
$p - value < 0.0001$
Le taux de mal classées du modèle logistique (BIC sél.) est significativement différent des autres.

Méthodes statistiques

CART

Bagging et forêts aléatoires sont des "améliorations" des arbres de régression et classification (CART - packages `rpart` et `maptree`)

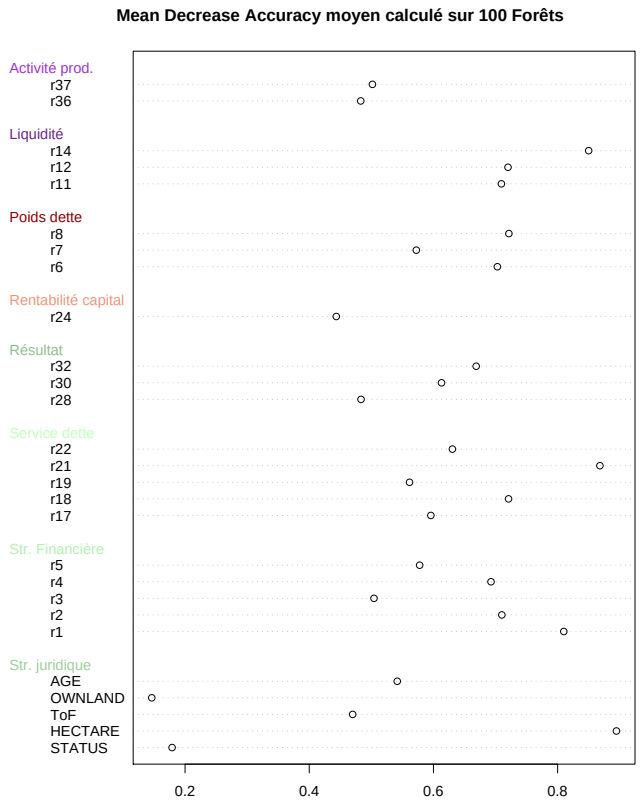


Principe de ces méthodes

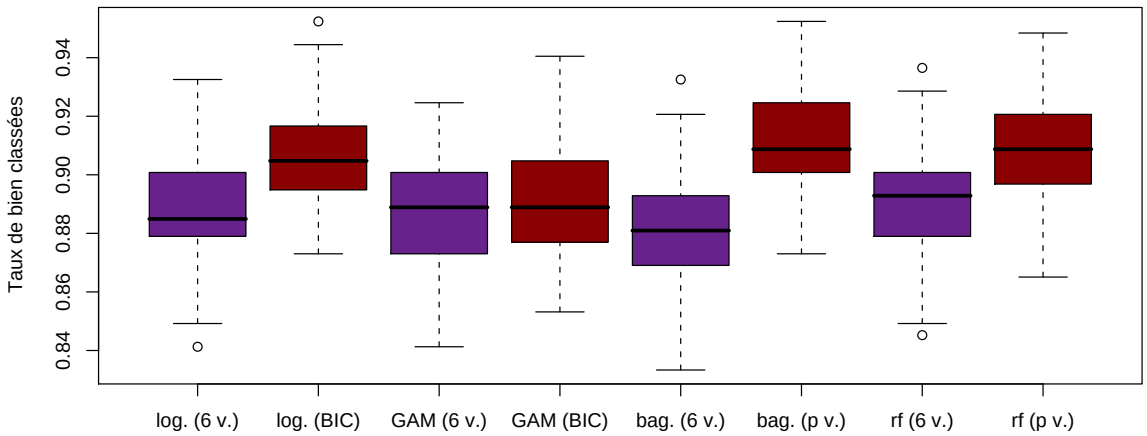
1. Bagging : pour chaque k allant de 1 à 100, on construit 25 échantillons boostrap pour chacun desquels on construit un arbre de classification. On calcule ensuite sur l'échantillon test le nombre de fois où une exploitation a été prédite en difficulté sur ces 25 arbres. Si ce nombre est supérieur à c (c estimé comme dans les méthodes précédentes), on prédit l'exploitation en faillite. Fonction *bagging()* du package *ipred*.
2. Forêts aléatoires : à chaque nouveau noeud, la variable qui dichotomise une branche est choisie parmi un ensemble q de variables tirées aléatoirement. Chaque arbre de classification est élagué de façon à conserver un faible nombre l de noeuds. Fonction *rf()* du package `randomForest`.

Interprétation

- Méthodes construites par aggrégation $\simeq$ pas d'interprétation directe
- Critère (Mean Decrease Accuracy) mesure l'importance d'une variable reposant sur le calcul de la perte de qualité de la prédiction induite par une permutation aléatoire des valeurs de la variable considérée.



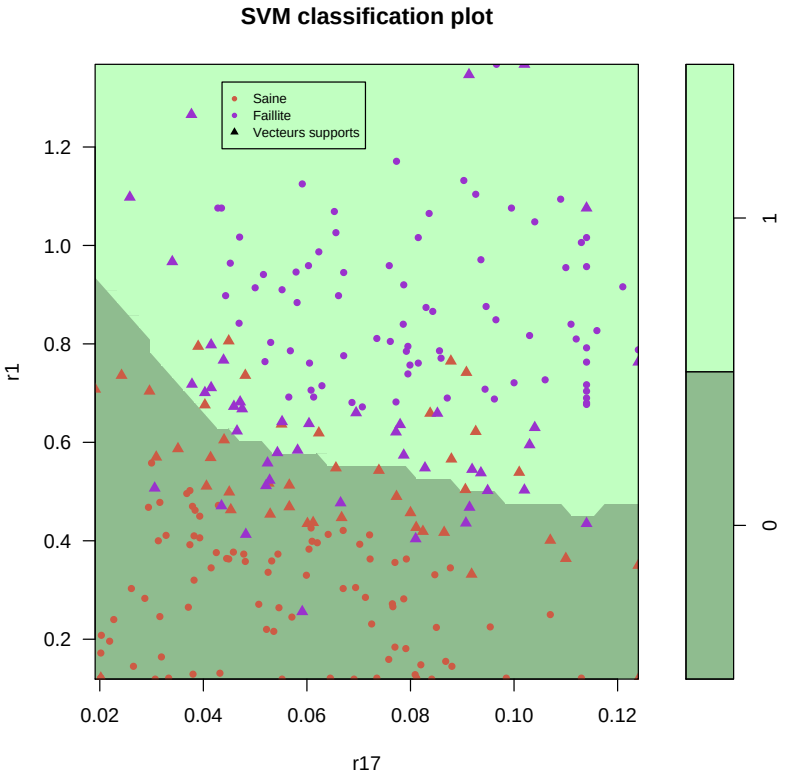
Résultats


Taux de bien classées
0.84
0.86
0.88
0.90
0.92
0.94
log. (6 v.)
log. (BIC)
GAM (6 v.)
GAM (BIC)
bag. (6 v.)
bag. (p v.)
rf (6 v.)
rf (p v.)

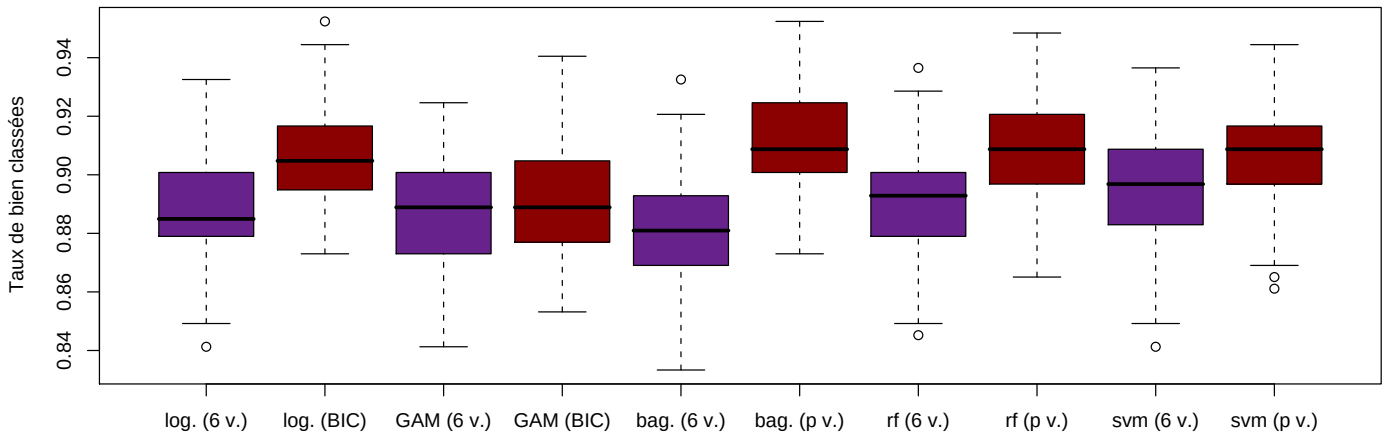
Méthodes statistiques

Support Vecteur Machines

- Construire l'hyperplan optimal qui sépare les classes
- Recherche de surfaces séparatrices non linéaires obtenue par l'introduction d'une fonction noyau
- fonction *svm()* du package e1071



Résultats


Taux de bien classées
0.84
0.86
0.88
0.90
0.92
0.94
log. (6 v.)
log. (BIC)
GAM (6 v.)
GAM (BIC)
bag. (6 v.)
bag. (p v.)
rf (6 v.)
rf (p v.)
svm (6 v.)
svm (p v.)

Conclusion/Perspectives

Récapitulatifs

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8876	0.0018	486.10	0.0000
Logi. (BIC)	0.0185	0.0026	7.15	0.0000
GAM (6 v.)	-0.0012	0.0026	-0.48	0.6339
GAM (BIC)	0.0035	0.0026	1.37	0.1717
bagg. (6 v.)	-0.0060	0.0026	-2.31	0.0214
bagg. (p v.)	0.0238	0.0026	9.22	0.0000
forêts (6 v.)	0.0035	0.0026	1.37	0.1717
forêts (p v.)	0.0216	0.0026	8.36	0.0000
svm (6 v.)	0.0076	0.0026	2.95	0.0032
svm (p v.)	0.0199	0.0026	7.71	0.0000

Conclusions/Perspectives

- Les méthodes d'agrégation bagging et forêts aléatoires donnent de bons résultats appliquées sur toutes les variables. Inconvénient : interprétations limitées.
- SVM, concurrent sérieux des méthodes d'agrégation, mais reste une boîte noire.
- Le réglage des paramètres des méthodes bagging, forêts aléatoires et SVM peut être amélioré avec la fonction *tune()*.
- GAM, en-dessous des espérances, peut-être à cause d'un sur-ajustement des données dans l'échantillon d'apprentissage. Réglage des paramètres à revoir.
- Régression logistique donne des résultats satisfaisants après minimisation du critère BIC. Facile à interpréter.

Conclusions/Perspectives

Toutes ces méthodes vallent la peine d'être testées et selon le jeu de données, apporteront des éléments de réponse complémentaires aux questions soulevées par un problème de discrimination.

Conclusions/Perspectives

Toutes ces méthodes vallent la peine d'être testées et selon le jeu de données, apporteront des éléments de réponse complémentaires aux questions soulevées par un problème de discrimination.

Merci de votre attention

Bibliographie

[1] Besse, P. (2008) *Apprentissage Statistique et Data Mining*, `http://www.math.univ-toulouse.fr/~besse/enseignement.html`.
[2] Breiman, L. (1996) Bagging predictors, *Machine Learning*, 26(2) :123-140.
[3] Desbois, D. (2008) Introduction to Scoring Methods : Financial Problems of Farm Holdings, *CS-BIGS*, 2(1) : 56-76.
[4] Hastie, T. et Tibshirani, R. (1986) Generalized Additive Models, *Statistical Science 1*, 297-318.
[5] Vapnik, V.N. (1999) *Statistical learning theory*, Wiley Inter science.
[6] Wood, S. (2006) *Generalized Additive Models : An Introduction with R.* , Chapman & Hall/CRC.