TP Bilan Data Mining

Rémi MUSSARD - Thomas ROBERT

8 janvier 2014

Pour ce TP, nous avons essayé des méthodes de SVM et de régression logistique sur 2 jeux de données *abalone* et *spam base*. Le code du fichier Matlab est organisé dans le même ordre que ce compte-rendu.

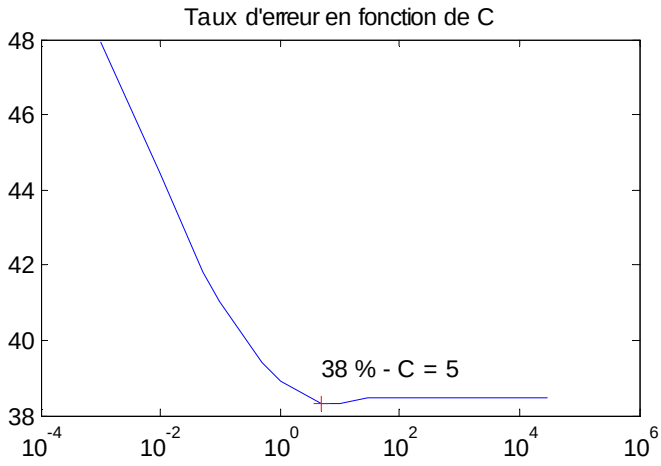
Jeu de données *abalone*

Introduction

Le jeu de données Abalone continent des données sur les ormeaux. La première variable représente le sexe de l'« individu » et a donc été supprimé car ne permet pas vraiment de déterminer la frontière. Nous avons fait des essais avec et sans mais cela n'améliore pas les résultats. Le label représente l'âge de l'individu (« jeune » ou « vieux »).

SVM

Nous avons commencé par faire une séparation par SVM en essayant de nombreuses valeurs de C. Le taux d'erreur varie entre 48 et 38%.



Regression logistique

Face à ce taux d'erreur relativement important, nous avons décidé d'essayer une seconde méthode : la régression logistique.

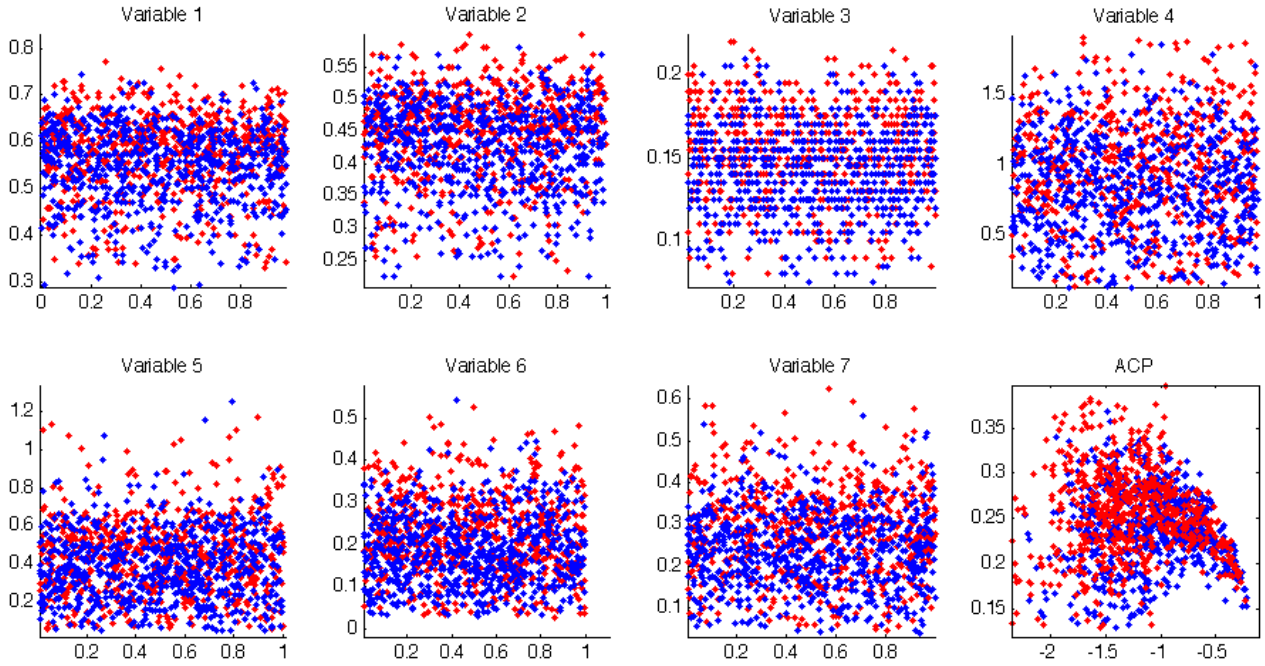
Nous avons essayé de réaliser la régression logistique avec une frontière linéaire $(\varphi = [1\ X])$ puis quadratique $\left(\varphi = \left[1\ \ x_1\ \ \ldots\ x_p\ \ x_1^2\ \ \ldots\ x_p^2\right]\right)$. Les taux d'erreurs obtenus sur de 38 et 39%, des résultats équivalents au SVM.

Affichage des données

Etonnés que les résultats ne s'améliorent pas en quadratique, nous avons essayé d'afficher les données de plusieurs façons.

Nous avons fait une ACP qui contient 98% d'information sur la 1ère variable et 99% sur les deux premières. En affichant les données sur les deux premiers axes, on voit que ces deux axes ne nous permettent pas vraiment de tracer une frontière entre les deux ensembles.

Ensuite, nous avons affiché que variable indépendant, pensant que certaines variables (la taille par exemple) devrait sans doute être très liée à l'âge de l'individu, et permettrait donc de départager les données. Nous avons donc regardé la répartition des valeurs de la première variable sur une droite verticale, colorées selon l'âge. Pour mieux voir les données afin qu'elles ne se superposent pas toutes sur une ligne, nous les avons éparpillées selon l'axe horizontal par un simple *rand*. On aurait donc aimé voir un groupe coloré en haut et un groupe coloré en bas, mais ce n'est pas du tout le cas, et on voit qu'aucune variable ne permet de séparer les données.



Affichages des données abalone

Conclusion

Arrivés à là, nous nous sommes dit qu'il serait plutôt difficile d'améliorer le score de 38% d'erreur, nous avons donc décidé d'en rester là et de tenter notre chance avec un second jeu de données.

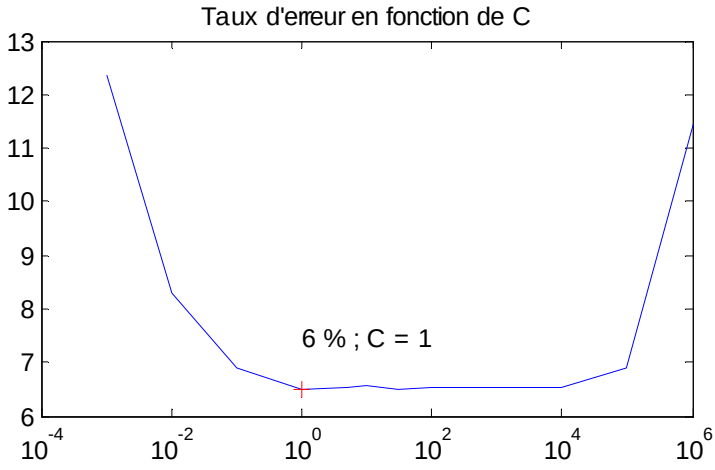
Jeu de données *spam base*

Introduction

Le jeu de données *spam base* contient des données sur des mails. Les variables représentent les diverses caractéristiques qui permettent de classifier ces mails en tant que spam ou non spam. Les labels représentent la classification du mail : spam ou non spam.

SVM

Nous avons commencé par faire une séparation par SVM. Après avoir testé différentes valeurs de C, on obtient un taux d'erreur minimal pour $C = 1$. Avec une classification par SVM, on obtient donc un taux d'erreur de classification de 6,5 %, un bon score.



Régression logistique

On a par la suite testé la méthode de régression logistique, même si la classification par SVM nous donne un taux d'erreur relativement faible.

On a tout d'abord testé une méthode de régression logistique linéaire, en prenant l'ensemble des données comme données d'apprentissage. On obtient ainsi une erreur de 6,9 %.

On a ensuite testé une méthode de régression logistique quadratique, car en général, il est rare que la frontière de décision soit linéaire. Avec la régression quadratique, on obtient un taux d'erreur de 5,6 %.

Notons que la fonction `ma_reg_log` a dû être adaptée. En effet, le conditionnement de la matrice W formée lors de la régression n'était pas bon. Pour résoudre ce problème, on a ajouté une valeur sur la diagonale de W. Après avoir testé différentes valeurs, on retiendra une valeur de $0{,}1$.

Régression logistique avec séparation test et apprentissage

Afin de tester les performances de la méthode, on simule un cas réel avec un ensemble d'apprentissage (25% du jeu de données) et un ensemble de test (75% du jeu de données).

On utilise la méthode la plus efficace, c'est-à-dire la régression logistique quadratique.

On trouve alors des performances assez similaires à celles trouvées précédemment, c'est-à-dire 1,2% d'erreur en apprentissage, 6,3% d'erreurs en test. *(Résultats trouvés en moyennant les résultats sur 50 tirages de la séparation du jeu)*

Conclusion

On obtient donc une classification avec une erreur minimale de 5,6 % avec la Régression Logistique Quadratique sur l'ensemble du jeu de données.