Nous travaillons sous R. Tous les tests sont à 5%.

1 Supports

Le site de notre cours est http://eric.univ-lyon2.fr/~ricco/cours/cours_regression_logistique.html

Plus spécifiquement pour cette séance, nous nous référerons à :

TUTO 1 – http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique_polytomique.pdf

TUTO 2 – http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf

TUTO 3 - http://tutoriels-data-mining.blogspot.com/2012/03/introduction-r-arbre-de-decision.html

N'oubliez pas que vous avez accès aux corrigés de nos précédentes séances !

2 Données

Nous travaillons sur une version de la base WAVEFORM où la variable cible « classe » comporte 3 modalités {A, B, C}. Le classeur Excel « **waveform.xlsx** » comporter 2 feuilles : la première correspond aux données d'apprentissage, la seconde au test. C'est un jeu de données artificiel, j'ai exagéré la taille de l'échantillon test pour obtenir une estimation stable des performances.

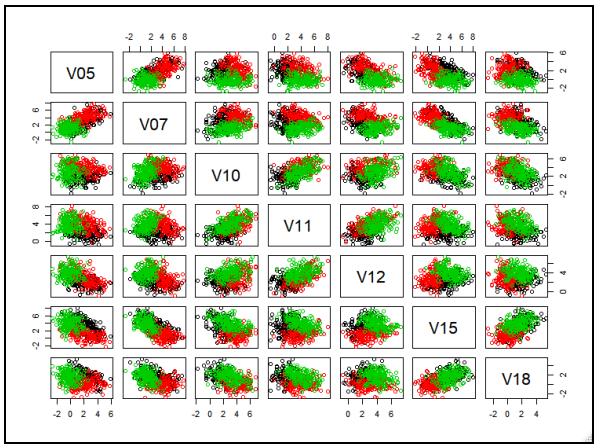
3 Exercices – Régression logistique multinomiale

3.1 Importation et préparation des données

0. Chargez la première feuille du fichier « **waveform.xlsx** » (read.xlsx si package « xlsx »). Affichez les propriétés de la base (str). Nombre d'observations et de variables ? (500, 8).

1. Affichez les statistiques descriptives (summary). Quels sont les effectifs de chaque classe {A, B, C}. Les variables explicatives potentielles sont-elles centrées ?

2. Réalisez un graphique nuage de points par paires de variables explicatives, coloriez les individus selon leur classe d'appartenance (pairs). Avez-vous des commentaires à faire ?



3. Nous souhaitons centrer et réduire les variables explicatives pour notre analyse, parce que le package de machine learning pour la régression que nous utiliserons plus loin requiert cette préparation. Faites appel à la fonction scale(), veillez à préciser les options adéquates pour que l'outil opère bien un centrage et une réduction.
4. Calculez les statistiques descriptives de la base des variables transformées (summary). Que constatez-vous ?
5. L'objet fourni par scale() possède d'autres propriétés. Affichez-les (attributes).
6. Que représentent les deux dernières propriétés. Comment y accède-t-on ? (attr)

3.2 Régression logistique avec « nnet »

7. Installez et chargez le package « nnet ».
8. Réalisez la régression visant à expliquer « classe » en fonction des variables centrées et réduites (multinom). Que lisez-vous dans les résultats ? (print) Combien d'équation de régression obtenons-nous ? Quelle classe a été définie comme modalité de référence ? (**TUTO 1**, pages 4 et 5).
9. Quels sont les propriétés de l'objet régression obtenu ?
10. Produisez l'objet « summary » de la régression (summary). Affichez-le. De quelles informations supplémentaires disposons-nous maintenant ?
11. Affichez les propriétés de l'objet « summary ».
12. Affichez la propriété « $coefficients » par exemple. Quel est le type d'objet associé ? (class).
13. Affichez les premières lignes de la propriété « $fitted.values ». De quoi s'agit-il ? Quelles sont les dimensions de la structure ? (dim) (500, 3).
14. Calculez les sommes par ligne de « $fitted.values ». Ce que vous obtenez est conforme à ce que l'on attendait (**TUTO 1**, page 5 ; « on peut vérifier que… ») ?
15. Nous souhaitons reproduire le calcul des probabilités d'affectation aux classes pour le premier individu à partir de sa description (variables centrées et réduites) et les coefficients des équations LOGIT.
    a. Calculer les LOGIT pour les modalités B et C (-4.089, -6.258).
    b. Passez-les à l'exponentielle (0.0167, 0.0019).
    c. Appliquez la fonction de transformation (**TUTO 1**, page 5 ; cf. exemple, page 7) pour obtenir les probabilités de B et C (0.0164, 0.00187).
    d. En déduire la probabilité de A (0.981).

16. Testez la significativité globale de la régression en opposant le modèle courant avec le modèle trivial composé uniquement de la constante.
    a. Construisez le modèle trivial (classe ~ 1).
    b. Calculez la statistique du rapport de vraisemblance LR en opposant les déviances ($deviance) (740.1113).
    c. Calculez les degrés de liberté via les degrés de liberté des régressions ($edf) (14).
    d. Reste alors à calculer la p-value du test (pchisq) (7.01E-149).
    e. Le modèle est-il globalement significatif à 5% ?
17. Testez la pertinence de chaque variable via les tests de significativité des coefficients c.-à-d. pour chaque variable à évaluer, ses coefficients sont-ils tous nuls dans l'ensemble des LOGIT ? Utilisez le test du rapport de vraisemblance (pour ma part, j'ai fait une boucle où j'ai retiré tour à tour les explicatives, et j'ai confronté la déviance avec le modèle initial comportant toutes les variables [LR(V05) = 63.76 ; LR(V07) = 29.80 ; LR(V10) = 23.18 ; etc.])

3.3 Prédiction et évaluation sur l'échantillon test

18. Chargez la seconde feuille du fichier « **waveform.xlsx** » qui fera office d'échantillon test. Affichez-en les caractéristiques (5000 obs., 8 variables).
19. Calculez et affichez les statistiques descriptives.
20. Centrez et réduire les variables explicatives avec les paramètres calculés sur l'échantillon d'apprentissage. Pourquoi devons-nous procéder ainsi ? Pour réaliser l'opération, regardez attentivement la documentation de scale.
21. Calculez et affichez de nouveau les statistiques descriptives. Que notez vous au niveau de la moyenne des variables notamment ? Pourquoi ?
22. Effectuez la prédiction sur les données transformées (predict ; regardez attentivement du côté de l'option « type »).
23. Affichez la distribution des prédictions (A : 1655, B : 1721, C : 1624).
24. Construisez et affichez la matrice de confusion :
    a. En déduire le taux de reconnaissance (taux de succès).
    b. Puis le taux d'erreur.
    c. Calculez le rappel (ou sensibilité) par classe (rappel de « A », de « B », de « C »).
    d. La précision par classe.

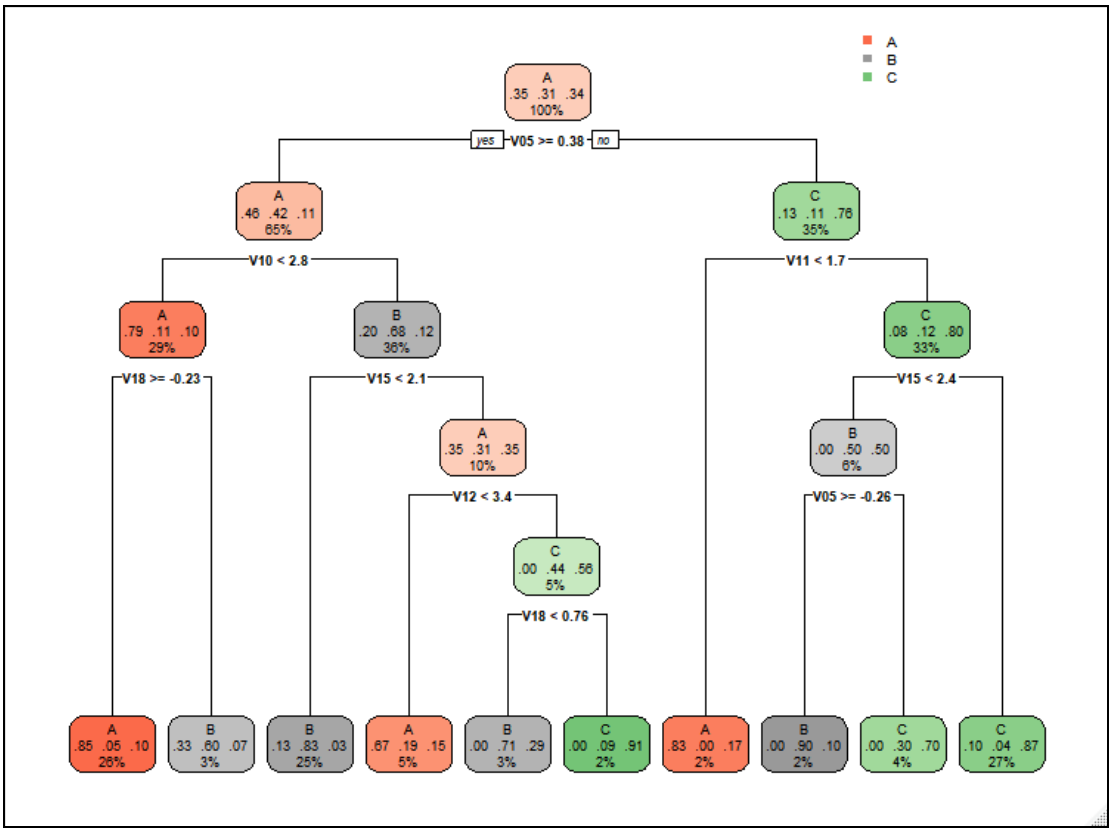
25. Pour vérifier vos résultats, faites appel à la fonction confusionMatrix() du package « caret » (qu'il faut éventuellement installer, puis charger).

3.4 Traitement avec le package « VGAM »

26. Nous souhaitons réitérer le traitement avec le package « VGAM ». Installez et chargez la librairie.
27. Réalisez la régression sur données centrées et réduites (vglm). Attention, vous devez porter une attention particulière à l'option « family » pour réaliser une régression multinomiale ou « A » est la modalité de référence. Affichez les résultats. Les coefficients obtenus sont-ils cohérents avec ceux de « nnet » ? (oui, heureusement)
28. Affichez le summary() de l'objet. De quelles informations supplémentaires disposons-nous ?
29. Réalisez la prédiction sur l'échantillon test (predictvglm, regardez attentivement du côté de l'option « type »).
30. Qu'obtenons-nous qui soit exploitable réellement ? (les probabilités d'affectation aux classes).
31. Comment convertir ces informations en classes prédites ? (pour chaque ligne, il faudrait trouver le maximum et assigner à la modalité correspondante).
32. Affichez la distribution des prédictions (A : 1655, B : 1721, C : 1624).
33. Calculez la matrice de confusion et les indicateurs de performances (utilisez directement « caret »). Que constatez-vous ? Notamment par rapport à la régression de « nnet » ?

3.5 Comparaison avec un arbre de décision

34. Les arbres de décision constitue une alternative possible pour la prédiction sur un problème multi-classes. Chargez la librairie « rpart » (elle est installée par défaut).
35. Sur les données d'apprentissage, lancez la modélisation où « classe » est la variable cible toujours (**TUTO 3**, page 9). Est-ce qu'il est nécessaire de travailler sur les données centrées et réduites ? (non) Pourquoi ?
36. Affichez l'arbre (print). L'arbre a produit combien de règles ? (10)
37. Pour disposer d'une visualisation plus sympathique, utilisez la fonction rpart.plot() du package « rpart.plot » (qu'il faut au préalable installer et charger). Combien de règles conduisent à la modalité « A » (« B », « C ») ? Combien d'opérations de segmentation distinguons-nous dans l'arbre ? (9)



38. Affichez les propriétés de l'objet « arbre » (attributes). Affichez le contenu de la propriété « $frame ». Qu'observez-vous ?

39. A partir de ces éléments, affichez la liste des variables qui apparaissent dans l'arbre.

40. Quelle variable, présente parmi les explicatives potentielles, n'est pas utilisée dans l'arbre ? (V07). Cette variable était-elle pertinente dans la régression logistique ? (oui) Comment expliquer cette apparente contradiction ? (voir du côté de summary peut-être…).

41. Effectuez la prédiction sur l'échantillon test (**TUTO 3**, page 10).

42. Calculez les performances prédictives. L'arbre est-il meilleur ou moins bon que la régression logistique sur les données « waveform » ? (nettement moins bon).