ISUP1 - Projets R **Année 2023-2024**

F.-X. Lejeune [1]

Projet R de Classification binaire avec gestion de valeurs manquantes et sélection de variables

---

Ce projet pourra se faire en binôme ou en trinôme et sera à me remettre via Moodle à partir du 13 Mai 2024 (date à préciser) avec 3 fichiers :

- *votre rapport sous la forme d'un diaporama au format pdf (maximum de 20 diapositives incluant les tables et figures),*
- *vos scripts R dans un fichier .R ou .Rmd,*
- *votre fichier de données au format .Rdata.*

La note prendra largement en compte la qualité de présentation du jeu de données, des analyses avec R et des figures, ainsi que les explications sur les sorties et scripts R d'analyses. Vous citerez les bibliothèques R et les sources (livres, articles, sites web, etc.) utilisées à la fin du rapport.

1. lejeufra@yahoo.fr

1 Timeline du projet

☞ Séance 1 - Lancement des projets : Introduction du projet, de la régression logistique et initiation à R Shiny

Mercredi 31/1 - 13h45-15h45 - Salle 14/15 103

☞ Séance 2 - Suivi de projet, Q/R sur le 1er objectif

Mercredi 14/2 - 14h-17h - Salle 55/65 102

☞ Séance 3 - Suivi de projet

Mercredi 27/3 - 13h45-15h45 (G2) & 16h-18h (G1) - Salle 13/14 108

☞ Séance 4 - Présentations orales des analyses via une interface graphique R Shiny (20 min/groupe, ordre de passage à déterminer plus tard)

Lundi 6/5 - 9h-12h & 14h-17h - Salle 23/24 207

☞ Remise des diaporamas et scripts R d'analyse

Semaine du 13/5 (date précise à déterminer)

Note finale = Présentation orale + Diaporama/scripts R

2 Description

Ce projet R vise à explorer la classification binaire par une approche de régression logistique multivariée dans une situation faisant intervenir des valeurs manquantes (VMs) dans les variables explicatives.

Dans ce projet, vous aurez à rechercher un tableau de données complètes se prêtant au problème avec une variable facteur Y à 2 niveaux indiquant l'appartenance des individus à 2 groupes et un ensemble de plusieurs variables explicatives X susceptibles d'expliquer l'appartenance des individus à ces groupes (au moins 5 variables, toutes quantitatives).

Pour nous compliquer un peu plus la tâche, nous supposerons la présence de VMs disséminées dans les variables X (celles-ci seront en fait générées de manière aléatoire par nos soins pour 2 taux de VMs fixés à 5%, 10% et 15%). La présence de VMs peut être fréquente en pratique. Elle est généralement un problème avec les méthodes classiques de régression, où en l'absence de stratégie pour combler (imputer) ces VMs, le logiciel doit retirer entièrement les observations/individus comportant au moins une VM (malgré les autres valeurs disponibles). Pour explorer les solutions possibles, le projet propose d'étudier 2 stratégies d'imputation (imputation simple par la moyenne et imputation basée sur l'approche des k plus proches voisins) et de les comparer.

Une fois notre stratégie de gestion des VMs fixée, le problème de classification sera traité en utilisant la régression logistique, combinée à une procédure de sélection de variables « pas à pas », en considérant le tableau complet, les tableaux à valeurs manquantes et les tableaux imputés.

Les performances de classification binaire seront évaluées avec les critères usuels de sensibilité, de spécificité, de précision et d'AUC (aire sous la courbe ROC) sur un échantillon de test. Les données du jeu de données choisi et leurs distributions seront présentées via une interface interactive développée en R Shiny.

3 Régression logistique binaire

En régression logistique binaire [1], la variable de sortie Y est une variable catégorielle à deux niveaux notés $\{0, 1\}$. Étant donné une variable Y binaire que l'on cherche à l'aide de p variables explicatives $X = (X_1, \dots X_p)^T$, on suppose alors que :

$$\mathbb{P}(Y = 1|X = \mathbf{x}) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)} = \frac{\exp(\beta_0 + \beta^T \mathbf{x})}{1 + \exp(\beta_0 + \beta^T \mathbf{x})}$$

où β_0 est la constante (intercept) et $\beta = (\beta_1, \dots, \beta_p)^T$ le vecteur des coefficients appliqués aux p variables.

3.1 Présentation du jeu de données

Questions

Q1. Choix et présentation du jeu de données pour le projet (à trouver vous-même ou dans les suggestions en références en fin de sujet). La base de données devra se prêter à un problème de classification binaire sur une variable cible catégorielle (Y) à 2 niveaux. Pour la suite du projet, il est préférable d'éviter autant que possible d'avoir un trop grand déséquilibre d'effectifs entre les 2 niveaux de Y. Considérer pour le projet un nombre de variables explicatives du jeu de données entre 5 minimum et 10 maximum, toutes de type numérique (quantitative). Indiquer aussi la source du jeu de données utilisé (référence bibliographique, package R, site web, etc.). Pour votre jeu de données, indiquer quelle serait une hypothétique question d'« intérêt pratique » à traiter par cette analyse de classification.

Q2. En utilisant la fonction `table1` du package furniture (à télécharger sur le CRAN), donner les principales statistiques descriptives des variables numériques X en fonction des 2 niveaux de la variable Y. Ajouter les résultats des tests sous la forme de p-valeurs en indiquant le test utilisé par la fonction. Conclure sur les tests.

3.2 Présentation des méthodes

Questions

Q3. Expliquer brièvement le principe de la régression logistique binaire : fonction de lien, estimation des valeurs β, prédiction des valeurs de Y, critères indiquant les qualités discriminantes du modèle (c'est-à-dire la capacité du modèle à « bien reclasser » les unités statistiques dans leur groupe d'appartenance), etc.

Q4. Dans le cadre de la régression logistique binaire, les qualités du classifieur peuvent s'évaluer à l'aide de critères numériques tels que le taux de bonnes prédiction (accuracy), la sensibilité et la spécificité. Expliquer brièvement ces 3 critères.

Q5. Les qualités du classifieur peuvent aussi s'évaluer graphiquement à l'aide d'une courbe ROC et d'une valeur d'AUC. Expliquer brièvement comment est construite la courbe ROC, sa lecture graphique en lien avec les critères numériques de la question 4, et son interprétation basée sur l'AUC avec le critère de discrimination donné par Hosmer et Lemeshow.

3.3 Génération de valeurs manquantes

Question

Q6. À partir du tableau de données original, générer des valeurs manquantes de type « MCAR »(missing completely at random), supposant que la probabilité d'absence est la même pour toutes les valeurs du tableau. On considérera deux situations avec des taux de 5%, 10% et 15% de valeurs manquantes dans l'ensemble des valeurs du tableau de données. Expliquer les lignes de commande R utilisées pour créer les valeurs manquantes. Indiquer aussi le nombre d'individus concernés par ces données manquantes pour les 3 tableaux générés (5%, 10% et 15%).

On continuera avec les 4 tableaux de données complètes et manquantes (tableau avec 5%, 10% et 15% de données manquantes retirées aléatoirement).

3.4 Imputation des valeurs manquantes

Mise en œuvre des techniques d'imputation simple (par la moyenne) et d'imputation par la méthode des k plus proches voisins (k-NN, k-Nearest Neighbors). La qualité de l'imputation peut s'évaluer de plusieurs manières, la plus directe étant de comparer les valeurs imputées aux vraies valeurs (initialement présentes dans le tableau complet). Par exemple, par des critères de distance mesurant l'écart entre les valeurs imputées et les vraies valeurs, tels que l'erreur quadratique moyenne (root mean square error, RMSE), l'erreur absolue moyenne (mean absolute error, MAE) et le coefficient de détermination (R2).

Questions

Q7. Expliquer brièvement le principe des 2 techniques d'imputation proposées (moyenne et k-NN) et indiquer des bibliothèques et fonctions de R dédiées pour les mettre en œuvre. Décrire brièvement les 3 critères de qualité et leur calcul avec R.

Q8. Comparer les 2 approches d'imputation (moyenne et k-NN) avec les 3 critères de distance RMSE, MAE et R2.

Q9. En répétant plusieurs fois l'étape de génération de valeurs manquantes, par exemple 100 fois et toujours avec les taux de 5%, 10% et 15%, vous pouvez recalculer 100 valeurs des 3 critères pour chaque approche pour obtenir une valeur moyenne et un écart type sur l'ensemble des répétitions. Quel serait selon vous l'objectif d'une telle démarche ? Donner pour votre tableau de données ces valeurs de moyennes et écarts-types. Dire quelle méthode vous parait la meilleure au sens de ces critères.

On continuera avec les 7 tableaux de données complètes, manquantes et imputées par la méthode de votre choix (avec 5%, 10% et 15% de données manquantes et imputées).

☞ Les questions 1 à 9 constitueront l'objectif à compléter pour la Séance 2 de suivi de projet.

3.5 Régression logistique (glm)

Dans la suite, on utilisera 80% des observations des 7 tableaux de données complètes, manquantes et imputées comme données d'apprentissage pour ajuster le modèle logistique, les 20% d'observations restantes serviront après de jeu de données test pour évaluer les qualités discriminantes du modèle construit.

Questions

Q10. En utilisant les données d'apprentissage : donner les estimateurs du maximum de vraisemblance des paramètres $\beta = (\beta_1, \ldots, \beta_p)^T$ fournis par la fonction `glm`.

Q11. Indiquer quels sont les coefficients $\hat{\beta}$ estimés significativement différents de zéro. Interpréter alors ces coefficients $\hat{\beta}$ en termes d'odds-ratio. Donner les valeurs d'odds-ratio avec leurs intervalles de confiance.

À l'aide du modèle, on peut aussi estimer pour chaque individu sa probabilité prédite $\hat{\mathbb{P}}(Y = 1|X = x_i)$ ou probabilité a posteriori d'appartenance à la classe $Y = 1$ lorsque X vaut x_i.

Questions

Q12. Calculer les probabilités prédites à l'aide de la fonction `predict.glm` appliquée à l'échantillon d'apprentissage. Dans un premier temps, on choisit d'affecter les individus à la classe $Y = 1$ si la probabilité estimée est supérieure à 0.5 et à la classe $Y = 0$ sinon pour obtenir les prédictions de classe. Donner alors pour ce seuil de 0.5 la matrice de confusion croisant les niveaux 0/1 observés et prédits de Y par le modèle et indiquer le taux global de bonnes prédictions (accuracy) et les taux de bonnes prédictions obtenus pour chacun des 2 niveaux de Y (valeurs de sensibilité et de spécificité).

Q13. Donner la courbe ROC et la valeur AUC obtenue avec le jeu de données d'apprentisage. Donner la valeur de seuil « optimale » des probalités prédites maximisant à la fois la sensibilité et spécificité (méthodes de « Youden » et « closest.topleft » du package pROC à télécharger sur le CRAN). Donner alors les valeurs de sensibilité et de spécificité obtenues avec ce seuil. Conclure.

L'évaluation du classifieur directement sur les données d'apprentissage (celles ayant servi à la construction du modèle) fournit généralement un aperçu trop optimiste de ses qualités discriminantes. En pratique, on préfère baser cette évaluation sur le jeu de données test (n'ayant pas contribué à la construction du modèle). Quand cela est possible, un jeu de données complètement indépendant (autre expérience ayant produit des données similaires) peut servir aussi à évaluer la capacité de généralisation du modèle (hors projet).

Questions

Q14. En utilisant maintenant la fonction `predict.glm` sur l'échantillon test, redonner pour la valeur de seuil 0.5 la matrice de confusion croisant les niveaux 0/1 observés et prédits de Y. Que valent maintenant les valeurs d'accuracy, de sensibilité et de spécificité ?

Q15. Idem question 13. Conclure.

3.6 Sélection du meilleur modèle par méthodes pas-à-pas

Les méthodes pas-à-pas visent à rechercher le meilleur sous-ensemble de variables explicatives à l'aide d'un critère donné, par exemple le critère d'information d'Akaike (AIC) que l'on souhaite minimiser. L'utilisation de ces méthodes peut se faire par élimination (méthode descendante de type « backward ») ou ajout (méthode ascendante de type « forward ») successif des variables, la méthode « stepwise » étant une combinaison des deux précédentes méthodes (direction = "both").

Questions

Q16. Appliquer les trois méthodes de sélection à vos modèles logistiques et indiquer dans chaque cas le meilleur sous-ensemble de variables avec la valeur d'AIC associé retenu par la procédure. Y a-t-il alors consensus dans les sélections opérées par les 3 approches ? En cas de non-consensus, garder pour la question suivante la sélection correspondant au plus petit AIC obtenu.

Q17. Idem questions 13 et 15. Conclure.

4 Application en R Shiny

Question

Q18. Mettre en œuvre une application en R Shiny prenant en paramètres d'entrée : une ou plusieurs variables numériques ou catégorielles de la base de donnée, une variable catégorielle de la base de données servant de variable de groupes des individus ; et retournant en sorties une table des statistiques descriptives et des graphiques appropriés des distributions des variables en fonction de la variable de groupe.

☞ Les questions 1 à 17 constitueront l'objectif à compléter pour la Séance 3 de suivi de projet.

Présentation orale (20 min/groupe) des analyses et de votre interface R Shiny
Lundi 6/5 - 9h-17h

Remise des diaporamas et scripts R d'analyse
Semaine du 13/5 (date précise à déterminer)

Note finale = Présentation orale + Projet écrit

Bon courage !!!

5 Références

5.1 Liste non exhaustive de datasets à choisir pour le projet (1 par trinôme)

Le dataset de votre choix devra nécessairement comporter une variable Y catégorielle à 2 niveaux, la variable cible, sur laquelle on cherchera à entraîner le modèle de régression logistique avec un ensemble de variables X quantitatives (au moins 5 variables). La liste suivante de datasets est donnée à titre d'exemple.

- Breast Cancer Wisconsin (Diagnostic) Dataset
  `https://www.kaggle.com/uciml/breast-cancer-wisconsin-data`
- Spam Dataset
  `http://search.r-project.org/library/kernlab/html/spam.html`
  `https://www.kaggle.com/monizearabadgi/spambase`
- Pima Indians Diabetes Database
  `https://www.kaggle.com/uciml/pima-indians-diabetes-database`
- Sonar Dataset Mines vs. Rocks
  `https://www.kaggle.com/ypzhangsam/sonaralldata`
- BankNote Authentication Dataset
  `https://www.kaggle.com/shantanuss/banknote-authentication-uci`
- Churn Dataset
  `https://search.r-project.org/CRAN/refmans/liver/html/churn.html`
- Etc.

5.2 Tutoriels d'initiation à R Shiny

- `https://shiny.rstudio.com/tutorial/`
- `http://perso.ens-lyon.fr/lise.vaudor/Tuto_Shiny/tuto_Shiny_fr_Part1.html#(7)`
- `https://geoviz.sciencesconf.org/data/pages/GeoViz2018_R_shiny.pdf`

Références

[1] David W. Hosmer and Stanley Lemeshow. *Applied logistic regression.* John Wiley and Sons, 2000.