

Le jeu de données « telecat.csv » comporte les résultats d'une enquête de satisfaction menée auprès de 150 clients d'une chaîne câblée.

Pour chaque client un score de satisfaction a été établi, noté Y, sur une échelle qui va de moins satisfait (valeurs négatives) à plus satisfait (valeurs positives).

Les variables explicatives étudiées sont les temps passés sur différentes chaînes, combinés aux nombres de visites sur ces chaînes. Ces covariables ont été normalisées. Il y en a p=160.

Les 160 chaînes étudiées sont :

20 chaînes proposant des films, notées « Film »,
20 chaînes proposant des séries, notées « Serie »,
20 chaînes de sport, notées « Sport »,
20 chaînes orientées science/santééconomie, notées « Science »,
10 chaînes d'actualité/politique, notées « Actu »,
20 chaînes musicales, notées « Music »,
20 chaînes de jeux, notées « Jeux »,
10 chaînes d'histoire/géographie/documentaire, notées « Hist »,
20 chaînes diverses.

Il y a aussi une variable « Sexe » qui détermine si l'individu est un homme (codé 1) ou une femme (codé 0).

On veut expliquer le score des abonnés par rapport aux scores des chaînes. Plus précisément on veut pouvoir prédire des scores, et on veut également sélectionner les chaînes ayant le plus d'influence sur le score.

Pour cela on va mettre en oeuvre trois méthodes :

une régression aléatoire (Random Regression),
une régression bayésienne de type A,
une régression LASSO bayésienne,
une régression SSVS (Stochastic Search Variable Selection).

Tout d'abord nous allons découper en deux l'échantillon pour avoir un jeu d'apprentissage « training » qui servira à construire les modèles et un jeu « test » qui servira à comparer nos modèles. Vous découperez ainsi l'échantillon en deux aléatoirement (avec l'instruction *sample* après avoir utilisé une clé *seed*) :

- 100 observations pour le jeu de données training
- 50 observations pour le jeu de données test.

Vous indiquerez en début de rapport la clé (seed) utilisée pour découper l'échantillon !!!

Dans toute la suite vous pourrez utiliser une méthode visuelle (seuillage), ou le boxplot, ou un intervalle de confiance a posteriori, pour sélectionner les meilleures covariables.

1. RR-BLUP

- 1.1 Estimation – Donner ou représenter les estimations des paramètres aléatoires du modèle obtenues sur le training.
- 1.2 Prédiction - A l'aide des estimations obtenues, calculer la corrélation des prédictions pour les observations du jeu de validation (test). Que peut-on en conclure ?
- 1.3 Sélection - A l'aide des estimations obtenues, sélectionner les variables explicatives qui paraissent les plus pertinentes (on pourra utiliser un critère basé sur le boxplot).

2. Bayes A
 - 2.1 Quelle est la différence (avantage ?) entre le modèle Bayes A et le modèle Random Regression ?
 - 2.2 A l'aide de la fonction BayesA vue en TD, donner ou représenter les estimations des paramètres du modèle. Comment sont-elles obtenues ?
 - 2.3 Pourquoi vaut-il mieux prendre une taille de burn in grande ?
 - 2.4 Examiner les traces de quelques paramètres après burn in (on pourra prendre seulement 500 simulations après burn in). Que nous peuvent nous indiquer ces trajectoires ?
 - 2.5 Proposez un choix d'hyper-paramètres qui auraient peu d'influence sur le modèle. Pourquoi ?
 - 2.6 A l'aide des estimations obtenues, donner la corrélation des prédictions avec les observations du jeu de validation et comparer cette corrélation à celle précédente du Random Regression.
 - 2.7 A l'aide des estimations obtenues, sélectionner les variables explicatives qui paraissent les plus pertinentes. On pourra comparer les meilleures variables retenues ici par rapport aux meilleures retenues avec RR-BLUP.

3. LASSO bayésien
 - 3.1 Quelle est la différence entre le LASSO bayésien et l'approche Bayes A ?
 - 3.2 Quel rôle joue le paramètre lambda ? Comment choisir lambda pour diminuer ou rétrécir (effet de shrinkage) les coefficients ?
 - 3.3 Donnez (ou représentez) les estimations du modèle à partir d'hyper-paramètres choisis autour de 1 et 2 (pourquoi ?).
 - 3.4 Comparez quelques distributions a posteriori à celles a priori.
 - 3.5 Prédiction – Donnez la corrélation des prédictions avec les observations du jeu de validation et comparez aux valeurs précédentes (obtenues avec RR et Bayes A).
 - 3.6 Sélectionnez les variables explicatives qui paraissent les plus pertinentes.
 - 3.7 Comparez les meilleures variables à celles obtenues avec les méthodes précédentes.
 - 3.8 Qu'est-ce qui pourrait expliquer la différence entre les variables retenues par le LASSO Bayésien et celles retenues par Bayes A (encore appelé le ridge bayésien).

4. SSVS
 - 4.1 Est-ce qu'on utilise un Metropolis-Hasting ou un Gibbs sampler dans notre approche SSVS ? Quelle est la loi (appelée « proposal ») qui choisit le nouveau sous-ensemble de variables ?
 - 4.2 En utilisant la fonction « selection_SSVS » vu en TP, sélectionnez les variables les plus pertinentes du modèle.
 - 4.3 Faites varier quelques hyper-paramètres pour voir l'influence sur la sélection (on pourra ainsi sélectionner plus ou moins de variables grâce aux a priori).

5. Comparaisons
 - 5.1 Classez les quatre approches précédentes en deux grands groupes de méthodes.
 - 5.2 En examinant les résultats obtenus avec RR-Blup, Bayes A et LASSO bayésien, y a-t-il une méthode où l'effet « shrinkage » (rétrécissement) est plus prononcé, c'est-à-dire, où les coefficients sont plus faibles ?

- 5.3 A l'aide des corrélations entre valeurs prédictes et observées, comparez ces trois méthodes.
- 5.4 Comparez les variables retenues avec les quatre méthodes : RR-BLUP, LASSO bayésien, Bayes A et SSVS.
- 5.5 Comparer les intervalles de confiance a posteriori des paramètres les plus importants.
- 5.6 Utilisez une méthode de régression pénalisée non bayésienne pour sélectionner des variables (par exemple LASSO, ou Ridge ou ElasticNet).
- 5.7 Comparez avec les résultats précédents.
- 5.8 Finalement, proposez un modèle de régression «standard» avec quelques variables issues des précédentes sélections bayésiennes.
- 5.9 Analysez le résultat de la régression obtenue (p-value, effet des variables, résidus, ...).
- 5.10 En conclusion, que pouvez-vous proposer à la chaîne de programmes pour améliorer les scores de satisfaction ?

Pour la suite, vous devez choisir :

- soit faire les questions 6-11,
- soit faire la 12,
- soit faire une approche avec Stan (le modèle linéaire, ou logistique en divisant Y en deux catégories, ou Poisson en regroupant les valeurs de Y comme des entiers).

6. Ajout de la variable Sexe.

Utilisez le package BLR en ajoutant la variable Sexe (vous pourrez construire une matrice indiquant le sexe de l'individu) et essayez d'analyser son effet.

7 Compléments sur SSVS

7.1 Dans l'algorithme SSVS, si on prend la probabilité P_i (la probabilité a priori de chacune des variables d'être sélectionnées) trop grande et si on prend trop de variables initiales non nulles, on risque de rencontrer un problème d'inversion de matrice. Lequel ? Dans quelle loi a priori ?

7.2 Que pourrait-on ajouter comme paramètre pour pallier ce problème d'inversion ?

7.3 Ecrire le schéma hiérarchique bayésien qui en découle.

8. Complément de modélisation

8.2 Proposez un modèle bayésien de régression de Poisson pour une variable Y de comptage.

8.2 Donnez le schéma bayésien qui en découle.

8.3 A quel endroit pourrait-on introduire un paramètre gamma de sélection de type SSVS ?

9. Elastic Net (question optionnelle)

9.1 Proposez un schéma hiérarchique bayésien Elastic Net à partir du package EBLglmnet.

9.2 Appliquez ce modèle sur le jeu de données précédent.

10. Modèle de régression bayésienne : BAYES A

10.1 Notons $\beta(j)$ la jème composante du vecteur de coefficient β . Dans le modèle Bayes A, complétez la partie du cours dans l'algorithme de Gibbs en montrant que la loi de la variance de $\beta(j)$ sachant $\beta(j)$ est une inverse gamma (dont on précisera les paramètres).

10.2 De la même manière, montrez que la loi de la variance de ϵ sachant (Y, β, μ) est aussi une inverse gamma (dont on précisera également les paramètres).

11. Complément sur les probabilités de sélection en SSVS

Dans l'approche SSVS, proposez une modification de l'algorithme pour accélérer la sélection des variables significatives.

12. Approche ABC (Approximate Bayesian Computation)

Reprendre le programme ABC étudié en TP pour simuler la moyenne et la variance a posteriori. Et transformer le critère d'acceptation par un test d'adéquation de la loi observée à la loi simulée.