Régression pénalisée :logistique et Poisson

15-16 septembre 2025

MSP *2025-2026*

Logistique pénalisée :

Le jeu de données telelogit.txt comporte les résultats d'une enquête de satisfaction menée auprès de 150 clients d'une chaîne câblée. Pour chaque client, on relève la satisfaction (1 satisfait, 0 non satisfait), notée Y.

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/santé/économie, 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.

Pour essayer d'expliquer la satisfaction des abonnés par rapport aux scores des chaînes, on vous demande de réaliser :

1- Une régression logistique Ridge sur ce jeu de données.

2- Une régression logistique Lasso sur ce jeu de données.

3- Une régression logistique ElasticNet sur ce jeu de données.

4- Comparer les résultats obtenus.

5- Réaliser des régressions logistiques sur les variables sélectionnées par les précédentes méthodes.

6- Comparer les résultats de ces dernières régressions.

Group Lasso

L'objectif de cette partie est de se familiariser avec la régression Group Lasso. On considère un jeu de données avec $n = 100$ observations et $p = 500$ variables réparties en $\frac{p}{d} = 100$ groupes de taille $d = 5$ chacun. On va simuler les variables de chaque groupe par une loi normale centrée de matrice de variance-covariance V, avec 1 sur la diagonale et ρ ailleurs. Deux variables du même groupe sont donc corrélées de la même manière, avec une corrélation égale à ρ. On suppose que les groupes ne sont pas corrélées. Les 500 variables

sont donc simulées suivant une loi normale centrée de matrice de variance-covariance bloc-diagonale, avec des blocs $d \times d$ égaux à V. Nous allons avoir besoin de charger les librairies (mnormt, grplasso, gglasso, glmnet).

On va considérer deux situations :

- Cas 1 : l'indice de sparsité (le nombre de vraies variables) est égal à 10. Et les 10 vraies variables sont réparties dans 10 groupes différents.
- Cas 2 : l'indice de sparsité vaut 50 et ces 50 vraies variables sont réparties dans 10 groupes.

7- On demande de réaliser que régression Group Lasso dans les deux cas précédents. On regardera notamment le pourcentage de variables vraies retenues. Et le pourcentage de vrais groupes retenus.

8- On comparera les résultats précédents à ceux obtenus avec un LASSO sans spécifier de groupes. On notera notamment le nombre de vraies variables retenues.

Retour au jeu de données télé

9- Charger le jeu de données tele.txt à l'origine du jeu de données telelogit.txt avant binarisation de la variable Y et proposer une analyse en Group LASSO.

Régression pénalisée sur données de compatages

Le jeu de données achat.txt contient des résultats liés à une étude portant sur des achats de produits via internet. On a relevé le nombre mensuel Y d'achats de produits appartenant à une certaine marque distribuée sur le web. Pour chaque individu dont on a suivi les achats, on a relevé également le nombre de visites sur des sites proposant des publicités de cette marque et le temps passé sur ces sites durant le mois.

- On note Y_i le nombre d'achats par l'individu i sur le mois d'étude.
- x_i le vecteur du nombre de visites/mois sur les sites présentant les publicités concernées.
- z_i le vecteur du temps passé par mois sur les sites présentant les publicités concernées.
- $n = 100$ individus ont été observés (avec accord préalable, anonymat, CNIL, etc. . .).

Les résultats obtenus sont dans le fichier achat.txt. On vous demande de trouver des liens entre les visites sur les différents sites, les temps passés sur les sites, et les achats des produits de la marque.

10- Faire une régression de Poisson (ou binomiale négative) ridge sur ce jeu de données

11- Faire une régression de Poisson (ou binomiale négative) Lasso

12- Faire une régression elasticnet

13- En retenant les variables avec les modèles précédents, comparer les MSE d'un modèle linéaire généralisé.