

# Big Data

## TD/TP 6 : Régression Pénalisée

### BUT 3

Guillaume Metzler et Antoine Rolland  
Institut de Communication (ICOM)  
Université de Lyon, Université Lumière Lyon 2  
Laboratoire ERIC UR 3083, Lyon, France

[guillaume.metzler@univ-lyon2.fr](mailto:guillaume.metzler@univ-lyon2.fr); [antoine.rolland@univ-lyon2.fr](mailto:antoine.rolland@univ-lyon2.fr)

Dans cette dernière fiche, on s'intéresse à la régression linéaire multiple gaussienne et nous allons, plus précisément, étudier une méthode de pénalisation dite **Lasso** dans le cadre de la régression et la comparer à de la régression Ridge.

## Régression

Etant donnée une matrice de données  $\mathbf{X} \in \mathcal{M}_{m,p}(\mathbb{R})$ , un vecteur  $\boldsymbol{\beta} \in \mathbb{R}^p$  et vecteur d'observations  $\mathbf{y}$  vérifiant le modèle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où  $\boldsymbol{\varepsilon}$  est un  $m$ -échantillon gaussien de moyenne nulle et de variance  $\sigma^2$ .

1. On considère la valeur  $p = 20$  et  $m = 100$  dans un premier temps. Simuler une matrice  $\mathbf{X} \in \mathcal{M}_{m,p}(\mathbb{R})$  dont les coordonnées suivent une loi uniforme  $\mathcal{U}([-5, 5])$  de façon indépendante. Simuler un vecteur  $\boldsymbol{\beta}$  dont les entrées sont soit 0 soit 1 suivant une loi  $\mathcal{B}(0.1)$ . Simuler un vecteur  $\mathbf{y}$  à l'aide de la relation précédente, en prenant  $\sigma^2 = 1$ .
2. Estimer les paramètres du modèle à l'aide d'un modèle linéaire classique (fonction LM).
3. Estimer les paramètres du modèle à l'aide du package GLMNET. Identifier les coordonnées non nulles dans la nouvelle estimation des paramètres.
4. Refaire la même expérience avec  $m = 100$  et  $p = 200$  et décrivez vos observations.
5. Pour chaque paramètre  $\lambda$  fourn par la fonction GLMNET, calculer le faux de vrais positifs (nombre moyen de valeurs estimées non nulle à raisons) et de faux positifs (le nombre moyen de valeurs estimées non nulle à tort). Construisez le nuage de point avec en abscisse le taux de faux négatif et en ordonnée, le taux de faux positifs. Cette courbe, si on relie les points, est appelée la **courbe ROC** (*receiver operating characteristic*). À votre avis, quel type de courbe signifie que la procédure marche bien ? C'est-à-dire qu'elle détecte bien les vrais positifs et pas trop les faux négatifs.

## Classification

On pourra appliquer cette méthode de pénalisation à un jeu de données de classification. Votre objectif sera alors semblable à la section précédente, mais on cherchera à établir un modèle de *régression logistique pénalisée* avec de bonnes capacités de généralisation.