

Challenge UP2 - Méthodes de Régression Avancées

On dispose d'un échantillon statistique de taille $n = 100$ d'une variable y et de variables $x^{(1)}, \dots, x^{(p)}$ avec $p = 100$. Ces données sont dans le fichier '**data.txt**' (données d'apprentissage).

Le but est de construire le meilleur modèle prédictif de la variable y à partir des variables $x^{(1)}, \dots, x^{(p)}$ sachant qu'il faudra au final calculer et retourner les prédictions associées au fichier '**Xtest.txt**' de taille 100×100 (100 prédictions à réaliser, une par ligne, cf. dernière étape ci-dessous).

Le critère utilisé pour évaluer ce challenge est le critère usuel (**Root Mean Square Error**) :

$$\text{RMSE} = \sqrt{\text{MSE}} \quad \text{où} \quad \text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2$$

où $n_{\text{test}} = 100$ désigne le nombre de prédictions à réaliser, y_i les valeurs réellement observées pour y et \hat{y}_i vos prédictions. On pourra s'aider au début des indications suivantes :

0. Charger uniquement le fichier de données '**data.txt**' qui sera utilisé pour mettre au point un modèle de prédiction.
1. Calculer l'écart-type de la variable y comme premier **RMSE de référence** qui consisterait à prédire par la moyenne de y (prédiction constante qui ne tient compte d'aucun prédicteur).
2. Envisager différentes techniques de régression : RLM, ridge, LASSO, PCR, PLS, pas à pas (stepwise), etc.
3. Charger enfin le fichier '**Xtest.txt**' et calculer les prédictions correspondant à la technique retenue.

On déposera sur Campus le fichier texte associé à vos prédictions (fichier comportant donc une seule « colonne » formée de vos 100 prédictions). Mettre ce fichier au format **NOM.txt**. On donne un script R qui calcule de telles prédictions pour le prédicteur constant de la question 1.

☞ **Date limite de dépôt sur Campus : vendredi 29 novembre (au soir)**