FTML 2025 – Exercice 4

Régression supervisée sur données réelles

Objectif de l'exercice

L'objectif de cet exercice est de résoudre une tâche de régression supervisée à partir d'un jeu de données réel. Deux modèles doivent être comparés :

- Régression linéaire (moindres carrés ordinaires, OLS)
- Régression Ridge (régularisation L2)

L'évaluation repose sur les métriques classiques : **MSE (erreur quadratique moyenne)** et **MAE (erreur absolue moyenne)**, et une **analyse qualitative des prédictions** sur les données de test.

Jeu de données

Les fichiers fournis sont :

- `X_train.npy`, `y_train.npy`
- `X_test.npy`, `y_test.npy`

Dimensions des matrices :

- $X_{\text{train}} \in \mathbb{R}^{200\times200}$ ($n = 200$ observations, $d = 200$ variables)
- $X_{\text{test}} \in \mathbb{R}^{200\times200}$

Les variables sont centrées autour de 0. La variable cible y est continue et suit une distribution relativement normale comme illustré ci-dessous.

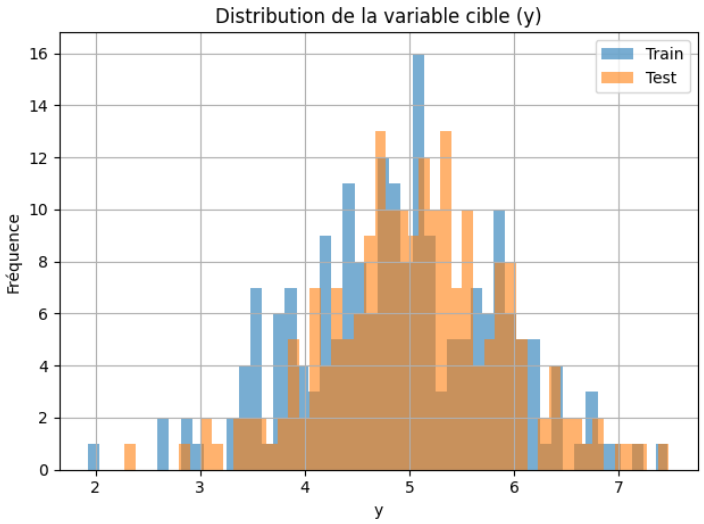

Figure 1: Distribution de la variable cible y pour l'ensemble train et test

Méthodologie

Modèles entraînés

- **OLS** : entraîné via `LinearRegression()` de `sklearn`.
- **Ridge** : entraîné avec validation croisée sur les hyperparamètres `alphas = [0.1, 1.0, 10.0]` via `RidgeCV()`.

La régression Ridge consiste à minimiser la fonction suivante :

$$\min_{\beta} \|X\beta - y\|^2 + \lambda\|\beta\|^2$$

où λ est un hyperparamètre de régularisation contrôlant la pénalisation sur la norme des poids. Plus λ est grand, plus la solution est contrainte. Le modèle Ridge a sélectionné automatiquement $\lambda^* = 1.0$ via validation croisée.

Évaluation des performances

Les prédictions ont été comparées sur l'ensemble de test à l'aide de :
- MSE : $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- MAE : $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

Résultats numériques

Modèle	**MSE**	**MAE**
OLS	8.0383	2.3423
Ridge	0.2090	0.3643

Table 1: Performances des deux modèles sur l'ensemble de test

Analyse : Le modèle Ridge surperforme très largement OLS. Cela montre l'importance de la régularisation sur ce jeu de données de haute dimension ($d = 200$) où la régression non pénalisée conduit à un surapprentissage et une forte variance.
On observe une **variance très élevée** du modèle OLS : malgré un ajustement parfait sur l'ensemble d'apprentissage, ses prédictions sur le test sont très dispersées, ce qui se traduit par un MSE élevé et une MAE plus de 6 fois supérieure à Ridge.

Comparaison visuelle

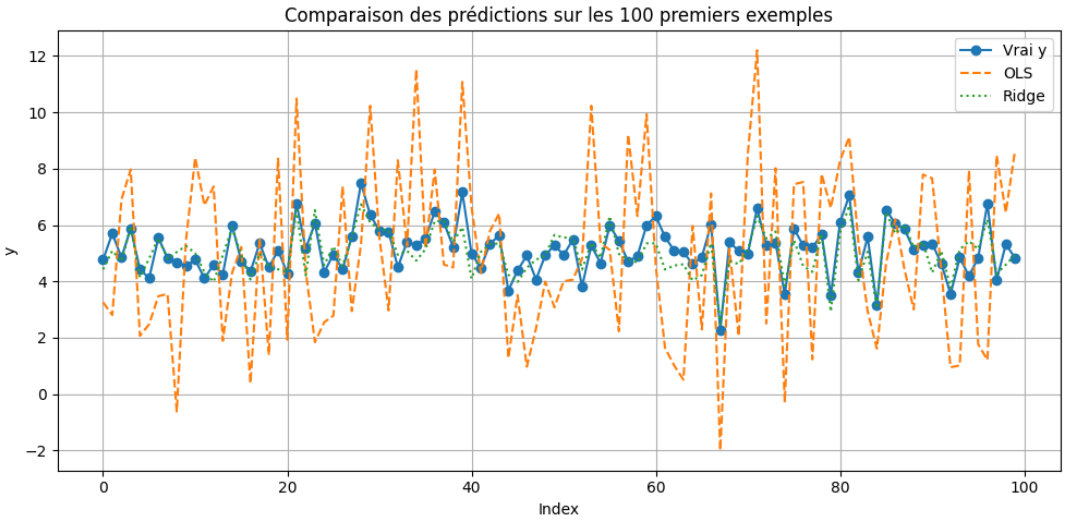

Figure 2: Comparaison des prédictions sur les 100 premiers exemples de test

L'OLS génère des prédictions très bruitées avec de fortes amplitudes d'erreur, tandis que Ridge reste plus proche de la vérité.

Conclusion

Cet exercice illustre concrètement la puissance de la régularisation L2 dans des contextes à haute dimension. La régression Ridge démontre une forte capacité de généralisation en limitant les oscillations dues au surapprentissage, contrairement à l'OLS.

- **OLS** échoue à généraliser dans un cadre $n \approx d$ à cause de sa variance élevée.
- **Ridge** impose une contrainte sur la norme des poids, limitant les coefficients aberrants.
- Le modèle Ridge sélectionne ici automatiquement $\lambda^* = 1.0$.
- Ce cas confirme l'importance des méthodes pénalisées en machine learning réel.