Thierry Fotchou Saadem

MINI-PROJET

Des données synthétiques ($n = 20$ observations) simulant une relation linéaire entre une variable réponse y et $p = 20$ variables explicatives ont été générées afin de comparer la performance prédictive de deux méthodes de régression réguliarisée : Ridge (pénalisation L2) et Lasso (pénalisation L1). La matrice des prédicteurs $X \in \mathbb{R}^{n \times p}$, est constituée de variables aléatoires indépendantes suivant une loi normale standard $\mathcal{N}(0, 1)$. Le vecteur de coefficients $\beta \in \mathbb{R}^p$ est défini de manière parcimonieuse : seuls les $k = 8$ premiers coefficients sont non nuls, tirés selon une loi $\mathcal{N}(0, 1)$, les autres étant fixés à zéro. Le bruit additif $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, avec $\sigma = 1.0$, est ajouté à la réponse. Pour chaque valeur du paramètre de régularisation λ, les modèles Ridge et Lasso sont entraînés sur un sous-échantillon aléatoire, puis évalués sur un ensemble test à l'aide de la racine de l'erreur quadratique moyenne (RMSE). L'expérience est répétée 50 fois pour chaque valeur de λ, et la figure finale présente la RMSE moyenne observée, mettant en évidence les différences de comportement entre les deux approches. Le code source de ce mini-projet est disponible sur GitHub à l'adresse suivante : `https://github.com/Saadem/Mini-Projet-IFT-7021.git`

La figure ci dessous révèle que, pour de très faibles valeurs de λ, les modèles Ridge et Lasso affichent une RMSE faible et très similaire, en raison d'une régularisation quasi absente. À mesure que λ augmente, la performance se maintient un temps, puis se détériore rapidement sous l'effet d'une pénalisation excessive. Dans la zone intermédiaire, Lasso surpasse Ridge grâce à sa capacité à éliminer les variables non pertinentes. Toutefois, lorsque λ devient très grand, les deux modèles voient leur RMSE augmenter fortement et leurs performances tendent à redevenir similaires, traduisant un sous-ajustement marqué dans les deux cas.

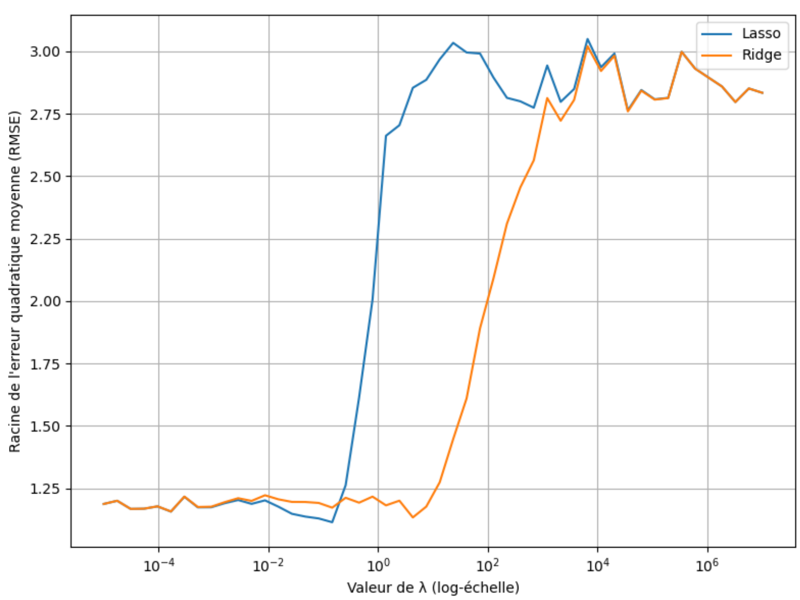

FIGURE 1 – RMSE moyenne en fonction de la régularisation λ pour Ridge et Lasso.