Le chemin des solutions du LASSO généralisé : Application à l'apprentissage statistique

Patrick Tardivel, institut de mathématiques de Bourgognes, Dijon, France

Le modèle de régression linéaire $Y = X\beta + \varepsilon$, où $Y \in \mathbb{R}^n$ est la réponse du modèle, $X \in \mathbb{R}^{n\times p}$ est la matrice de régression, $\beta \in \mathbb{R}^p$ est le paramètre inconnu des coefficients de régression et $\varepsilon \in \mathbb{R}^n$ est le résidu aléatoire est l'un des modèles en statistique les plus connus. Le paramètre β peut être estimé via l'estimateur LASSO généralisé qui est défini comme une solution du problème d'optimisation convexe suivant

$$\min_{b\in\mathbb{R}^p} \left\{ \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|Db\|_1 \right\}, \tag{1}$$

où $D \in \mathbb{R}^{m\times p}$. Le choix le plus standard pour la matrice D est l'identité, le terme de pénalité est alors la norme ℓ_1 et l'estimateur obtenu est le LASSO (acronyme signifiant « Least Absolute Shrinkage and Selection Operator ») qui est connu pour favoriser la parcimonie. De nombreuses autres matrices D sont également pertinentes pour cet estimateur (l'article [3] fournit de nombreux exemples). L'estimateur LASSO généralisé dépend du paramètre de régularisation λ ; on notera $\widehat{\beta}(\lambda)$ une solution du problème (1). Une approche classique en apprentissage statistique pour choisir ce paramètre de régularisation est de minimiser la somme des carrés résiduel sur un jeu de données de validation $X^{\text{val}}, Y^{\text{val}}$:

$$\lambda > 0 \mapsto \|Y^{\text{val}} - X^{\text{val}}\widehat{\beta}(\lambda)\|_2^2$$

La minimisation de cette expression nécessite de calculer le chemin des solutions de l'estimateur LASSO généralisé c'est-à-dire la fonction $\lambda > 0 \mapsto \widehat{\beta}(\lambda)$. L'étudiante ou l'étudiant choisissant ce stage pourra, à sa guise, explorer les points suivants :

- Étudier les résultats théoriques de l'article [2] traitant du chemin des solution du LASSO.
- Étudier les résultat théorique des articles [1, 3] traitant du chemin des solution du LASSO généralisé.
- Recoder sur R ou Python les algorithmes permettant de calculer le chemin des solution du LASSO ou LASSO généralisé et appliquer ces algorithmes sur des jeu de données réelles.
- Redémontrer que l'application $\lambda > 0 \mapsto \widehat{\beta}(\lambda)$ est affine par morceau.
- Étudier le chemin des solutions sur d'autres d'autres modèles (modèle de régression logistique ou modèle graphique par exemple...).

Références

[1] Taylor B Arnold et Ryan J Tibshirani : Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

[2] Julien Mairal et Bin Yu : Complexity analysis of the lasso regularization path. *In Proceedings of the 29th International Conference on Machine Learning*, pages 353–360, 2012.

[3] Ryan J. Tibshirani et Jonathan Taylor : The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335 – 1371, 2011.