TP 4 : Le Chemin de Régulatisation du Lasso

Stéphane Canu

12 octobre 2025, MLA, ITI, INSA Rouen

Le but du TP est d'étudier le chemin de régularisation du Lasso.

Pour le faire fonctionner, vous êtes supposé avoir déjà installé CVX (que vous pourrez télécharger à cette adresse : `http://cvxr.com/cvx/`)

Ex. 1 — Le Chemin de Régulatisation du Lasso

1. Reprendre les données 'prostate' du premier TP les centrer et les réduire

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

url_data = "https://web.stanford.edu/~hastie/ElemStatLearn/datasets/prostate.data"
df = pd.read_csv(url_data, delimiter='\t')

variables = df.columns[1:9]
df[variables] = df[variables].apply(lambda x: (x - x.mean()) / x.std())

# Get the training and test sets
Y_train = df.loc[df["train"]=="T", 'lpsa'].to_numpy()
X_train = df.loc[df["train"]=="T", variables].to_numpy()
print("Training set : n = {} samples and p = {} dimensions".format(X_train.shape[0],
    X_train.shape[1]))

Y_test = df.loc[df["train"]=="F", 'lpsa'].to_numpy()
X_test = df.loc[df["train"]=="F", variables].to_numpy()
print("Test set : n = {} samples and p = {} dimensions".format(X_test.shape[0], X_test.
    shape[1]))

Xn = (X_train - X_train.mean(axis=0))/X_train.std(axis=0)
n, p = Xn.shape
yn = (Y_train - Y_train.mean())/Y_train.std()
```

2. Résoudre le problème du Lasso sous la forme

$$\min_{\beta \in \mathbf{R}^p} J_\lambda(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \lambda ||\beta||_1 \tag{1}$$

avec $\lambda = 11.69$

Ex. 2 — Les deux bouts du chemin de régularisation du lasso

Soit $\widehat{\beta}_\lambda$ la solution du lasso (1), pour un λ donnée.

1. Le plus petit λ.
   a) En reprenant la formulation (1), à partir de la solution des moindres carrés et en utilisant les conditions d'optimalités, calculer λ_{min} tel que pour tout $\lambda \leq \lambda_{min}$ tous les tous les coefficients soient non nul .
   b) Quelle va être a première variable à être déclarée "non pertinente" ?
2. Le plus grand λ
   a) En reprenant la formulation (1), calculer λ_{max} tel que pour tout $\lambda \geq \lambda_{max}$ tel que tous les tous les coefficients soient nul ($\widehat{\beta}_\lambda = 0$). Pourquoi peut-on qualifier cette valeur de "plus grand λ" ?
   b) Quelle va être la première composante du vecteur de β non nulle à être activée

3. Calculez et représentez graphiquement en utilisant `matplotlib` la totalité du chemin de régularisation en utilisant SKlearn. Vérifiez que les résultats que vous avez sont bien conformes à ceux calculées aux question précédentes.
4. Dans la pratique, pour calculer les chemin de régularisation par quel bout est-il préférable de commencer ?

Ex. 3 — Le lasso simplifié : le lasso positif.

1. Pré traitement
   a) Que devient $\widehat{\beta}_\lambda$ la solution du Lasso quand une colonne de la matrice X change de signe ?
   b) Comment transformer la matrice X pour que toutes les composante du vecteur solution des moindre carrés (quand $\lambda = 0$) soient positifs ?
2. Dans le cas particulier ou l'on sait que tous les coefficients non nul de la solution du lasso sont positifs, le problème du lasso peut se réécrire

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^p} & \frac{1}{2} \|X\beta - y\|^2 \\ s.t. & \sum\limits_{j=1}^{p} \beta_j \leq t \\ et & \beta_j \geq 0 \quad j = 1, \ldots, p \end{cases}$$

   a) Résoudre ce problème et vérifier que la solution est bien la même que celle de la première question.
   b) Vérifier que les valeurs des multiplicateurs de Lagrange des contraintes sont bien conformes à la théorie.
   c) Montrez que les conditions de complémentarité font apparaitre une partition de l'ensemble de composants à travers deux ensembles $I_0 = \{j \in \{1, \ldots, p\} | \widehat{\beta}_j = 0\}$ et $I_1 = \{j \in \{1, \ldots, p\} | \widehat{\beta}_j > 0\}$.
   d) Démontrez numériquement que cette solution est bien optimale en vérifiant les KKT du problème.
   e) En utilisant les KKT, comment retrouver les valeurs de l'estimateur du lasso et λ connaissant, X, y, t et $I_0 = \{j | \widehat{\beta}_j = 0\}$, l'ensemble des indices des composantes nules de la solution du lasso.
   f) En utilisant les KKT, donnez les valeurs $\lambda_{min} < \lambda < \lambda_{max}$ telle qu'une composante quitte l'ensemble I_0 ou entre dedans.
3. En déduire le principe d'un algorithme de suivit du chemin de régularisation associée à une suite de λ, partant de λ_{min} et $I_0 = \{\emptyset\}$ et finissant à la solution des moindres carré avec $\lambda = 0$ et $I_0 = \{1, \ldots, p\}$