TP 4 : Conditions d'optimalité

Stéphane Canu

29 septembre 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. les conditions d'optimalités du Lasso
   a) Résoudre le problème du Lasso sous la forme
   $$\min_{\beta \in \mathbb{R}^p} J_L(\beta) = \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$
   avec $\lambda = 23.38$
   b) démontrez numériquement que cette solution est bien optimale en vérifiant que $0 \in \partial_\beta J_L(\beta)$ avec
   $$\partial_\beta J_L(\beta) = 2X^\top(X\beta - y) + \lambda \partial_\beta \|\beta\|_1$$
   où $\partial_\beta \|\beta\|_1$ est la sous différentielle de la norme 1 du vecteur β par rapport à lui même.
   c) Le lasso simplifié : le lasso positif. Dans ce cas particulier, tous les coefficients non nul étant positifs, le problème du lasso peut se réécrire
   $$\begin{cases} \min_{\beta} & \sum_{i=1}^{n} \left(y_i - x_i^\top \beta\right)^2 \\ s.t. & \sum_{j=1}^{p} \beta_j \leq t \\ and & \quad \beta_j \geq 0 \quad j = 1, \ldots, p \end{cases}$$

Résoudre ce problème et vérifier que la solution est bien la même que celle de la première question. Démontrez numériquement que cette solution est bien optimale en vérifiant les KKT du problème. Vérifier que les valeurs des multiplicateurs de Lagrange des contraintes sont bien conformes à la théorie.

3. Le plus grand λ
   a) En reprenant la formulation (1), à partir de la solution des moindres carrés et en utilisant les conditions d'optimalités, calculer le plus petit λ tel que tous les tous les coefficients soient nul. 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
4. Le plus petit λ
   a) Toujours avec la la formulation (1), calculer le plus grand λ tel que tous les tous les coefficients soient non nul.
   b) Quelle va être a première variable à être déclarée "non pertinente" ?
5. 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.