Le Lasso
Le gradient proximal

S.Canu
APPS ITI /INSA de Rouen Normandie

23 octobre 2023

Le Lasso

- Données
  - $X \in \mathbb{R}^{n \times p}$ les p variables observées n fois
  - $y \in \mathbb{R}^n$ la variable réponse
- Inconnues
  - $\beta \in \mathbb{R}^p$
- hypothèses : pas de biais (pas de terme constant)
  - $mean(X) = 0$
  - $\|X_j\|^2 = 1$
  - $mean(y) = 0$

le cout pénalisé

$$J_\lambda(\beta) \;=\; \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \;,$$

Les conditions d'optimalité

$$0 \in \partial J_\lambda(\beta) \;=\; X^\top(X\beta - y) + \lambda\partial(\|\beta\|_1) \;,$$

idée 1 : Le proximal L1

$$prox_{l1}(w, \lambda) = argmin_u \quad \lambda||u||_1 + \frac{1}{2}||u - w||^2$$

$$prox_{l1}(w, \lambda) = sign(w)max(|w| - \lambda, 0)$$

idée 2 : la linéarisation

$$\begin{array}{rccc} J: & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \boldsymbol{x} & \longmapsto & J(\boldsymbol{x}) \end{array}$$

Développement de Taylor au premier ordre

$$J(\beta) = J(\beta) + J(\beta^0) + (\beta - \beta^0)^t \nabla_\beta J(\beta^0) + o(\|\beta - \beta^0\|)$$

$$\begin{array}{ccccc} J(\beta) & = & J(\beta^0) & +(\beta - \beta^0)^t \nabla_\beta J(\beta^0) & +o(\|\beta - \beta^0\|) \\ \frac{1}{2}\|X\beta - y\|^2 & = & \frac{1}{2}\|X\beta^0 - y\|^2 & +(\beta - \beta^0)^t X^t (X\beta^0 - y) & +o(\|\beta - \beta^0\|) \end{array}$$

Développement de Taylor au second ordre

$$J(\beta) = J(\beta) + J(\beta^0) + (\beta - \beta^0)^t \nabla_\beta J(\beta^0) + \frac{1}{2}(\beta - \beta^0)^t H (\beta - \beta^0) + o(\|\beta - \beta^0\|^2)$$

idée 3 : la méthode de la région de confiance

Pour un β^0 donné

$$\begin{array}{ll} \min_{\beta} & \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \ , \\ \text{avec} & \|\beta - \beta^0\|^2 \leq \varepsilon \end{array}$$

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \ + \frac{1}{2t}(\|\beta - \beta^0\|^2 - \varepsilon)$$

idée 3 : la méthode de la région de confiance

Pour un β^0 donné

$$\begin{array}{ll} \min_{\beta} & \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \,, \\ \text{avec} & \|\beta - \beta^0\|^2 \leq \varepsilon \end{array}$$

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \; + \frac{1}{2t}(\|\beta - \beta^0\|^2 - \varepsilon)$$

idée 2+idée 3

Pour un β^0 donné

$$\min_\beta \tfrac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \ + \frac{1}{2t}(\|\beta - \beta^0\|^2 - \varepsilon)$$

$$\begin{aligned} J(\beta) = \tfrac{1}{2}\|X\beta - y\|^2 \quad &= J(\beta^0) + (\beta - \beta^0)^t \nabla_\beta J(\beta^0) \\ &= \tfrac{1}{2}\|X\beta^0 - y\|^2 + (\beta - \beta^0)^t X^t (X\beta^0 - y) \end{aligned}$$

$$\min_\beta (\beta - \beta^0)^t X^t (X\beta^0 - y) + \lambda\|\beta\|_1 \ + \frac{1}{2t}\|\beta - \beta^0\|^2$$

$$\min_\beta \ \lambda\|\beta\|_1 \ + \frac{1}{2t}\|\beta - \beta^0 + tX^t(X\beta^0 - y)\|^2$$

idée 2+idée 3+idée 1

Pour un β^0 donné

$$\min_{\beta} \; t\lambda\|\beta\|_1 \; + \frac{1}{2}\|\beta - \beta^0 + tX^t(X\beta^0 - y)\|^2$$

$$\begin{array}{ll} \min\limits_{\beta} & t\lambda\|\beta\|_1 \; + \frac{1}{2}\|\beta - v\|^2 \\ avec & v = \beta^0 - tX^t(X\beta^0 - y) \end{array}$$

L'algorithme du gradient proximal

Pour un β^0 donné

$t = \frac{1}{\|X^t X\|}$
Tant qu'on n' pas convergé....

$$
\begin{aligned}
v = & \quad \beta^k - tX^t(X\beta^k - y) \\
\beta^{k+1} = & \quad \arg\min_{\beta} \; t\lambda\|\beta\|_1 \; + \frac{1}{2}\|\beta - v\|^2
\end{aligned}
$$

Comment choisir t ?

J est gradient Lipschitz s'il existe une constante $0 < L$ telle que, $\forall \beta, \beta^0$

$$\|\nabla_\beta J(\beta) - \nabla_\beta J(\beta^0)\| \leq L \|\beta - \beta^0\|$$

Choix de t : argument de convergence

Si

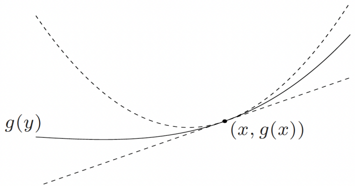
$$t \leq \frac{1}{L}$$

L'algorithme du gradient proximal converge

```
stepsize = 1/np.linalg.norm(X.T@X)
```

Le Lasso

Quadratic upper bound from Lipschitz property



- affine lower bound from convexity

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) \qquad \forall x, y$$

- quadratic upper bound from Lipschitz property

$$g(y) \leq g(x) + \nabla g(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \qquad \forall x, y$$