Variantes du Lasso

S.Canu
MLA ITI /INSA de Rouen Normandie

4 novembre 2024

Reweigthed least square (Lasso and Ridge)

$$J_\lambda(\beta) \;=\; \tfrac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \frac{\beta_j^2}{|\beta_j|} \;,$$

the idea : iterate the weighted rigde towards a fixed point

$$\beta^{(k+1)} \;=\; \arg\min_{\beta} \tfrac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} \frac{\beta_j^2}{|\beta_j^{(k)}|} = w_j \; \beta_j^2 \;,$$

with $w_j = 1/|\beta_j^{(k)}|$, that is

$$\beta^{(k+1)} \;=\; (X^\top X + \lambda W)^{-1} X^\top y \qquad \text{with } diag(W) = \frac{1}{|\beta_j^{(k)}|}$$

W = Identité
Tantque on n'a pas convergé
  $\beta = \arg\min \tfrac{1}{2}\|X\beta - y\|^2 + \lambda \beta^\top W \beta$
  $W = diag(1/|\beta|)$
fin de tantque

Cout et pénalité

- Ce que l'on veut : le modèle le plus simple qui s'ajuste aux données

$$\min_{\beta} \; cout(\beta) \text{ ET penalité}(\beta)$$

- Les deux critères sont contradictoires (le problème est mal posé)
- Le compromis que l'on doit faire (λ, t ou ε)

$$\min_{\beta} \; cout(\beta) \; + \; \lambda \text{ penalité}(\beta)$$

$$\begin{array}{ll} \min_\beta & cout(\beta) \\ \text{avec} & \text{penalité}(\beta) \leq t \end{array} \qquad \qquad \begin{array}{ll} \min_\beta & \text{penalité}(\beta) \\ \text{avec} & cout(\beta) \leq \varepsilon \end{array}$$

- exemple : le lasso

$$cout(\beta) \; = \; \frac{1}{2}\|X\beta - y'\|^2 \qquad \text{penalité}(\beta) \; = \; \|\beta\|_1$$

Lasso et Elastic Net

Le Lasso

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

Elastic Net

$$J_{\text{el}}(\beta) \;=\; \tfrac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 + \gamma\|\beta\|_2^2 \;,$$

Lasso = Elastic Net with $\gamma = 0, X = X', y = y'$

Elastic Net = Lasso with $X' = (X^\top, \sqrt{\gamma} I_p)^\top, y' = (y^\top, 0_p)^\top$

$$\|X\boldsymbol{\beta} - y\|_2^2 + \underbrace{\gamma\,\|\boldsymbol{\beta}\|_2^2}_{\|\sqrt{\gamma}\boldsymbol{\beta}\|_2^2} = \left\| \underbrace{\begin{bmatrix} X \\ \sqrt{\gamma} I_p \end{bmatrix}}_{=:X'} \boldsymbol{\beta} - \underbrace{\begin{bmatrix} y \\ 0_p \end{bmatrix}}_{=:\underline{y}'} \right\|_2^2 = \|X'\boldsymbol{\beta} - y'\|_2^2 .$$

Regularization and variable selection via the elastic net H Zou, T Hastie, 2005

Mise en œuvre de l'Elastic Net

- c'est un QP : on peut utiliser CVX, ou un autre solveur
- il existe aussi un chemin de régularisation
- Component wise

$$\partial_\beta J_{el}^{(1d)}(\beta) = \boldsymbol{x}^\top(\boldsymbol{x}\beta - \boldsymbol{r}) + 2\gamma\beta + \begin{cases} \lambda\alpha & \text{if } \beta = 0 \\ \lambda \text{sign}(\beta) & \text{else.} \end{cases}$$

$$\beta = \text{sign}(c)\max\big((|c| - \lambda), 0\big) \qquad c = \frac{\boldsymbol{x}^\top \boldsymbol{r}}{\|x\|^2 + 2\gamma}$$

- et le gradient proximal
  $t = \frac{2}{\|X^t X + 2\gamma I_p\|}$
  Tant qu'on n'a pas convergé....

$$\begin{aligned} v &= \beta^k - t\big(X^t(X\beta^k - y) + 2\gamma\beta^k\big) \\ \beta^{k+1} &= sign(v)max(|v| - t\lambda, 0) \end{aligned}$$

The adaptive Lasso

$$J_a(\beta) \;=\; \tfrac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j| \;,$$

$$w_j = \frac{1}{|\hat{\beta}_j^{(\mathrm{ls})}|^\gamma}$$

```
w = np.linalg.solve(X.T@X,X.T@y)
w = 1/np.sqrt(np.abs(w))
X_c = X/w
c = mon_lasso(X_c,y,lambda)
beta = c/w
```

The Adaptive Lasso and Its Oracle Properties, H. Zou, JASA, 2012

Mise en œuvre du Lasso adaptatif

- c'est un QP : on peut utiliser CVX, ou un autre solveur
- il existe aussi un chemin de régularisation
- Component wise

$$\partial_\beta J_a^{(1d)}(\beta) = \mathbf{x}^\top(\mathbf{x}\beta - \mathbf{r}) \; + \begin{cases} \lambda w \alpha & \text{if } \beta = 0 \\ \lambda w \operatorname{sign}(\beta) & \text{else.} \end{cases}$$

$$\beta = \operatorname{sign}(c)\max\big((|c| - \lambda w), 0\big) \qquad c = \frac{\mathbf{x}^\top \mathbf{r}}{\|\mathbf{x}\|^2}$$

- et le gradient proximal
  $t = \frac{2}{\|X^t X + 2\gamma I_p\|}$
  Tant qu'on n'a pas convergé....

$$\begin{aligned} v =& \quad \beta^k - t\big(X^t(X\beta^k - y)\big) \\ \beta^{k+1} =& \quad sign(v) max(|v| - t\lambda w, 0) \end{aligned}$$

The Grouped Lasso

Exemple : X code sous la forme de "one hot vector" (codage disjonctif complet) G variables catégorielles, chacune avec n_g catégories. I_g désigne l'ensemble des indices associés de sorte que $\beta_{I_g} = \{\beta_j | j \in I_g\}$.

$$J_\lambda(\beta) \;=\; \tfrac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{g=1}^{G} \sqrt{n_g}\|\beta_{I_g}\|_2 \;,$$

- si $n_g = 1$, on retrouve le lasso

Yuan and Lin, 2006 ; Bakin, 1999 ; Cai, 2001 ; Antoniadis and Fan, 2001

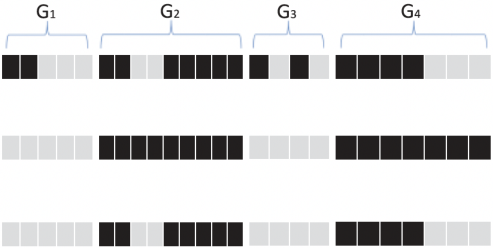
The Grouped Lasso

$$\min_{\beta} \; cout(\beta) \text{ ET } \text{penalité}(\beta) \; \min_{\beta} \; cout(\beta) \; + \; \lambda \, \text{penalité}(\beta)$$

Le terme d'attache aux données

$$cout(\beta) = \frac{1}{2}\|X\beta - y\|^2$$

La pénalité



Lasso $\|\beta\|_1$

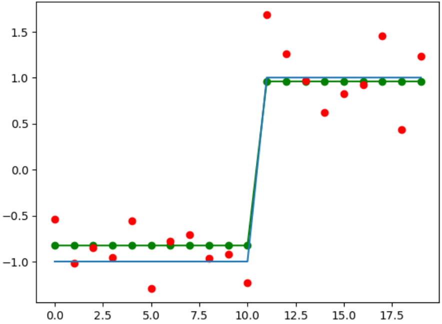
Group Lasso $\sum_{g=1}^{G} \sqrt{n_g}\|\beta_{I_g}\|_2$

Sparse Group Lasso $\|\beta\|_1 + \gamma \sum_{g=1}^{G} \sqrt{n_g}\|\beta_{I_g}\|_2$

The Fused Lasso

$$J_f(\beta) \;=\; \frac{1}{2}\|X\beta - y\|^2 + \mu \sum_{j=1}^{p-1} |\beta_j - \beta_{j+1}| \left(+\, \lambda ||\beta||_1\right) ,$$

Exemple : recherche de composantes continues



$$n = p$$

$$X = I, \lambda = 0$$

- y les points rouges
- β les points verts

(Tibshirani et al., 2005)

Exemples d'autres termes d'attache aux données

- The Dantzig selector

$$J_\lambda(\beta) \;=\; \|X^T(X\beta - y)\|_\infty + \lambda||\beta||_1 \;,$$

Candès and Tao, 2007

- LAD lasso (least absolute deviation)

$$J_\lambda(\beta) \;=\; \|X\beta - y\|_1 + \lambda||\beta||_1 \;,$$

Wang et al., 2007

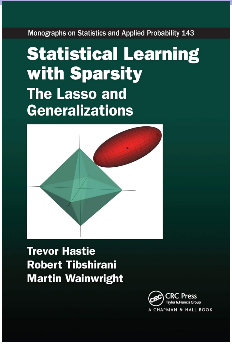
- The logistic lasso

$$J_\lambda(\beta) \;=\; -\sum_{i=1}^{n}\Big(y_i x_i^T\beta - \log\big(1 + \exp^{x_i^T\beta}\big)\Big) + \lambda||\beta||_1 \;,$$

The L1 penalty for logistic regression, Tibshirani,1996

Conclusion

- machine learning : optimisation bi critère
  - terme d'attache aux données (vraisemblance)
  - pénalité (a priori)
- Sélection de variables
  - Parcimonie = singularité
  - Parametric QP (pénalité convexe)
- Lasso when p is very large : Sceening
- Sélection d'individus ?



Variant	Key Feature	Best Use Case
Elastic Net	Combines L1 and L2 penalties	Highly correlated predictors
Adaptive Lass	Weighted penalties for coefficients	High-dimensional data
Group Lasso	Selects groups of variables	Grouped predictors
Fused Lasso	Encourages sparsity in changes	Time-series or spatial data
Dantzig selec	L_∞ data related loss	Highly correlated