École Nationale de la Statistique et de l'Administration Économique

Rapport du Projet de Statistique

Janvier 2017

Introduction à la régression pénalisée et à l'estimateur lasso

Auteurs :
Mehdi Abbana Bennani
Julien Mattei

Superviseur :
Edwin Grappin

28 décembre 2016

1 Régression OLS

Question 1 :

$$\min_{\beta} \|y - X\beta\|^2 \qquad \text{(On minimise l'erreur quadratique)}$$

$$\nabla_\beta \|y - X\beta\|^2 (\beta^\star) = 0 \qquad \text{(Le gradient est nul en } \beta^\star)$$

$$\nabla_\beta (y - X\beta)^T (y - X\beta)(\beta^\star) = 0$$

$$\boxed{\beta^\star = (X^T X)^{-1} X^T y} \qquad (1)$$

Dans la suite on notera $\beta^\star$ l'estimateur des moindres carrés.

Question 2 : Voir code.

On obtient une RMSE de 3.85, sachant que la variance de y est 90 à peu près.

2 Régression pénalisée : le lasso

Question 3 :

Démonstration. On ne peut pas utiliser la régression par moindres carrés si le nombre de variables est supérieur au nombre d'observations
X est une matrice de dimensions (n,p)
Donc $X^T X$ est une matrice carrée de dimension p
On a $rg(X^T X) = rg(X)$
Donc $rg(X^T X) \leq min(n, p)$
Par hypothèse, le nombre de variables est plus grand que le nombre d'observations, soit $p > n$
Donc $rg(X^T X) < p$
Or $X^T X$ est une matrice de taille p
Donc $X^T X$ n'est pas inversible
Donc on ne peut pas utiliser la méthode précédente.
La matrice $X^T X$ n'est plus inversible car son rang est inférieur au minimum des dimensions de X. □

Question 4 :

Sous l'hypothèse de p=1, on résoud le problème d'optimisation :

$$\min_{\beta} (y - x\beta)^2 + \lambda|\beta|$$

A partir de l'expression (1), on a $\hat{\beta} = \frac{y}{x}$

On injecte $\hat{\beta}$ dans le problème d'optimisation et après simplification :

$$\min_{\beta} \frac{x^2\beta^2}{2} - x^2\beta\hat{\beta} + \lambda|\beta|$$

On remarque que si $\hat{\beta} < 0$ Alors forcément $\beta^{lasso} < 0$

De même si $\hat{\beta} > 0$ Alors forcément $\beta^{lasso} > 0$

On va traîter les deux cas

Si $\hat{\beta} > 0$, on résoud :

$$\min_{\beta} \frac{x^2\beta^2}{2} - x^2\beta\hat{\beta} + \lambda\beta$$

On obtient $\beta^{lasso} = \hat{\beta} - \frac{\lambda}{x^2}$

Or β^{lasso} forcément positif, donc $\beta^{lasso} = max(0, \hat{\beta} - \frac{\lambda}{x^2})$

Si $\hat{\beta} < 0$, à l'aide d'un raisonnement similaire $\beta^{lasso} = max(0, \hat{\beta} + \frac{\lambda}{x^2})$

On en déduit que $\beta^{lasso} = sg(\hat{\beta})(|\hat{\beta}| - t)_+$ avec $t = \frac{\lambda}{x^2}$

(2)

Le seuillage doux permet de regler l'intervalle sur lequel on veut estimer $\beta^\star$. Ainsi, on choisit λ, si notre estimateur $\beta^\star$ est compris dans l'intervalle $[-\lambda, \lambda]$ il est ramené à 0. On ne retient que les valeurs de $\beta^\star$ en dehors de cet intervalle (même principe qu'un filtre).
L'estimateur $\hat{\beta}^{lasso}$ est nul pour tout les $\beta^\star$ tels que $\beta^\star - t < 0$. Donc cela induit une estimation nulle plus souvent que l'estimateur des moindres carrés (qui se produit seulement lorsque $\beta^\star = 0$)

Question 5 : Dans la régression pénalisée, on pénalise les valeurs de β pour limiter leur valeurs possibles, dans le cas du lasso par exemple, toutes les estimations par moindres carrées inférieures à t sont ramenées zeros, l'intervalle d'estimation est $\{0\}\mathrm{U}[t, \infty[$. Dans le cas de la régression Ridge par exemple, on pénalise les coefficients trop grands.
L'estimateur Lasso est adapté à notre problème car nous allons pouvoir réduire la taille de l'espace des paramètres jusqu'à ce que le nombre de coefficients de beta non nuls soit égal à n le nombre d'observations, en changeant la valeur de λ. Plus celui-ci est grand, plus le nombre de paramètres nuls sera grand.

Question 6 : Quelles sont les avantages du lasso sur la pénalisation ‘ 0 ? Même question avec la pénalisation Ridge

Question 7 : A implémenter

Question 8 : Le λ permet de régler la "force" de pénalisation du paramètre β.

Question 9 : Dans la méthode de cross validation, on divise le dataset en trois parties : Train Set, Validation Set, et Test Set. On entraine le modèle sur le train set puis on estime la qualité de la prédiction sur le validation set pour plusieurs valeurs de λ, on choisit le lambda le plus performant sur le validation set pour entrainer le modèle sur le nouveau train set comprenant le train set précédent et le validation set.

Question 10 : A implémenter

Combien d'éléments de β estimez vous non-nuls ? Evaluez l'erreur de votre modèle. Trouvez de bonnes mesures pour évaluer votre méthode