Cours de Machine Learning

Régression régularisée, MARS

M2 Informatique pour la Science des Données –2020-2021
Université Paris Saclay, D. Jeannel
ML.M2.ISD.Orsay@gmail.com

Régression régularisée, MARS

1. Introduction
2. Régression pénalisée
   1. Principe
   2. Régressions Ridge et Lasso
   3. Différences Ridge et Lasso
   4. Applications numériques
3. MARS
   1. Introduction
   2. Modélisation
   3. Applications numériques

1. Introduction

- MCO : régression Moindre Carré Ordinaire (OLS)
  - Technique découverte par Legendre (1805) et Gauss (1809) pour résoudre des problèmes astronomiques
  - Fondements statistiques établis par Fisher en 1920
  - Utilisé par les calculateurs électromécaniques en 1950

- Modèle de la forme :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

  - Réponse exprimée par un hyperplan
  - Constante β_0
  - Estimateurs $\beta_1, \beta_2, \ldots, \beta_p$
  - Estimateurs obtenus en minimisant la somme des carrés des résidus
  - Existence de techniques de sélections de variables (forward, stepwise….)

1. Introduction

- Précautions et conditions d'emploi de l'usage des MCO
  - Nettoyage des données
  - Gestion des données manquantes
  - Transformation des données
  - Choix des Sélection des variables
  - Connaissance des interactions significatives
  - Conditions de validité (risque de multicolinéarité, hypothèse de normalité des résidus)

  ➔ Régression pénalisée (Ridge, LASSO), MARS

2. Régression pénalisée - Principe

- Alternatives pour construire des modèles linéaires pour
  - Obtenir une solution stable en présence de multicolinéarité,
  - Prendre critère additionnel d'optimisation que l'erreur quadratique moyenne
  - Éviter les problèmes de sur-ajustement (overfit)
  - Obtenir une solution unique lorsque le volume de données est important
- Formes de modélisation :

Régression MCO

Minimisation

$$SCR + \lambda * \text{Coefficient de complexité}$$

Minimisation

Régression pénalisée

Coefficient de complexité :

- Ridge : somme des carrés des coefficients
- Lasso : somme des coefficients en valeur absolue

2. Régression pénalisée – Régressions Ridge & Lasso

- Régression Ridge et Lasso

Régression Ridge $$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2}$$ ← Coefficient « shrinkage penalty »

Régression LASSO $$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \boxed{\lambda \sum_{j=1}^{p} |\beta_j|}$$ ← Coefficient « lasso penalty »

  - λ :paramètre de réglage (tuning parameter)
    - $\lambda = 0$ ➔ estimateur des MCO
    - $\lambda \rightarrow \infty$, coefficients $\beta \rightarrow 0$
    - Constante non affectée par λ (valeur moyenne de la réponse lorsque les coefficients β sont nuls)

2. Régression pénalisée – Régressions Ridge & Lasso

- Estimation des coefficients dépendante de l'échelle des variables
  - Centrage et réduction des variables continues

- Estimation par MCO pour régression Ridge

$$\hat{\beta}_{\lambda}^{ridge} = (\mathbb{X}'\mathbb{X} + \lambda I_p)^{-1}\mathbb{X}'Y$$

  - Choix critique du paramètre de réglage λ
  - Critère de validation croisée
  - Variance estimateur β_{RIDGE} < Variance estimateur β_{MCO}

2. Régression pénalisée – Régressions Ridge & Lasso

- Estimation des coefficients pour régression LASSO
  - A cause de la valeur absolue, estimation par algorithme quadratique pour régression Lasso
  - Hyp : si variable xj orthonormée

$$\left[\hat{\beta}_{\lambda}^{lasso}\right]_j = (x^j)'Y\left(1 - \lambda/\left(2\left|(x^j)'Y\right|\right)\right)_+$$

  - Estimateur LASSO peut-être nul pour un certain nombre de variables ➔ sélection de variables

2. Régression pénalisée – Différences Ridge & Lasso

- Avantages régression Ridge
  - Si p> n, pas de solution unique pour estimateur MCO alors que estimateur Ridge peut améliorer légèrement le biais pour une grande baisse de la variance
  - Estimateurs Ridge préférable à estimateurs MCO lorsque Var βmco est très grande
  - Pas besoin de faire 2^p modèles pour sélectionner le meilleur modèle (avantage calculatoire)
  - En présence de relation linéaire entre réponse et prédicteurs, bais faible estimateur MCO mais variance peut-être élevée

- Défaut régression Ridge
  - Prise en compte de l'ensemble des p variables
  - Pas d'indicateurs pour distinguer les variables les plus influentes

2. Régression pénalisée – Différences Ridge & Lasso

- Avantages régression Lasso
  - Sélection des variables. Effets des variables peu importantes sont estimés à o
  - Interprétation du modèle Lasso plus facile que modèle Ridge
  - Modèles plus simples (nombre de variables < p)

- Défaut régression Lasso
  - En présence de variables explicatives corrélées, sélection arbitraire d'une variable influente par rapport aux autres (effets nuls mis aux autres)
  - Algorithme de calcul pour régression Lasso

Variable
| $\mathbb{X}$, Y, λ : *matrice* $n \times p$, *vecteur de taille* n, *réel* > 0
Début
| Initialiser $\beta = \beta_{in}$
| **Répéter**
| | **Pour** j **variant de** 1 **à** p
| | | Calculer $R_j = (x^j)'(Y - \sum_{k \neq j} \beta_k x^k)$
| | | Calculer $\beta_j = R_j(1 - \lambda/(2|R_j|))_+$
| | **FinPour**
| **Jusque** Convergence de β
| **Retourner** β
Fin

2. Régression pénalisée – Différences Ridge & Lasso

- Intuition géométrique sélection de variable par régression Lasso

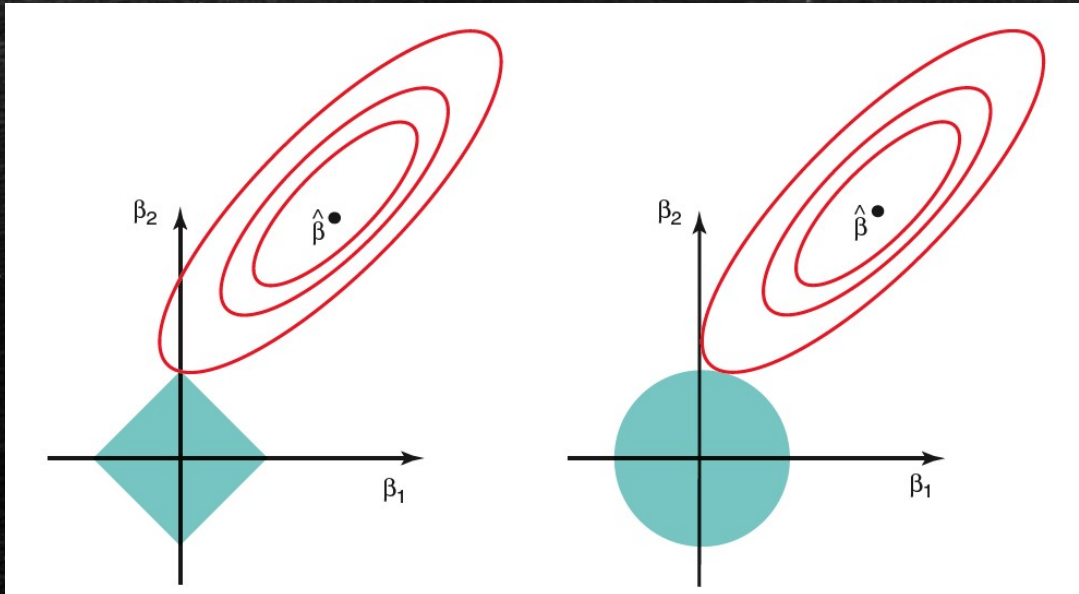


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

2. Régression pénalisée – Différences Ridge & Lasso

- Exercice intuitif sur le comportement des estimateurs Moindres Carrés Ordinaires, Ridge et Lasso :
  - Soient n observations, p variables (p = n), matrice des variables explicatives X = diag(1,…,1)

  Question : déterminer l'estimateur βj pour les différents types de régressions suivantes :

  - Régression Moindres Carrés Ordinaires : $\sum_{j=1}^{p}(y_j - \beta_j)^2$
  - Régression Ridge : $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$
  - Régression Lasso : $\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

Question : représentation graphique de l'estimateur βj en fonction de yj

3. Régression multivariée par spline adaptative

- Introduction
  - Procédure adaptive pur régression développé en 1991 par J. Friedman
  - Technique non-paramétrique adaptée pour données à grande dimension et la déterction de non-linéarité

- Modèle de la forme

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X)$$

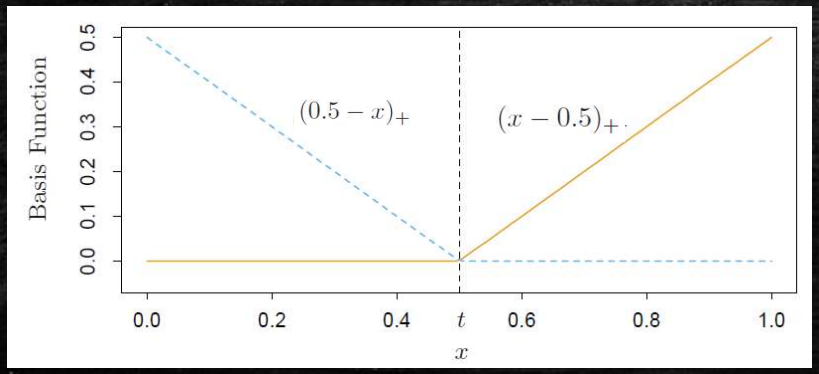
  - $h_m(X)$ fonctions de base linéaires par morceaux
  - Pas d'hypothèses sur les données de base
  - Traitement des données manquantes et des données qualitatives (différence réseau de neurones)
  - Quasi-aussi précis qu'un réseau de neurones mais interprétable par rapport à réseau de neurones

3. Régression multivariée par spline adaptative

- Modélisation
  - Fonctions de base linéaires par morceaux

$$(x-t)_+ = \begin{cases} x-t, & \text{if } x > t, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad (t-x)_+ = \begin{cases} t-x, & \text{if } x < t, \\ 0, & \text{otherwise.} \end{cases}$$

  - Exemple :



3. Régression multivariée par spline adaptative

- Modélisation
  - Construction des fonctions de base linéaires par morceaux

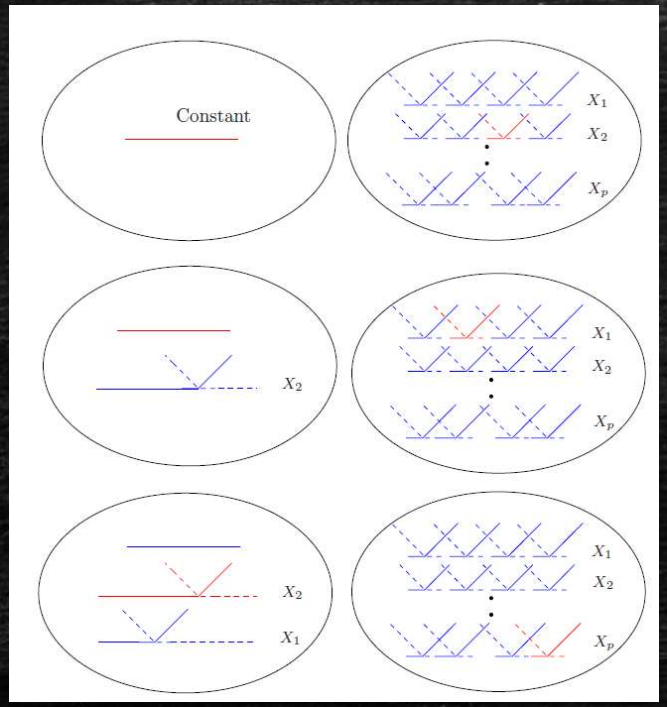
$$\mathcal{C} = \{(X_j - t)_+,\ (t - X_j)_+\} \quad \begin{matrix} t \in \{x_{1j}, x_{2j}, \ldots, x_{Nj}\} \\ j = 1, 2, \ldots, p. \end{matrix}$$

    - t représente des nœuds construits comme sur le principe de CART

  - Construction du modèle
    - Ajout de fonctions de base de telle sorte que la SCR du modèle diminue (parallèle avec procédure FORWARD)
    - Fonctions ajoutées peuvent être issues de C ou être des produits de 2 ou plus de fonctions

3. Régression multivariée par spline adaptative

- Modélisation
  - Visualisation graphique de la construction d'un modèle MARS



Etape 1 : X2 est ajoutée car elle diminue fortement la SCR

A chaque étape, on considère toutes les fonctions candidates et oin retient celle qui diminue significativement la SCR (fonction en rouge)

3. Régression multivariée par spline adaptative

- Modélisation
  - Procédure de validation croisée pour déterminer le nombre de paramètres dans le modèle
  - Critère de validation croisée généralisée (GCV) :

$$\mathrm{GCV}(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\lambda(x_i))^2}{(1 - M(\lambda)/N)^2}$$

  - M(λ) : nombre de paramètres du modèle M
  - N : nombre de données
  - On choisit le modèle qui minimise le GCV

4. Applications Numériques

- Régression RIDGE, LASSO :

Type de régression :
0 : RIDGE
1 : LASSO

Valeur du paramètre de régularisation (grille de valeur ou valeur seule)

```
library (glmnet )
grid =10^ seq (10,-2, length =100)
regression_penalisee =glmnet (x,y,alpha =1, lambda =grid)
regression_penalisee
```

- MARS :

```
library(earth)
modele_MARS = earth(x,y,deg=1)
modele_MARS
evimp(modele_MARS, trim=T)
plotmo(modele_MARS,ylim=NA)
```

Niveau d'interaction des variables explicatives
deg=1 (effet simple), deg=2 (interaction double),…

Importance des variables

Visualisation des « Partial Dependant Plot »