Pratique de l'apprentissage statistique

2. Régression régularisée

V. Lefieux

École des Ponts ParisTech

Plan

Plan

Généralités

Régression Ridge

Régression LASSO

Compléments

Introduction

Les méthodes de régularisation permettent de répondre à plusieurs problématiques :

- Sélection (ou pondération) de variables.
- Traitement de la colinéarité dans les modèles linéaires.
- Traitement des « fat matrix » : $n < p$ (méthodes « sparses »).

变量的选择（或加权）。
处理线性模型中的共线性。
处理“脂肪矩阵”：n < p（“稀疏”方法）。

Données considérées

- On dispose d'un échantillon de (X, Y) :

$$\mathcal{D}_n = (X_i, Y_i)_{i \in \{1,\ldots,n\}}$$

  où $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ et $Y \in \mathbb{R}$.

- On note :

$$d_n = (x_i, y_i)_{i \in \{1,\ldots,n\}} \ .$$

Modèle

On considère le modèle de régression linéaire suivant :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

où ε désigne l'erreur du modèle de régression.

Plan

Estimateur Ridge

- On appelle estimateur Ridge de $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ le vecteur $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}$ solution de :

误差小 系数绝对值之和小

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

où $\lambda \geq 0$ est un paramètre de régularisation (à déterminer).

- On considère ici une pénalité ℓ^2.

Problème d'optimisation équivalent

On peut également voir cet estimateur comme solution du problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \qquad \text{误差平方之和}$$

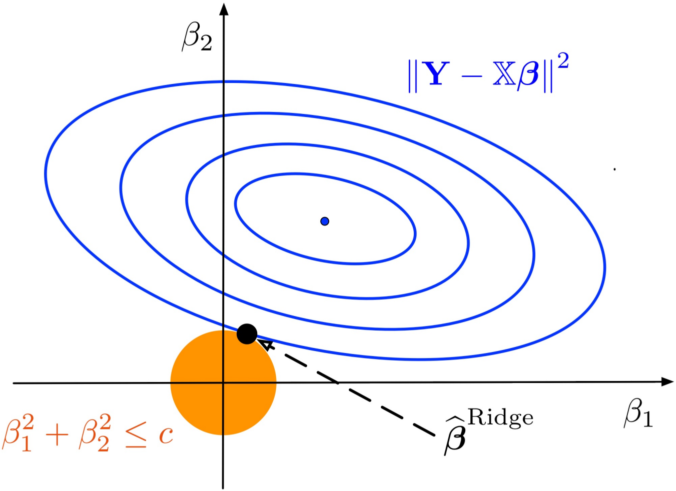
$$\text{sc} \quad \sum_{j=1}^{p} \beta_j^2 \leq c \ .$$

Remarques

我们首先将变量 (Y;X1;:::;Xp) 居中。
惩罚常数被排除（如果存在）。
通过交叉验证得到最优参数。

- On centre au préalable les variables $(Y, X_1, \ldots, X_p)$.
- On exclut la constante de la pénalisation (si elle est présente).
- Le paramètre optimal est obtenue par validation croisée.

Illustration



Explicitation de la solution

- On peut réécrire ce problème sous forme matricielle :

$$\min_{\boldsymbol{\beta}} \left(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\right)^\top \left(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}\right) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta} \ .$$

où :

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbb{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np}. \end{bmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \ .$$

- La solution est :

$$\widehat{\beta}^{\text{Ridge}} = \left(\mathbb{X}^\top \mathbb{X} + \lambda \, \mathsf{I}_p\right)^{-1} \mathbb{X}^\top \mathbf{Y} \ .$$

- Il s'agit encore d'un estimateur linéaire en $\mathbf{Y}$.

Mise en évidence du seuillage I

阈值

- On suppose que $n > p$. n:样本数 p:系数个数
- La décomposition en valeurs singulières permet d'obtenir :

分解成奇异值

$$\mathbb{X} = U\,D\,V^\top \; .$$

où :

- U est une matrice orthogonale de dimensions (n, n) : n*n正交阵

$$U\,U^\top = U^\top\,U = \mathrm{I}_n$$

- V est une matrice orthogonale de dimensions (p, p) : p*p正交阵

$$V\,V^\top = V^\top\,V = \mathrm{I}_p$$

- D est une matrice de dimensions (n, p) ne contenant que des termes positifs sur sa « diagonale » : n*p对角阵

$$d_1 \geq \ldots \geq d_p \geq 0 \; .$$

Mise en évidence du seuillage II

- On peut écrire :

$$\mathbb{X}\,\widehat{\beta}^{\text{Ridge}} = UD\left(D^\top D + \lambda\, \mathsf{I}_p\right)^{-1} D^\top U^\top \mathbf{Y}\ .$$

- On a :

$$\mathbb{X}\,\widehat{\beta}^{\text{Ridge}} = \sum_{j=1}^{p} u^j \left(\frac{d_j^2}{d_j^2 + \lambda^2}\right) {u^j}^\top \mathbf{Y}$$

  où $\left(u^1, \ldots, u^p\right)$ désignent les colonnes de la matrice U.

- Pour $\lambda = 0$, on retrouve bien la solution des MCO :

$$\mathbb{X}\,\widehat{\beta} = \sum_{j=1}^{p} u^j {u^j}^\top \mathbf{Y}\ .$$

Mise en évidence du seuillage III

- L'élément j de la base est « seuillée » par $\frac{d_j^2}{d_j^2+\lambda^2}$. 阈值确定
- Les plus petits coefficients sont les plus seuillés.
- Plus λ est grand, plus le seuillage est important.
- L'effet biais augmente avec λ. 偏差增大
- L'effet variance diminue avec λ. 方差减小
  Dans le cas où λ est faible, on peut être confronté à du sur-apprentissage.
- On définit le degré de liberté par :

$$\mathrm{ddl}\,(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda^2}\,.$$

自由度

Mise en évidence du seuillage IV

- Dans le cas où les variables sont centrées, la matrice de variance-covariance des variables vaut :

$$\frac{1}{n}\mathbb{X}^\top\mathbb{X} = \frac{1}{n}VD^\top DV^\top .$$

- On sait que :

$$D^\top D = \mathrm{diag}\left(d_1^2, \ldots, d_p^2\right) .$$

- Si $\left(v^1, \ldots, v^p\right)$ désignent les colonnes de la matrice V, on peut montrer que $\mathbb{X}\,v^j$ est la j-ème composante principale de $\mathbb{X}$, de variance $\frac{d_j^2}{n}$.
- La régression Ridge seuille donc peu les premières composantes principales (d_j grand) mais davantage les dernières.

Plan

Estimateur LASSO

最小绝对收缩和选择算子

- On appelle estimateur LASSO (Least Absolute Shrinkage and Seletion Operator) de $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ le vecteur $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$ solution de :

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

où $\lambda \geq 0$ est un paramètre de régularisation (à déterminer).

- On considère ici une pénalité ℓ^1.

Problème d'optimisation équivalent

On peut également voir cet estimateur comme solution du problème d'optimisation suivant :

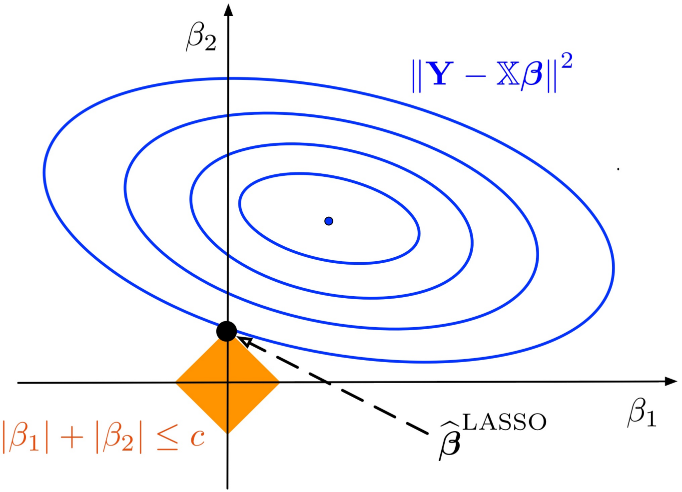
$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2$$

$$\text{sc} \quad \sum_{j=1}^{p} |\beta_j| \leq c \ .$$

Remarques

- On centre au préalable les variables $(Y, X_1, \ldots, X_p)$.
- On exclut la constante de la pénalisation (si elle est présente).
- Le paramètre optimal est obtenue par validation croisée.

Illustration



Propriétés

- L'estimateur LASSO n'est pas linéaire en $\mathbf{Y}$.
- En toute généralité, on ne dispose pas d'expression explicite de l'estimateur LASSO (non-dérivabilité du critère ℓ^1).
- Si $\lambda \to +\infty$, on tend à annuler tous les paramètres $(\beta_j)_{j\in\{1,\dots,p\}}$.

Cas particulier

- On considère le cas où la matrice $\mathbb{X}$ est orthogonale :

$$\mathbb{X}^\top \mathbb{X} = \mathsf{I}_p \ .$$

- On considère le problème :

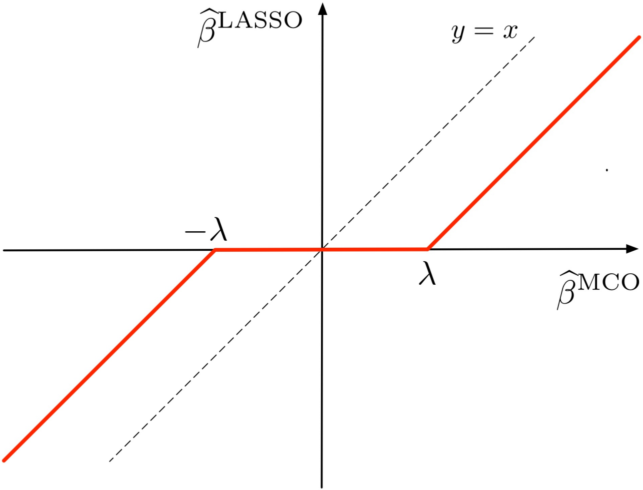
$$\min_{\boldsymbol{\beta}} \ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 + 2\lambda \sum_{j=1}^{p} |\beta_j| \ .$$

- Pour $j \in \{1, \ldots, p\}$, on obtient la solution explicite suivante :

$$\widehat{\beta}_j^{\text{LASSO}} = \text{signe}\left(\widehat{\beta}_j^{\text{MCO}}\right) \left(\left|\widehat{\beta}_j^{\text{MCO}}\right| - \lambda\right) \mathbb{1}_{\left|\widehat{\beta}_j^{\text{MCO}}\right| \geq \lambda} \ .$$

- On parle de « seuillage doux » (*soft thresholding*) de l'estimateur des MCO.

Seuillage doux



Plan

Critère ℓ^q

Les estimateurs Ridge et LASSO sont des cas particulier du problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

où $\lambda \geq 0$ est un paramètre de régularisation (à déterminer par validation croisée) et $q \in \mathbb{R}^+$.

Méthode Elastic Net

La méthode Elastic Net combine les régressions Ridge et LASSO via le problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right)$$

où $\lambda \geq 0$ et $\alpha \in]0, 1[$ sont des paramètres de régularisation (à déterminer par validation croisée).

Méthode LARS : principe

La régression LARS (Least Angle Regression) est une méthode :

- de type forward,
- qui n'intègre pas complètement une variable explicative : elle l'intègre au niveau de son « mérite ».

LARS回归（最小角度回归）是一种方法：
前向型，
它没有完全整合一个解释变量：它在它的“优点”水平上整合它。

Méthode LARS : algorithme

1. Initialisation :
   - Centrage et réduction des covariables. 协变量的居中和归约
   - $\mathbf{e} = \mathbf{Y} - \bar{y}\,\mathbf{1}_n$
     où $\mathbf{1}_n$ est une vecteur de dimension n constitué de 1.
   - $\boldsymbol{\beta} = 0$.

   找到与残差 e 最相关的协变量 Xj
2. Trouver la covariable X_j la plus corrélée avec le résidu $\mathbf{e}$.
3. « Déplacer » β_j vers $\operatorname{corr}(\mathbf{X}_j, \mathbf{e})$, où $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})^\top$, jusqu'à ce qu'une autre covariable X_k ait une corrélation plus importante avec le résidu.
4. « Déplacer » β_j et β_k dans une direction donnée par leur estimation des MCO conjointe jusqu'à ce qu'une autre covariable X_ℓ ait une corrélation plus importante avec le résidu.
5. Poursuivre jusqu'à intégration de toutes les covariables.

Références

Hastie, T., R. Tibshirani et J. H. Friedman. 2009, *The elements of statistical learning. Data Mining, inference, and prediction*, 2e éd., Springer Series in Statistics, Springer.

Hastie, T., R. Tibshirani et M. Wainwright. 2015, *Statistical learning with sparsity. The Lasso and generalizations*, Monographs on Statistics & Applied Probability (143), CRC. Chapman & Hall.