Ecole Centrale Casablanca

Apprentissage Statistique et Réseaux de Neurones

TP 2 : Régression linéaire

Auteur :
Mohamed EL MAIMOUNI

17 octobre 2020

Nomenclature

X	Matrice de dimension n x p
Y	Vecteur de $\mathbb{R}^n$
β	Vecteur de $\mathbb{R}^p$
ϵ	Bruit blanc gaussien de variance 1

Introduction

Ce rapport synthétise et analyse les résultats obtenus à travers le deuxième TP du module "**Apprentissage statistique et réseaux de neurones**". Après un TP où nous avons présenté les différents facteurs qui nous permettent d'évaluer la performance d'une régression linéaire

L'objectif de ce TP est d'étudier un cas particulier dans lequel nous disposons de nombre de critères plus que d'observations (par exemple, essayer de prédire le prix d'une voiture en fonction de ses fonctionnalités, alors que nous ne disposons que de quelques échantillons de voitures avec leurs prix)

1 Étude théorique

En continuité avec ce qui a été évoqué dans le rapport du premier TP (cf. rapport TP 1), les données dont nous disposons sont représentées par une matrice de n lignes et p colonnes, où les lignes et les colonnes représentent les observations et les critères respectivement.
Nous allons alors étudier le cas où $n < p$

Génération du modèle

Dans ce TP nous allons générer une matrice X à travers une distribution de normale centrée de variance $\sigma^2 = 4$
Puis Y est générée selon la relation suivant :

$$Y = X\beta + \epsilon$$

avec $\beta = (-2, 7, 0, .., 0) \in \mathbb{R}^p$ et ϵ est un bruit blanc gaussien de variance 1
On va estimer $\hat{\beta}$ à l'aide de la formule du cours

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

En implémentant cette relation dans **R**, on obtient l'erreur :
`le système est numériquement singulier : conditionnement de la réciproque = 4.10241e-20`

Cette erreur est dû à la non-inversibilité de $X^t X$. En effet, X n'est pas de plein rang. Comme nous avons vu en cours, on peut recourir à une régularisation du problème.

La fonction glmnet

Dans cette section, nous allons estimer β à l'aide de la formule du cours, puis utiliser la fonction glmnet de **R** et comparer les résultats.

D'après la formule du cours, dans le cas d'une régression régularisée Ridge :

$$\hat{\beta} = (X^t X + \lambda I_p)^{-1} X^t Y$$

On peut alors calculer cette grandeur car $(X^t X + \lambda I_p)$ est inversible.

On rappelle que la régression régularisée consiste à minimiser :

$$\beta \to ||Y - X\beta|| + \lambda[\alpha(||\beta||_1 + (1-\alpha)||\beta||_2^2]$$

Ridge	$\alpha = 0$
Lasso	$\alpha = 1$
Elastic Net	$\alpha \notin \{0, 1\}$

D'autre part, la fonction `glmnet` peut résoudre le problème Elastic net pour une α donnée et une sèquence de λ.
Soit $\lambda = 1$, d'abord, le vecteur $\hat{\beta}_{Ridge}$ calculé à travers la formule du cours, contient 100 valeurs, alors que $\hat{\beta}_{ElasticNet}$ est un vecteur de $\mathbb{R}^p$ avec peu de valeurs, en l'occurrence quelques éléments et le reste sont des cases vides (.)

$$\hat{\beta}_{Ridge} = (-0.27248, 1.18610, ..., 0.06050)$$
$$\hat{\beta}_{Elsatic\ Net} = (-1.16478, 6.58757, 0, \beta_i ...)$$

avec $\hat{\beta}_{Elsatic\ Net,36} = 0.25177,\ \hat{\beta}_{Elsatic\ Net,59} = 0.45527, ..etc$

On peut interpréter le fait que la méthode Elastic Net retourne un vecteur avec peu de valeurs non nulles, par la tendance de la penalty l^1 d'éliminer les critères les moins importantes (colonnes de X) de l'étude dans le cas où p » n

De plus, on remarque que $\hat{\beta}_{Elsatic\ Net}$ approche notre β théorique plus que $\hat{\beta}_{Ridge}$

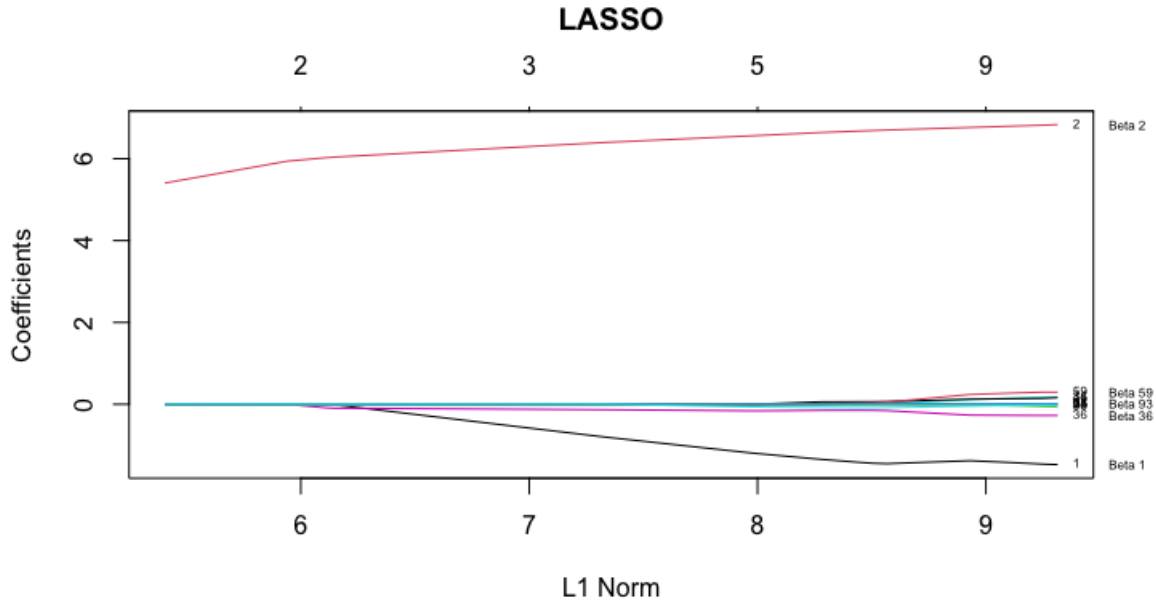

FIGURE 1 – Plot des coefficients des β_i en fonction de la $||\beta^\lambda||_1$

Nous remarquons que $\beta_2 \to 7$ lorsque $||\hat{\beta}^\lambda||_1$ augmente, i.e λ diminue, de même $\beta_2 \to -2$. De plus, nous remarquons aussi que les autres coefficients ne seront pas nuls

Colonnes de X corrélées

Dans cette section, nous allons générer une matrice à moitié colonnes dupliquées à peu près (a deuxième partie diffère de la première d'un petit bruit gaussien de variance $\sigma^2 = 0.01$)

Fixant $\lambda = 1$, et effectuant trois régressions avec $\alpha \in \{0, 0.5, 1\}$

Validation croisée

Dans cette section nous allons essayer d'examiner la pertinence d'avoir fixé $\lambda = 1$ via la méthode de la validation croisée (*cross validation*).
En implémentant la fonction `cv.glmnet()` de **R**, on obtient $\lambda_{optimale} = 1.1568$ (optimalité au sens de minimisation MAE).
L'estimateur de notre modéle en utilisant $\lambda_{optimale}$:

$$\beta = (1.16449, 6.58097, 0, ..., \beta_i, ..)$$

avec $\beta_{36} = -0.35851, \beta_{59} = 0.44253, \beta_{85} = -0.23664, ..etc$

On remarque que notre $\hat{\beta}_{ElsaticNet}$ simulée avec $\lambda = 1$ est très proche de celle calculée à travers la validation croisée. En effet, nous nous sommes retrouvés dans un cas où notre λ choisie arbitrairement au début est proche de celle trouvée grâce à la validation croisée. Or on remarque qu'un centième près les valeurs données par $\lambda = 1$ approche bien notre β,on peut interpréter ce constat par le fait que notre jeu de données n'est constitué que de quelques observations, en l'occurrence n = 10, donc la validation croisée n'a pas pu avoir un grand effet sur notre modèle.

2 Données réelles

Boston Housing Data

Dans cette section, nous allons appliquer une régression LASSO par validation croisée au jeu de données de **Boston** étudié en TP1.

L'implémentation dans **R**, nous donne $\lambda_{optimale} = 0.05130$, et

$\beta = (12.6666, 0.03623, -0.07079, -0.58526, -6.96697, 0.22949, *, -0.78781, 0.51435, *,$

-0.18801, -0.00754, 0.12527, -0.15792)

* Signifie que le modèle a supprimer ce critère (caractère anguleux de la *penalty* l_1 cf. section fonction glmnet).
$\beta \in \mathbb{R}^{14}$ car le modèle a ajouté l'ordonnée à l'origine, de plus le critère ratio élèves par enseignants n'étant pas éliminé par le modèle, $\beta_{ptratio} = -0.18801$,

on ne peut pas dire alors qu'il est significativement non nul.

Effectuant cette fois une régression LASSO avec $\lambda = 1$, on obtient :

$$\beta = (-0.503749, *, *, *, *, *, *, *, *, -0.00285, 0.11608, -0.02140)$$

On remarque que cette fois ci, le ratio élèves enseignants est éliminé par le modéle, donc il n'est pas significativement non nul.

Forest Fires

Nous allons étudier la base de donnée **Forest Fires**, pour plus de détails sur ce jeu de données : Lisez le fichier forestfires.names

Dans ce deuxième TP, nous allons effectuer des régressions avancées (régularisées), comme mentionné dans le rapport du TP1, les colonnes *day, month* sont catégorielles. Or la fonction glmnet n'encode pas ces données automatiquement comme la fonction lm, nous allons donc les supprimer de la base de données qu'on va étudier.
Il serait donc moins pertinent de comparer nos résultats avec ceux du TP1, pour cette raison nous allons présenter une régression similaire à celle faite dans TP1 dans ce rapport avec notre jeu de données modifié

Le tableau suivant décrit les deux modèles de régression qu'on va appliquer

Elastic Net	$\alpha = 0.5$ et $\lambda = 2.1247$
LASSO	$\alpha = 1$ et $\lambda = 6.222$

À noter que les valeurs de λ sont obtenues via une validation croisée

Les coefficients sont résumés dans les figures suivantes

	LASSO alpha = 1
Interept	-3.0219243
X	1.5060734
Y	0.1916102
FFMC	0.0000000
DMC	0.0472505
DC	0.0000000
ISI	-0.2708310
temp	0.5582152
RH	-0.1781454
wind	0.6379911
rain	0.0000000

	Elastic Net alpha = 0.5
Intercept	-2.3181345
X	1.4333504
Y	0.1788809
FFMC	0.0000000
DMC	0.0442051
DC	0.0000000
ISI	-0.2121900
temp	0.5342467
RH	-0.1691938
wind	0.5273309
rain	0.0000000

```
Call:
lm(formula = area ~ X + Y + FFMC + DMC + DC + ISI + temp + RH +
    wind + rain, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
 -39.36  -15.89   -8.54   -0.29 1065.45

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.369315  63.019477  -0.101    0.920
X            1.907945   1.448367   1.317    0.188
Y            0.569181   2.736327   0.208    0.835
FFMC        -0.039200   0.660756  -0.059    0.953
DMC          0.077335   0.067161   1.151    0.250
DC          -0.003295   0.016452  -0.200    0.841
ISI         -0.713739   0.771506  -0.925    0.355
temp         0.800213   0.787185   1.017    0.310
RH          -0.230645   0.237305  -0.972    0.332
wind         1.557431   1.670088   0.933    0.352
rain        -3.404037   9.680970  -0.352    0.725

Residual standard error: 63.58 on 506 degrees of freedom
Multiple R-squared:  0.02164,   Adjusted R-squared:  0.002305
F-statistic: 1.119 on 10 and 506 DF,  p-value: 0.3454
```

FIGURE 2 – Coefficients de régression

D'abord, on remarque que les régressions LASSO et Elastic Net ont se permis de supprimer quelques critères du modèle (un coefficient avec 0 signifie qu'il a été eliminé par le modèle), la régression via la fonction lm nous a donnée un modèle **très médiocre** , aucun des coefficients n'a une signification statistique et on ce modèle ne peut pas être mieux que le modèle constant.
De plus, on remarque que les deux modèles de régression avancée, ont éliminé quatre critères, en l'occurrence ***Y, FFMC, DC rain***.

Conclusion

Dans ce TP, nous avons appliqué les différentes formes de la régression régularisée
Puis nous avons exposé la validation croisée et l'avons appliqué en guise de choisir les bons paramètres pour le modéle (*hyperparameter tuning*)
Enfin, nous avons appliqué ces techniques aux jeu de données déjà traités dans
Un script en **R** permettant de tester et d'appliquer ce qui a été fait dans ce rapport sera joint à ce document.