$\|X\|_2$ always L_2 norm.

TELECOM ParisTech
2018/2019

SDTSIA204
François Portier

EXERCISE CLASS : Linear regression

For $i = 1, \dots, n$, we consider $y_i \in \mathbb{R}$ and $x_i = (x_{i,0}, \dots, x_{i,p})^T \in \mathbb{R}^{p+1}$ with $x_{i,0} = 1$. The OLS estimator is any coefficient vector $\hat{\theta}_n = (\hat{\theta}_{n,0}, \dots, \hat{\theta}_{n,p})^T \in \mathbb{R}^{p+1}$ such that

$$\hat{\theta}_n \in \underset{\theta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2.$$

$x_1 = (x_{1,0}, \dots, x_{1,p})^T$
$x_1^T = (x_{1,0}, \dots, x_{1,p})$

With the notations

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{1,0} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,0} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}, \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

We have

$$\hat{\theta}_n \in \underset{\theta \in \mathbb{R}^{p+1}}{\arg\min} \|Y - X\theta\|. \tag{1}$$

Let $X = (1_n, \tilde{X})$ and introduce $\hat{\mu}_X = (\tilde{X}^T 1_n)/n$ and $\hat{\mu}_Y = (1_n^T Y)/n$. Define the centred version of Y and $\tilde{X}$, given by $Y_c = Y - 1_n \hat{\mu}_Y$ and $\tilde{X}_c = \tilde{X} - 1_n \hat{\mu}_X^T$, respectively. Consider the following alternative procedure :

$$\hat{\theta}_{n,c} = \underset{\theta \in \mathbb{R}^p}{\arg\min} \|Y_c - \tilde{X}_c \theta\|, \tag{2}$$

for which, the predictor at $\tilde{x} \in \mathbb{R}^p$ is given by $\hat{\mu}_Y + (\tilde{x} - \hat{\mu}_X)^T \hat{\theta}_{n,c}$.

Exercise 1. *Aim is to show that*

$$\min_{\tilde{\theta} \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \tilde{\theta}\| = \min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|.$$

$\leq \; \geq$

and, assuming that X has full rank, we have the following relationship between the traditional OLS and the OLS based on centred data,

$$\hat{\theta}_{n,0} = \hat{\mu}_Y - \hat{\mu}_X^T \hat{\theta}_{n,c}, \qquad (\hat{\theta}_{n,1}, \dots, \hat{\theta}_{n,p}) = \hat{\theta}_{n,c}^T. \tag{3}$$

Consequently, the 2 methods give the same predictor. $\|Z - z1_n\| =$

1) *Start by obtaining that the inequality $\geq$ holds true.*
2) *Then show that for any sequence (z_i), and for all $z \in \mathbb{R}$, it holds that $\|Z - z1_n\| \geq \|Z - \overline{z}^n 1_n\|$, where $Z = (z_1, \dots, z_n)$ and $\overline{z}^n = n^{-1} \sum_{i=1}^n z_i$.*
3) *Find $\hat{a}_n$ such that, for any $\theta_0 \in \mathbb{R}$ and $\tilde{\theta} \in \mathbb{R}^p$, $\|Y - \theta_0 1_n - \tilde{X}\tilde{\theta}\| \geq \|Y - \hat{a}_n(\tilde{\theta}) 1_n - \tilde{X}\tilde{\theta}\|$.*
4) *Conclude that $\min_{\theta \in \mathbb{R}^p} \|Y_c - \tilde{X}_c \theta\| = \min_{\theta \in \mathbb{R}^p, \theta_0 \in \mathbb{R}} \|Y - X(\theta_0, \theta^T)^T\|$*
5) *Use the Lebesgue projection theorem to conclude that whenever $\ker(X) = \{0\}$, we have (3).*

Exercise 2 (on-line ols and cross-validation). *The goal of this exercise is to show that the OLS estimator $\hat{\theta}_n$ associated with design matrix $X_{(n)} \in \mathbb{R}^{n \times (p+1)}$ and output $\boldsymbol{y}_{(n)} \in \mathbb{R}^n$ can be easily updated when a new pair of observation $(\boldsymbol{x}_{n+1}^T, y_{n+1}) \in \mathbb{R}^{(p+1)} \times \mathbb{R}$ is given. We apply the result to cross validation procedure in the end.*

To clarify the notation :

$$X_{(n+1)} = \begin{pmatrix} X_{(n)} \\ \boldsymbol{x}_{n+1}^T \end{pmatrix} \in \mathbb{R}^{(n+1) \times (p+1)}, \quad \text{and} \quad \boldsymbol{y}_{(n+1)} = \begin{pmatrix} \boldsymbol{y}_{(n)} \\ y_{n+1} \end{pmatrix} \in \mathbb{R}^{n+1}$$

We assume from now on that $X_{(n)}$ and $X_{(n+1)}$ are full column rank (i.e., the columns of each matrix are independent vectors).

NB : Some of the questions require some computation (in particular obtaining (4) and (6)). Even if you could not prove it, it can be use later.

1) *Let A, B, C, D be matrices with respective sizes (d,d), (d,k), (k,k), (k,d). Show that if A and C are invertible, then*

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}. \tag{4}$$

2) *Obtain that*

$$(X_{(n+1)}^T X_{(n+1)})^{-1} = (X_{(n)}^T X_{(n)})^{-1} - \frac{\zeta_{n+1}\zeta_{n+1}^T}{1+b_{n+1}} \tag{5}$$

where $\zeta_{n+1} = (X_{(n)}^T X_{(n)})^{-1}\boldsymbol{x}_{n+1}$ *and* $b_{n+1} = \boldsymbol{x}_{n+1}^T (X_{(n)}^T X_{(n)})^{-1}\boldsymbol{x}_{n+1}$.

3) *Express* $X_{(n+1)}^T \boldsymbol{y}_{(n+1)}$ *with respect to* $X_{(n)}^T \boldsymbol{y}_{(n)}$ *and* $y_{n+1}\boldsymbol{x}_{n+1}$.

4) *Show that the OLS estimator* $\hat{\boldsymbol{\theta}}_{n+1}$ *associated with design matrix* $X_{(n+1)}$ *and output* $\boldsymbol{y}_{(n+1)}$ *can be obtained as follows :*

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n + \frac{u_{n+1}}{1+b_{n+1}}\zeta_{n+1}, \tag{6}$$

where $u_{n+1} = y_{n+1} - \boldsymbol{x}_{n+1}^T\hat{\boldsymbol{\theta}}_n$.

5) *Keeping in memory* $(X_{(n)}^T X_{(n)})^{-1}$ *and* $\hat{\boldsymbol{\theta}}_n$, *explain how to update* $\hat{\boldsymbol{\theta}}_{n+1}$ *using a minimal number of operations of the kind : matrix* $(p+1, p+1)$ *times vector* $(p+1, 1)$. *How many such operation are needed ?*

6) *Using Equation* (5) *above, show that*

$$1 + b_{n+1} = \frac{1}{1-h_{n+1}}$$

where $h_{n+1} = \boldsymbol{x}_{n+1}^T (X_{(n+1)}^T X_{(n+1)})^{-1}\boldsymbol{x}_{n+1}$.

7) *The prediction of* y_{n+1} *given by the model is* $\hat{y}_{n+1} := \boldsymbol{x}_{n+1}^T\hat{\boldsymbol{\theta}}_{n+1}$. *With the following formula*

$$\hat{y}_{n+1} = \boldsymbol{x}_{n+1}^T\hat{\boldsymbol{\theta}}_n + \frac{u_{n+1}b_{n+1}}{1+b_{n+1}}.$$

prove that

$$y_{n+1} - \hat{y}_{n+1} = u_{n+1}(1-h_{n+1}).$$

8) *Given some data* $(\boldsymbol{y}, X)$, *leave-one-out cross-validation consists in computing the risk*

$$R_{cv} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}_{(-i)})^2$$

where $\hat{\boldsymbol{\theta}}_{(-i)}$ *is the OLS estimator based on* $(\boldsymbol{y}_{(-i)}, X_{(-i)})$, *i.e., the data* $(\boldsymbol{y}, X)$ *without the i-th line. Applying what have been done so far, show that*

$$R_{cv} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(1-\hat{h}_i)^2,$$

with $\hat{h}_i = \boldsymbol{x}_i^T(X^TX)^{-1}\boldsymbol{x}_i$ *and* $\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}_n$, $\hat{\boldsymbol{\theta}}_n$ *being the OLS estimator of* $(\boldsymbol{y}, X)$.