KERNEL METHODS - PRACTICAL SESSION № 1

AMMI 2020 18/05/2020

Linear Algebra Recap

Exercise 1

Let $A \in \mathbb{R}^{m \times n}$ be a real matrix. For each function f, **specify its codomain and compute** ∇f

(a) $f : \begin{cases} \mathbb{R} \to ? \\ x \to xA \end{cases}$

(b) f: $\begin{cases} \mathbb{R}^n \to ? \\ \mathrm{x} \to A\mathrm{x} \end{cases}$

(c) f: $\begin{cases} \mathbb{R}^{m \times n} \to ? \\ X \to Tr(A^T X) \end{cases}$

Exercise 2

Let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$.
Prove that $M = X^T X + \lambda I_p$ **is invertible.**
Hint: Prove that the eigenvalues of M *are larger than* λ

Linear Regression

Using notations from the slides:

- $Y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ the vector of outcomes
- $X = (\mathrm{x}_1, ..., \mathrm{x}_n)^T \in \mathbb{R}^{n \times p}$ the matrix (n rows=samples, p columns=features)
- $\beta \in \mathbb{R}^p$ the linear model's parameters

Exercise 3

(a) State the loss function for ordinary least squares (OLS)

(b) Formulate OLS as a minimization problem

(c) Compute the solution $\hat{\beta}^{OLS}$

(d) What happens if $X^T X$ is singular (not invertible)?

Exercise 4 - Bias and Variance of OLS

slide 16

Exercise 5 - Optimality of OLS

slide 17

Ridge Regression

Exercise 6 - Bias and Variance

slide 33

Exercise 7 - Performance

slide 34

Ridge Logistic Regression

Exercise 8

slide 46

EXERCISE 01

① $A \in \mathbb{R}^{m \times m}$

a] $f: \begin{cases} \mathbb{R} \longrightarrow \mathbb{R}^{m \times m} \\ x \longrightarrow xA \end{cases}$

$\nabla f_{ij} = f'_{ij}(x) = A_{ij}$; $f_{ij}(x) = A_{ij}\, x \Rightarrow \boxed{\nabla f = A}$

b] $f: \begin{cases} \mathbb{R}^m \longrightarrow \mathbb{R}^m \\ x \rightsquigarrow Ax \end{cases}$

$\nabla f_{ij} = \dfrac{\partial f_i}{\partial x_j}$

$i \in (1, m)$
$j \in (1, m)$

$f_i(x) = A_i^T x$
$= \sum_k A_{ik} x_k$
$= A_{ij} x_j$

$= A_{ij}$

$\Rightarrow \boxed{\nabla f = A}$

c] $f: \begin{cases} \mathbb{R}^{m \times m} \longrightarrow \mathbb{R} \\ X \rightsquigarrow Tr(A^T X) \end{cases}$

$Tr(M) = \Sigma$ of diagonal elements of M.

$A^T X : \mathbb{R}^{m \times m} \times \mathbb{R}^{m \times m} \to \mathbb{R}^{m \times m}$

$Tr : \mathbb{R}^{m \times m} \to \mathbb{R}$

$\nabla f = \dfrac{\partial f}{\partial x_{ij}}$ with $f(x) = \Sigma A_{ij} X_{ij}$

$\dfrac{\partial f}{\partial x_{ij}} = A_{ij} \Rightarrow \boxed{\nabla f = A}$

If we have $f: \mathbb{R}^m \longrightarrow \mathbb{R}$

$f(x) = f(x_0) + (x - x_0)^T \nabla f(x_0) + O(\|x - x_0\|)$
$\approx f(x_0) + f'(x_0)(x - x_0) + O(x - x_0)$

$\nabla' f_{ij} = \dfrac{\partial^2 f}{\partial x_{ij}}$

EXERCISE 2

Let $X \in \mathbb{R}$, $\lambda > 0$. Proove that $M = X^T X + \lambda I_p$ is invertible.

LEMMA $M' = X^T X$ eigen values u of M' are ≥ 0.

$M' = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ $\det A = \prod_{\lambda \in Eig(A)} \lambda$

$u \in Eig(M')$, y eigen vector

$M'y = uy$

$y^T M' y = y^T u y$, $u \geq 0$

$y^T M' y = u y^T y$

$y^T M' y = y^T X^T X y$

$= (Xy)^T Xy = \|Xy\|^2$

$y^T y = \sum y_i^2 = \|y\|^2$

$u = \frac{\|Xy\|}{\|y\|^2}$

$\Rightarrow u = \frac{\|Xy\|^2}{\|y\|^2} \geq 0$ $\Rightarrow$ we show that u are not neg.

$u \in Eig(M') \Leftrightarrow u + \lambda \in Eig(M)$, $u \geq 0$

if $\lambda' \in Eig(M)$, $\lambda' \geq \lambda > 0$

All the eigenValues of M are ≥ 0 $\Rightarrow$ M is inversible

EXERCISE 3

a) State the loss function for ordinary least squares (OLS)

$y = \beta^T x + \varepsilon$

$= \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$

$\varepsilon = y - \beta^T x$

The Criterion : $MSE(\beta) = \frac{1}{n} \sum (y_i - \beta^T x_i)^2$

$= \frac{1}{n} (y - X\beta)^T (y - X\beta)$

b) Formulate OLS as a minimization Problem.

OLS = ordinary Least Square

$$\min_{\beta} MSE(\beta), \quad MSE : \mathbb{R}^p \longrightarrow \mathbb{R}$$

c) Compute the solution $\hat{\beta}^{OLS}$

$$\nabla MSE(\tilde{\beta}) = 0$$

$$\nabla MSE = \frac{1}{n}\left[X^T(Y - X\beta)\cdot(Y - X\beta)^T X\right] = \frac{2}{n} X^T(X\beta - Y)$$

$$\nabla MSE_i = \frac{\partial MSE}{\partial \beta_i}$$

OLS convex quadratic reaches optimum when $\nabla MSE(\beta^*)$

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

$$X^T X \beta^* = X^T y$$

If $X^T X$ non singular $\beta^* = (X^T X)^{-1} X^T y$

$A^T B = B^T A$.

$$MSE = -\frac{1}{n}\left[X^T(Y - X\beta) + (Y - X\beta)^T X\right] = -\frac{2}{n} X^T(Y - X\beta)$$

d) What happens if $X^T X$ is singular (not invertible)

when $X^T X$ is singular

⇒ pseudo inverse → Find solution when there is no solution.

⇒ Regularize with small λ; ridge regression no closed forme / unstable

EXERCISE 4 - BIAS AND VARIANCE OF OLS (SLIDE 16)

Show that $\hat{\beta}^{OLS} = (X^T X)^{-1} X^T y$

Satisfies $\begin{cases} E(\hat{\beta}^{OLS}) = \beta^* \\ Var(\hat{\beta}^{OLS}) = E(\hat{\beta}^{OLS} - \beta^*)(\hat{\beta}^{OLS} - \beta^*)^T \end{cases}$

$$Var(A) = E\left[(A - E(A))(A - E(A))^T\right]$$

1-dimension $Var(A) = E(A - E(A))$

$$Var(A)_{ij} = CoV(A_i, A_j) = E(A_i, A_j) = E(A_i)E(A_j)$$

$$y = X\beta^* + \varepsilon \quad \text{with } E(\varepsilon) = 0 \; ; \; Var(\varepsilon) = \sigma^2 I$$

$$CoV(\varepsilon_i, \varepsilon_j) = 0 \quad \text{if } i \neq j$$

$$Var(\varepsilon_i) = \sigma^2$$

$$\hat{\beta}^{OLS} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^TX\beta^* + (X^TX)^{-1}X^T\varepsilon$$

$$= \beta^* + \underbrace{(X^TX)^{-1}X^T}_{A}\varepsilon$$

E linear

$$E[\hat{\beta}^{OLS}] = \beta^* + A\underbrace{E[\varepsilon]}_{E(A\varepsilon)} = \beta^* \Rightarrow \hat{\beta}^{OLS} \text{ is unbiased}$$

$$Var(A\varepsilon) = E\left((A\varepsilon)^*(A\varepsilon)^T\right) = E(A\varepsilon\varepsilon^TA) = A\underbrace{E(\varepsilon\varepsilon^T)}_{\sigma^2 I}A^T = \sigma^2 AA^T$$

$$A = (X^TX)^{-1}X^T$$

$$AA^T = (X^TX)^{-1}(X^TX)(X^TX)^{-1} = (X^TX)^{-1}$$

$$Var(\hat{\beta}) = Var(A\varepsilon) = \underline{\sigma^2(X^TX)^{-1}}$$

EXERCISE 5

Gauss Markov theorem

Least squares estimator $\hat{\beta}^{OLS}$ is a Best Linear unbiased Estimator.

Assumption

$$y = X\beta^* + \varepsilon$$

$$E\varepsilon = 0 \qquad Var(\varepsilon) = \sigma^2 I$$

if $\tilde{\beta} = \underset{\text{is a matrix}}{C}y$ unbiased $(E\tilde{\beta} = \beta^*)$ then $Var(\hat{\beta}^{OLS}) \leq Var(\tilde{\beta})$

$$\triangle \neq \hat{\beta}^{OLS}_{ij} \leq \tilde{\beta}_{ij}$$

$M = Var(\tilde{\beta}) - Var(\hat{\beta}^{OLS})$ is a positive semidefinite matrix

$\Leftrightarrow \forall x_0, x_0^T M$

$\hat{B} = Cy$ Candidate unbiased estimator

$D = C - (X^T X)^{-1} X^T$

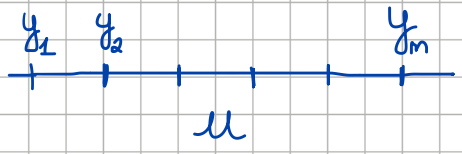
$E(\hat{B}) = B^* = E(Cy) = C E(y) = C X B^*$, $y \in XB^* + \varepsilon$ and $E(y) = E(XB^*) + E(\varepsilon) = XB^*$ ($E(\varepsilon) = 0$)

$\Rightarrow \hat{B}^{OLS}$ is unbiased (Ex 4)
is Best Linear unbiased

$i = [1, m]$ $\quad y_i \in \mathbb{R}$ $\quad y = u + \varepsilon$

$y_i = u + \varepsilon_i$ $\quad E(\varepsilon_i) = 0$

Find an unbiased estimator of u



$\Rightarrow \hat{y}_i = \frac{1}{n} \sum y_i$ $\quad y_i = u + \varepsilon_i$, $E[y_i] = u$

Unbiased

$E(\hat{y}) = \frac{1}{n} \sum E(y_i) = \frac{1}{n} \sum_{i=1}^{n} u = u$

$\hat{y} = y_i$ is also an unbiased estimator

$E(\hat{y}) = E(y_i) = u$ Now we have 2 unbiased estimator, Which one is better

Their Variance are different $\quad Var(\hat{y}_2) \leq Var(\hat{y}_1)$

$Var(y_2) \subset Var(\varepsilon)$

$Var\, \hat{y}_1 = E((\hat{y}_1 - E(y_1))^2) = E\left(\frac{1}{n} \sum (y_i - u)^2\right)$

$Var(\hat{y}_i) = \frac{1}{n} Var(\varepsilon) \leq Var(\hat{y}_2)$

$\hat{y}_i$ is better (Smaller over all error), Error = bias^2 + variance

Back to (Ex 5)

$\hat{\beta}^{OLS}$ is unbiased we want to know if it is the best linear unbiased estimator

we need to take $\tilde{\beta} = Cy$ linear unbiased estimator.
show that $\hat{\beta}^{OLS}$ is better than $\tilde{\beta}$, $Var(\hat{\beta}^{OLS}) \leq Var(\tilde{\beta})$

$\tilde{\beta} = Cy$ $\qquad D := C - (X^TX)^{-1}X^T$

unbiased $E(\tilde{\beta}) = \beta^*$

$$E(\tilde{\beta}) = E(Cy) = CE(y), \quad y = X\beta^* + \varepsilon$$
$$= CX\beta^* + C\underbrace{E(\varepsilon)}_{=0}$$
$$= CX\beta$$
$$\beta^* = CX\beta^*, \quad (CX - I)\beta^* = 0$$

$DX = CX - (X^TX)^{-1}X^TX = CX - I$

$DX\beta^* = 0$

$\tilde{\beta}$ unbiased for all $\beta^* \Rightarrow DX = 0$

$Var(\tilde{\beta}) = E((\tilde{\beta} - \beta^*)(\tilde{\beta} - \beta^*)^T)$

$\tilde{\beta} = Cy = (D + (X^TX)^{-1}X^T)(X\beta^* + \varepsilon) = \beta^* + D\varepsilon + (X^TX)^{-1}X^T\varepsilon$

$Var(\tilde{\beta}) = E((\tilde{\beta} - \beta^*)(\tilde{\beta} - \beta^*)^T) = D\varepsilon + (X^TX)^{-1}X^T\varepsilon$

$$= E(D\varepsilon(D\varepsilon)^T + 2(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1} + D\varepsilon\varepsilon^TX(X^TX)^{-1} + (X^TX)^{-1}X^T\varepsilon\varepsilon^TD^T)$$
$$= DE(\varepsilon\varepsilon^T)D^T(X^TX)^{-1}X^TE(\varepsilon\varepsilon^T)X(X^TX)^{-1} + DE(\varepsilon\varepsilon^T)X(X^TX)^{-1} + (X^TX)^{-1}X^TE(\varepsilon\varepsilon^T)D^T$$
$$= \sigma^2 DD^T + \sigma^2(X^TX)^{-1} + 0 + 0$$

$Var(\tilde{\beta}) = \sigma^2 DD^T + Var(\hat{\beta}^{OLS})$, we want to show that M is posit. definite

$Var(\tilde{\beta}) = Var(\hat{\beta}^{OLS}) = \sigma^2 DD^T$; $Var(\tilde{\beta}) \leq Var(\hat{\beta}^{OLS}) \Leftrightarrow X_0^TMX_0 \geq 0$

$X_0^TMX_0 = \sigma^2 X_0^T DD^T X_0 = \sigma^2 (D^TX_0)^T D^TX_0 = \sigma^2 \|D^Tx_0\|^2 \geq 0, \ \forall x_0$

$Var(\tilde{\beta}) \geq Var(\hat{\beta}^{OLS})$ then $\hat{\beta}^{OLS}$ is BLUE

Ridge Regression

EXERCISE 6

Slides 33

$$\begin{cases} \text{bias}\left(\hat{\beta}^{\text{Ridge}}\right) = ? = \frac{-\lambda}{1+\lambda}\beta^* \\ \text{Var}\left(\hat{\beta}_2^{\text{Ridge}}\right) = ? = \frac{1}{(1+\lambda)^2}\text{Var}\left(\hat{\beta}^{OLS}\right) \end{cases}$$

$$L(\beta) = MSE(\beta) + \lambda\|\beta\|^2 = \frac{1}{n}\sum (y_i - \beta^T x_i)^2 + \lambda\|\beta\|^2$$

Ridge regression loss

$$\nabla L(\beta) = 0$$

$$\hat{\beta} = (X^TX + \lambda n I)^{-1} X^T y$$

$\lambda > 0$, $X^TX + \lambda n I$ is inversible.

$$\nabla L(\beta) = \frac{2}{n} X^T(X\beta - y) + 2\lambda\beta$$

$$\nabla = 0 \Rightarrow (X^TX + \lambda n I)\hat{\beta} = X^T y$$

$$E(\hat{\beta}^{\text{Ridge}}) = (X^TX + \lambda n I)^2 X^T \underbrace{E(Y)}_{X\beta^*}$$

$$= (X^TX + \lambda n I)^{-1} \underbrace{X^TX}_{X^TX + \lambda n I = \lambda_n I} \beta^*$$

$$= \beta^* - \lambda n (X^TX + \lambda n I)^{-1}\beta^*$$

Assumption : $X^TX = nI$

$$E(\hat{\beta}^{\text{Ridge}}) = \beta^* - \frac{\lambda}{1+\lambda}\beta^*$$

$$\text{bias}(\hat{\beta}^{\text{Ridge}}) = -\frac{\lambda}{1+\lambda}\beta^*$$

$\Rightarrow$ when $\lambda = 0$ (no regularization), bias $= 0$, $\hat{\beta}_\lambda^{\text{Ridge}} = \hat{\beta}^{OLS}$

When $\lambda = +\infty$, bias $= -\beta^*$ because $\hat{\beta}_\lambda^{\text{Ridge}} = 0$

Assumption $X^T X = nI$

$$X_i^T X_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

$$\hat{\beta}_\lambda^{Ridge} = \frac{1}{n(1+\lambda)} X^T y$$

$$Var(\hat{\beta}_\lambda^{Ridge}) = \frac{1}{n^2(1+\lambda)^2} Var(X^T y) = \frac{1}{n^2(1+\lambda)^2} \underbrace{Var(X^T Z)}_{n\sigma^2}$$

$$Var(X^T Z) = E(X^T \varepsilon \varepsilon^T X) = X^T Var(\varepsilon) X = \sigma^2 X^T X = n\sigma^2$$

$$Var(\hat{\beta}_\lambda^{RIDGE}) = \frac{\sigma^2}{n(1+\lambda)^2} \quad ; \quad Var(\hat{\beta}^{OLS}) = (X^T X)^{-1}\sigma^2 = \frac{\sigma^2}{n}$$

$$Var(\hat{\beta}_\lambda^{Ridge}) = \frac{1}{(1+\lambda)^2} Var(\hat{\beta}^{OLS})$$

when $\lambda \to 0$: Same Variable (Same estimator)

$\lambda \to +\infty$: Variable $\to 0$ (estimator $\hat{\beta} x = 0$)

EXERCISE 7.

$$f(\lambda) \quad \mathbb{E}\left[bias^2(x^T \beta_\lambda^{Ridge}) + Var(x_0^T \hat{\beta}_\lambda^{Ridge})\right]$$

1 minimum for $\lambda \in \lambda^* = \frac{\sigma^2 p}{n \|\beta^*\|^2}$

(Compute $f'(\lambda)$)

$f(\lambda)$ = error function $\exists$ optimal λ^* that gives the best error

Cannot compute it directly. Find with cross-validation

$$f(\lambda^*) \leq \min(f(0), f(+\infty))$$

Using λ^* is better than not using regularization.