TD : Supervised Learning - Partie théorique

1. OLS

On a : $\beta^* = (x^T x)^{-1} x^T y = Hy$ où on suppose $y = x\beta + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$

on considère $\tilde{\beta} = (H + D) y$. Il vient

$\mathbb{E}(\tilde{\beta}) = (H+D) x\beta = (I + Dx)\beta$ où I est l'identité

On suppose que $\mathbb{E}(\tilde{\beta}) = \beta$. on en déduit donc que $Dx = 0$

on calcule ensuite :

$Var(\tilde{\beta}) = Var(Cy) = C\, Var(y)\, C^T = \sigma^2 C C^T$ d'après notre hypothèse

d'où $Var(\tilde{\beta}) = \sigma^2 ((x^T x)^{-1} x^T + D)((x^T x) x^T + D)^T$

En développant et en utilisant $Dx = 0$ il reste

$Var(\tilde{\beta}) = \underbrace{\sigma^2 (x^T x)^{-1}}_{Var(\beta^*)} + \sigma^2 (D D)^T$

On a ainsi prouvé que $Var(\tilde{\beta}) > Var(\beta^*)$

2. Ridge.

2.1) Cette fois-ci on prend $\beta^* = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c$

β^* ainsi défini est, pour rappel, l'estimateur de Ridge

En remarquant que $\mathbb{E}(y_c) = x_c \beta$, on a par linéarité :

$\mathbb{E}(\beta^*) = (x_c^T x_c + I\lambda)^{-1} x_c^T x_c \beta$

Par $\lambda \neq 0$, $\mathbb{E}(\beta^*) \neq \beta$.

et dans ces cas l'estimateur est biaisé.

2.2. ~~On~~ On écrit $x_c = UDV^T$ de sorte que

$\beta^* = ((UDV^T)^T (UDV^T) + \lambda I)^{-1} (UDV^T)^T y_c$

$= (((V D^T U^T)(UDV^T)) + \lambda I)^{-1} (UDV^T)^T y_c$

$= V (D^T D + \lambda I)^{-1} V^T V D^T U^T y_c = V (D^T D + \lambda I)^{-1} D^T U^T y_c$

Avec cette forme, on n'a plus besoin de recourir à l'inverse d'une matrice puisqu'on peut utiliser le fait que $(D^T D + \lambda I)^{-1} D^T$ est une matrice de la forme $\begin{cases} \dfrac{\lambda_h}{\lambda_h^2 + \lambda} & \text{sur la diagonale} \\ 0 & \text{sinon} \end{cases}$, où les λ_i sont les valeurs propres

2-3. On a que

$$Var(\beta^{\lambda}_{ridge}) = \left((x_c^T x_c + \lambda I)^{-1} x_c^T\right) \underbrace{Var(y_c)}_{6^2 I} \left((x_c^T x_c + \lambda I)^{-1} x_c^T\right)^T$$

$$= 6^2 (x_c^T x_c + \lambda I)^{-1} x_c^T \left(x_c (x_c^T x_c + \lambda I)^{-1}\right)$$

Comme $Var(\beta^{\alpha}_{ols}) = 6^2 (x^T x)^{-1}$, on a : $Var(\beta^{\alpha}_{ols}) \geq Var(\beta^{\alpha}_{ridge})$

2-4. Plus λ augmente, plus le biais augmente et la variance diminue.