TP Intro Supervised

gael.marcheville

March 2023

1 Theoretical questions

1.1 OLS

Pour calculer la valeur attendue de $\tilde{\beta}$, on a :

$$\mathrm{E}[\tilde{\beta}] = E[Cy] = CE[y] = (H+D)E[y] = HE[y] + DE[y]$$

Pour calculer la variance de $\tilde{\beta}$, on a :

$$\mathrm{Var}(\tilde{\beta}) = E[(\tilde{\beta} - E[\tilde{\beta}])(\tilde{\beta} - E[\tilde{\beta}])^\top]$$
$$= E[(Cy - E[Cy])(Cy - E[Cy])^\top]$$
$$= E[(Cy - (H+D)E[y])((Cy - (H+D)E[y])^\top]$$
$$= E[(Cy - HE[y] - DE[y])(y^\top C^\top - E[y]^\top (H+D)^\top)]$$
$$= E[((H+D)(y - E[y]))((y - E[y])^\top (H+D)^\top)]$$
$$= (H+D)E[(y - E[y])(y - E[y])^\top](H+D)^\top$$
$$= (H+D)Var(y)(H+D)^\top$$

Nous faisons alors l'hypothèse que $Var(y) = \sigma^2 I_n$. On a alors :

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2 (H+D)(H+D)^\top$$

Pour montrer que $Var(\beta^*) < Var(\tilde{\beta})$, on va utiliser que $(H+D)(H+D)^\top \geq HH^\top$, qui découle du fait que $\forall$ matrices A et B, on a $ABAB^\top \geq ABA^\top B^\top$ (théorème de Schur). Ainsi :

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2 (H+D)(H+D)^\top \geq \sigma^2 HH^\top = \sigma^2 Var(\beta^*)$$

Par conséquent, on a montré que $Var(\beta^*) < Var(\tilde{\beta})$, et l'hypothèse que nous avons utilisé est $Var(y) = \sigma^2 I_n$.

1.2 Ridge regression

- La forme explicite de la solution de Ridge est $\beta^*_{ridge} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c$, on a donc : $E[\beta^*_{ridge}] = E[(x_c^T x_c + \lambda I)^{-1} x_c^T y_c] = [(x_c^T x_c + \lambda I)^{-1} x_c^T] E[y_c] = [(x_c^T x_c + \lambda I)^{-1} x_c^T x_c]\beta$, ce qui est différent de β sauf si $\lambda = 0$ (cas OLS), c'est donc un estimateur biaisé.

- La décomposition SVD pour l'estimateur Ridge peut être écrite comme suit :

$$\beta^*_{ridge} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c = ([UDV^T]^T[UDV^T] + \lambda I)^{-1}(UDV^T)^T y_c$$

$$= (VD^TU^TUDV^T + \lambda I)^{-1}VD^TU^Ty_c = (VD^TDV^T + \lambda I)^{-1}VD^TU^Ty_c = V(D^TD + \lambda I)^{-1}V^TVD^TU^Ty_c$$

$$\beta^*_{ridge} = V(D^TD + \lambda I)^{-1}D^TU^Ty_c$$

  On utilise le fait que U et V sont des matrices orthogonales (l'inverse est égal à la transposée). Cette transformation est utile pour le calcul (optimiser la vitesse d'execution) car il n'est pas nécessaire d'inverser une matrice, puisque $(D^TD + \lambda I)^{-1}D^T$ est une matrice diagonale.

- La variance de l'estimateur Ridge peut être calculée comme suit :

$$Var(\beta^*_{ridge}) = Var((x_c^T x_c + \lambda I)^{-1} x_c^T y_c)$$

$$Var(\beta^*_{ridge}) = (x_c^T x_c + \lambda I)^{-1} x_c^T) Var(y_c)(x_c^T x_c + \lambda I)^{-1} x_c^T)^T$$

$$Var(\beta^*_{ridge}) = \sigma^2 (x_c^T x_c + \lambda I)^{-1} x_c^T x_c (x_c^T x_c + \lambda I)^{-1}$$

  Pour une valeur de lambda positive, $(x_c^T x_c + \lambda I)$ sera toujours supérieure à $x_c^T x_c$, par conséquent $(x_c^T x_c + \lambda I)^{-1} x_c^T x_c (x_c^T x_c + \lambda I)^{-1}$ sera toujours inférieure à $(x_c^T x_c)^{-1}$, donc $Var(\beta^*_{OLS}) \geq Var(\beta^*_{ridge})$.

- L'estimateur Ridge propose un compromis entre le biais et la variance. Lorsque λ augmente, le biais augmente et la variance diminue. Si nous prenons un λ très proche de zéro, la solution tendra vers la solution des MCO avec un biais de 0 et une variance élevée. Si λ est proche de l'infini, on aura une variance nulle, mais un grand biais.

- On a $\beta^*_{ridge} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c$, si $x_c^T x_c = Id$, $\beta^*_{ridge} = ((1+\lambda)I)^{-1} x_c^T y_c$.
  De plus, $\beta^*_{OLS} = (x_c^T x_c)^{-1} x_c^T y_c$, et comme $x_c^T x_c = Id$, $\beta^*_{OLS} = x_c^T y_c$
  Ainsi, $\beta^*_{ridge} = \frac{\beta^*_{OLS}}{1+\lambda}$

1.3 Elastic Net

Réécriture de l'équation 2 :

$$\beta^*_{ridge} = \arg\min_\beta (y_c - x_c\beta)^T(y_c - x_c\beta) + \frac{\lambda_2}{2}||\beta||_2^2 + \lambda_1||\beta||_1$$

Comme la fonction est strictement convexe, le minimum peut être obtenu en égalant le sous-gradient à zéro. ($\lambda_1||\beta||_1$ n'est pas différentiable en 0).

$$\frac{\partial f}{\partial \beta} = 2x_c^T(y_c - x_c\beta^*_{ElN}) + 2\lambda_2\beta^*_{ElN} + \lambda_1 * signe(\beta^*_{ElN}) = 0$$

$$2x_c^T(y_c - x_c\beta^*_{ElN}) + 2\lambda_2\beta^*_{ElN} \pm \lambda_1 = 0$$

$$2\beta^*_{OLS} - 2\beta^*_{ElN}(1 - \lambda_2) \pm \lambda_1 = 0$$

Comme $x_c^T x_c = Id$ et que $\beta^*_{OLS} = x_c^T y_c$, d'où :

$$\beta^*_{ElN} = \frac{\beta^*_{OLS} \pm \frac{\lambda_1}{2}}{1 - \lambda_2}$$