Introduction to Supervised Learning - IMA205

Nom: **OLIVEIRA MACHADO DE SOUSA** Prénom: **Lucas**

1 OLS

- $E[\tilde{\beta}] = E[Cy] = \beta(I_d + Dx)$, which must be β $(Dx = 0)$ since the estimator is unbiased.

$$\begin{aligned} Var(\tilde{\beta}) &= Var(Cy) = CVar(y)C^T = \sigma^2 CC^T \\ &= \sigma^2(H + D)(H + D)^T = \sigma^2(HH^T + HD^T + DH^T + DD^T) \\ &= \sigma^2(x^T x)^{-1} + \sigma^2(DD^T) \end{aligned}$$

The second term being positive (DD^T is symmetric semipositive), we've shown that OLS has the smallest variance and the inequality of the question holds. We have to assume that $E[\varepsilon] = 0$ or at least $E[x^T \varepsilon] = 0$ and x is deterministic.

2 Ridge Regression

- As we've seen in class, the unique solution for the Ridge Regression if $\beta^*_{ridge} = (x^T x + \lambda I_d)^{-1} x^T y$. Therefore,

$$\begin{aligned} E[\beta^*_{ridge}] &= E[(x_c^T x_c + \lambda I_d)^{-1} x_c^T y_c] \\ &= (x_c^T x_c + \lambda I_d)^{-1} x_c^T E[y_c] \\ &= (x_c^T x_c + \lambda I_d)^{-1} x_c^T x_c \beta \end{aligned}$$

This is not zero, since $\lambda \neq 0$.

- Now, let's write the ridge solution using the SVD decomposition of x_c. This solution is useful when working with big matrices, since we don't have to actually compute an inverse (which is computationally expensive), since $(D^T D + \lambda I)^{-1}$ is a diagonal matrix with its values defined by lambda and the eigenvalues of the matrix.

$$\begin{aligned} \beta^*_{ridge} &= (x_c^T x_c + \lambda I)^{-1} x_c^T y_c((UDV^T)^T(UDV^T)+)^{-1}(UDV^T)^T y_c \\ &= (VD^T U^T UDV^T + \lambda I)^{-1} VD^T U^T y_c \\ &= (VD^T DV^T + \lambda I)^{-1} VD^T U^T y_c \\ &= V(D^T D + \lambda I)^{-1} V^T VD^T U^T y_c \\ &= V(D^T D + \lambda I)^{-1} D^T U^T y_c \end{aligned}$$

- Then let's show that $Var(\beta^*_{OLS}) \geq Var(\beta^*_{ridge})$.

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

  Let's remember that OLS is the same as Ridge with $\lambda = 0$. For a positive value of λ, $(x_c^T x_c + \lambda I) > (x_c^T x_c) \implies (x_c^T x_c + \lambda I)^{-1} < (x_c^T x_c)^{-1}$. Therefore, $Var(\beta^*_{OLS}) \geq Var(\beta^*_{ridge})$.

- About the bias-variance tradeoff of the Ridge estimator, we can start from the fact that OLS is unbiased but has a high variance. If we take ridge with λ tending to zero, the solution is closer to the OLS solution, therefore he have lower bias and higher variance. As λ turns bigger, we get further from the OLS solution, increasing the bias and reducing the variance.

- Last but not least, if $x_c^T x_c = Id$, we have:

$$\beta^*_{ridge} = (Id + \lambda I_d)^{-1} x_c^T y_c = \frac{I_d}{1+\lambda} x_c^T y_c = \frac{\beta^*_{OLS}}{1+\lambda}$$