TP3 IMA205 - Supervised Learning

Luiz Augusto Facury de Souza

March 2023

1 OLS

$$E[\tilde{\beta}] = E[Cy] = \beta(I_d + D_x)$$

Since $\tilde{\beta}$ is unbiased, we can see that D_x must be zero, since $E[\tilde{\beta}] = \beta$ in this case. Therefore, we have:

$$Var(\tilde{\beta}) = Var(Cy) = CVar(y)C^T = \sigma^2 CC^T = \sigma^2((x^Tx)^{-1}x^T + D)(x(x^Tx)^{-1} + D^T)$$

$$= \sigma^2(x^Tx)^{-1} + \sigma^2(x^Tx)^{-1}(D_x)^T + \sigma^2(D_x)(x^Tx)^{-1} + \sigma^2(DD^T)$$

As D_x is zero, we can see that:

$$Var(\tilde{\beta}) = \sigma^2(x^Tx)^{-1} + \sigma^2(DD^T)$$

DD^T is always semi-positive and symmetric, so it is greater or equal to zero, being the OLS case when it is zero. Therefore, knowing that $Var(\beta^*) = \sigma^2(x^Tx)^{-1}$, we can see that:

$$Var(\tilde{\beta}) = Var(\beta^*) + \sigma^2(DD^T)$$

We can see that the variance of the OLS is the smallest if the assumptions hold: x is deterministic, and $E[\epsilon] = 0$

2 Ridge

- We know that the Ridge expression can be written as:

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

  Therefore, we have:

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

  If $\lambda = 0$, we have the same expression as the OLS. However, for any other value of λ, we have a different estimator, and it is biased.

- The SVD expression is given by:

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

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

  Using that U and V are orthogonal matrices. This decomposition us useful, since we do not need to invert a matrix, since $(D^T D + \lambda I)^{-1} D^T$ is equal to a diagonal matrix where each element is given by its eigenvalue divided by ($eigenvalue^2$ + lambda).

- The variance of the Ridge estimator can be expressed as:

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

  We can observe that, for a positive value of λ, the expression $(x_c^T x_c + \lambda I)$ will, in every case, be greater than $x_c^T x_c$. Therefore, we have $(x_c^T x_c + \lambda I)^{-1} x_c^T x_c (x_c^T x_c + \lambda I)^{-1}$ always smaller than $(x_c^T x_c)^{-1}$, which implies:

$$Var(\beta^*_{OLS}) > Var(\beta^*_{ridge})$$

- Any machine learning problem tries to balance the trade-off between bias and variance. The ridge estimator is an alternative to the OLS estimator that increases the bias with the increase of λ and decreases the variance. This can be verified with the fact that, when $\lambda = 0$, we have bias = 0.

- If $x_c^T x_c = I_d$, we have:

$$B^*_{ridge} = (I_d + \lambda I_d)^{-1} x_c^T y_c = ((1 + \lambda) I_d)^{-1} x_c^T y_c$$

  For the OLS, we have:

$$\beta^*_{OLS} = (x_c^T x_c)^{-1} x_c^T y_c = x_c^T y_c$$

  Therefore:

$$\beta^*_{ridge} = \frac{\beta^*_{OLS}}{1 + \lambda}$$

3 Elastic Net

We know that:

$$\beta^*_{ElNet} = argmin_\beta (y_c - x_c \beta)^T (y_c - x_c \beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$

Since we have a strictly convex function, it is possible to compute the minimum by calculating the gradient and equaling it to 0, in order to find the

minimum of the function. Since the Lasso term ($\lambda_1 \|\beta\|_1$) is not differentiable in 0, we need to use the subgradient technique:

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

Where $f(\beta) = 0$, if $\beta > 0$, -1, if $\beta < 0$, and $[-1, 1]$ if $\beta = 0$.
Therefore:

$$2x_c^T(y_c - x_c\beta) \pm \lambda_1 + 2\lambda_2\beta = 0 = 2x_c^T y_c - 2x_c^T x_c\beta \pm \lambda_1 + 2\lambda_2\beta$$

Knowing that $x_c^T x_c = I$, $\beta^*_{OLS} = x_c^T y_c$. Therefore, we can obtain:

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

Than, we have:

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