IMA205: Introduction Supervised Learning

Benjamin TERNOT

March 19th 2023

Theoretical questions

OLS

We start by considering the regular OLS estimator $\beta^* = HY$, where $H = (X^T X)^{-1} X^T$ is the projection matrix onto the column space of X. We assume that $\ker(X) = 0$ (i.e., X has full column rank), which implies that H is well-defined. We also assume that the errors ϵ_i are independent and normally distributed with mean 0 and variance σ^2.

Now suppose we have another linear unbiased estimator $\tilde{\beta} = CY$ for β. Since $\tilde{\beta}$ is unbiased, we have $\mathbb{E}[\tilde{\beta}] = \beta$. Moreover, we can write $\tilde{\beta}$ as $\tilde{\beta} = HY + DY$, where D is a matrix that satisfies $DX = 0$ and $\mathbb{E}[DY] = 0$.

We can calculate the variance of $\tilde{\beta}$ as follows :

$$
\begin{aligned}
\mathbb{V}[\tilde{\beta}] &= \mathbb{V}[CY] \\
&= C\mathbb{V}[Y]C^T \qquad \text{(since } \epsilon_i \text{ is independent and normally distributed)} \\
&= \sigma^2 CC^T \qquad \text{(since } \mathbb{V}[Y] = \sigma^2 I_n\text{)} \\
&= \sigma^2 (H + D)(H^T + D^T) \\
&= \sigma^2 (X^T X)^{-1} + \sigma^2 (DX)^T (X^T X)^{-1} + \sigma^2 X (X^T X)^{-1} D^T + \sigma^2 DD^T \\
&= \sigma^2 (X^T X)^{-1} + \sigma^2 DD^T \qquad \text{(since } DX = 0 \text{ and } X(X^T X)^{-1} = H^T\text{)} \\
&= \mathbb{V}[\beta^*] + \sigma^2 DD^T.
\end{aligned}
$$

Since DD^T is positive (D is not null), we have $\mathbb{V}[\tilde{\beta}] > \mathbb{V}[\beta^*]$, which means that β^* has the minimum variance among all linear unbiased estimators of β.

Ridge Regression

- Ridge estimator $\hat{\beta}_{\text{ridge}}$ is obtained by minimizing the following objective function :

$$\underset{\beta}{\text{argmin}} ||Y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

  We can express the ridge estimator as :

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_d)^{-1} X^T Y$$

The expected value of the ridge estimator is then :

$$
\begin{aligned}
\mathbb{E}[\hat{\beta}_{\text{ridge}}] &= \mathbb{E}[(X^TX + \lambda I_d)^{-1}X^TY] \\
&= (X^TX + \lambda I_d)^{-1}X^T\mathbb{E}[Y] \\
&= (X^TX + \lambda I_d)^{-1}X^TX\beta \\
&= (X^TX + \lambda I_d)^{-1}\lambda I_d\beta + (X^TX + \lambda I_d)^{-1}X^TX\beta - (X^TX + \lambda I_d)^{-1}X^TX\beta \\
&= \lambda(X^TX + \lambda I_d)^{-1}\beta + (X^TX + \lambda I_d)^{-1}X^TX\beta - \beta \\
&= \beta + \lambda(X^TX + \lambda I_d)^{-1}\beta - \beta \\
&= (I_d - \lambda(X^TX + \lambda I_d)^{-1})\beta
\end{aligned}
$$

With $\lambda > 0$, the model is then biased.

- The ridge estimator can be expressed in terms of the SVD decomposition of the data matrix X, which is $X = UDV^T$, where U and V are orthogonal matrices and D is a diagonal matrix.
  The ridge estimator is then given by :

$$
\begin{aligned}
\beta_{\text{ridge}} &= \left(VD^TU^TUDV^T + \lambda I_d\right)^{-1} VD^TU^TY \\
&= \left(V(D^TD + \lambda I_d)V^T\right)^{-1} VD^TU^TY \\
&= V\left(D^TD + \lambda I_d\right)^{-1} D^TU^TY
\end{aligned}
$$

  This expression avoids the need to invert a matrix when computing the ridge estimator. Instead, the inverse of the diagonal matrix $D^TD + \lambda I_d$ can be computed directly in linear complexity, which is a much simpler and computationally efficient operation.

- We know that $\mathbb{V}(\beta_{\text{OLS}}) = \sigma^2(X^TX)^{-1}$.
  For ridge regression, we have $\beta_{\text{ridge}} = (X^TX + \lambda I_d)^{-1}X^TY$. Using the properties of variance, we have :

$$
\begin{aligned}
\mathbb{V}(\beta_{\text{ridge}}) &= \mathbb{V}((X^TX + \lambda I_d)^{-1}X^TY) \\
&= (X^TX + \lambda I_d)^{-1}X^T\mathbb{V}(Y)X(X^TX + \lambda I_d)^{-1} \\
&= \sigma^2(X^TX + \lambda I_d)^{-1}X^TX(X^TX + \lambda I_d)^{-1}
\end{aligned}
$$

  Since X^TX is positive and $\lambda > 0$, we have $(X^TX + \lambda I_d) \geq X^TX$.
  This means that $(X^TX + \lambda I_d)^{-1} \leq (X^TX)^{-1}$, and thus :

$$
(X^TX + \lambda I_d)^{-1}X^TX(X^TX + \lambda I_d)^{-1} \leq (X^TX)^{-1}
$$

  Combining this with the expression for $\mathbb{V}(\beta^*_{\text{OLS}})$, we get :

$$
\mathbb{V}(\beta^*_{\text{ridge}}) \leq \sigma^2(X^TX)^{-1}
$$

And then :

$$\mathbb{V}(\beta^*_{\text{OLS}}) \geq \mathbb{V}(\beta^*_{\text{ridge}})$$

Thus, the variance of the OLS estimator is always greater than the variance of the ridge estimator, and the equality holds only when $\lambda = 0$.

- Increasing the regularization parameter λ in the ridge model corresponds to increasing the penalty applied to the magnitude of the coefficients.
  This has the effect of reducing the variance of the model, since it limits the ability of the model to fit the noise in the training data.
  On the other hand, it can increase the bias of the model, since it biases the estimates of the coefficients towards zero. Therefore, increasing λ leads to a trade-off between bias and variance. Specifically, as λ increases, the variance of the model decreases while the bias increases.
- If $X^TX = I_d$, then $\beta^*_{\text{OLS}} = X^TY$, and :

$$\begin{aligned}\beta^*_{\text{ridge}} &= ((1+\lambda)I_d)^{-1}X^TY \\ &= \frac{1}{\lambda+1}X^TY \\ &= \frac{\beta^*_{\text{OLS}}}{\lambda+1}\end{aligned}$$

Elastic Net

If $X^TX = I_d$, then we have $\beta^*_{\text{OLS}} = X^TY$.
Let's define $f(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda_2||\beta||_2 + \lambda_1||\beta||_1$, and $\beta^*_{\text{ElNet}} = \arg_\beta \min(f(\beta))$.
The first derivative of f with respect to β is :

$$\begin{cases} -\lambda_1 & \text{if } \beta < 0 \\ \lambda_1 & \text{if } \beta > 0 \end{cases}$$

The second derivative of f with respect to β is :

$$\frac{\partial^2 f}{\partial \beta^2} = 2X^TX + 2\lambda_2 > 0$$

This shows that $f(\beta)$ is convex and has a minimum where its gradient is null. Setting the gradient to zero, we get :

$$\begin{aligned}0 &= -2X^T(Y - X\beta^*_{\text{OLS}}) + 2\lambda_2\beta^*_{\text{OLS}} \pm \lambda_1 \\ &= -2\beta^*_{\text{OLS}} + 2(1+\lambda_2)\beta^*_{\text{OLS}} \pm \lambda_1 \\ &= 2\beta^*_{\text{OLS}} \pm \lambda_1 - 2(1+\lambda_2)\beta^*_{\text{OLS}} \\ &= \frac{\beta^*_{\text{OLS}} \pm \frac{\lambda_1}{2}}{(1+\lambda_2)}\end{aligned}$$

And so

$$\beta^*_{\text{ElNet}} = \frac{\beta^*_{\text{OLS}} \pm \frac{\lambda_1}{2}}{(1+\lambda_2)}$$