Theoretical Supervised Learning Questions IMA205

Giovanni Benedetti da Rosa

March 11, 2024

1 OLS

The OLS estimator can be defined as

$$\beta^* = (X^T X)^{-1} X^T Y = HY. \tag{1}$$

Another linear unbiased estimator of β is defined as

$$\tilde{\beta} = CY, \tag{2}$$

where C is a $d \times n$ matrix and $C = H + D$, D being a non-zero matrix.

Let $Y = \beta X + \epsilon$, where X is deterministic and the error ϵ follows $E[\epsilon] = 0$ with $Var(\epsilon) = \sigma I^2$.

Calculating expected value and variance of $\tilde{\beta}$:

- $E[\tilde{\beta}] = E[CY] = CE[Y] = (H + D)E[Y] = (H + D)\beta X = (I + DX)\beta$. This estimator is unbiased iff $DX = 0$.

- $Var(\tilde{\beta}) = Var(CY) = CVar(Y)C^T = \sigma^2 CC^T = \sigma^2(H^T H + H^T D + D^T H + D^T D)$

By the previous answer: $Var(\tilde{\beta}) = \sigma^2((X^T X)^{-1} + D^T D) = Var(\beta^*) + \sigma^2 D^T D$.

So, as $\sigma^2 \|x\|_2^2 > 0$:

$$Var(\tilde{\beta}) > Var(\beta^*), \forall D \neq 0$$

2 Ridge Regression

• Show that the estimator of ridge regression is biased:

To minimize the function, as it a strictly convex function, let's compute the gradient and evaluate the critical points, setting it to zero.

$$\nabla J(\beta) = -2X_c^T(y_c - X_c\beta) + 2\lambda\beta$$
$$-2X_c^T(y_c - X_c\beta) + 2\lambda\beta = 0$$
$$-2X_c^T y_c + 2X_c^T X_c\beta + 2\lambda\beta = 0$$
$$X_c^T X_c\beta + \lambda\beta = X_c^T y_c$$
$$(X_c^T X_c + \lambda I)\beta = X_c^T y_c$$
$$\beta_{\text{ridge}} = (X_c^T X_c + \lambda I)^{-1} X_c^T y_c$$

Now, let's compute the expectation:

$$\begin{aligned}
E[\beta_{\text{ridge}}] &= E[(X_c^T X_c + \lambda I)^{-1} X_c^T y_c] \\
&= (X_c^T X_c + \lambda I)^{-1} E[X_c^T y_c] \\
&= (X_c^T X_c + \lambda I)^{-1} (X_c^T X_c\beta) \\
&= (X_c^T X_c + \lambda I)^{-1} X_c^T X_c\beta
\end{aligned}$$

Finally, using the definition of bias $b(\beta_{\text{ridge}}) = E[\beta_{\text{ridge}}] - \beta$, which is zero if and only if $\lambda = 0$(OLS). Thus, it's a biased model for $\lambda \neq 0$

• Recall that the SVD decomposition is $X_c = UDV^T$. Write down by hand the solution ridge using the SVD decomposition. When is it useful using this decomposition ?

$$\begin{aligned}
\beta_{\text{ridge}} &= ((VDU^T)(UDV^T) + \lambda I)^{-1}(VDU^T)y_c \\
&= ((VD^2V^T) + \lambda I)^{-1}(VDU^T)y_c \\
&= ((VD^2) + \lambda I)^{-1}(V^TVDU^T)y_c \\
&= ((VD^2) + \lambda I)^{-1}(DU^T)y_c \\
&= ((VD^2) + \lambda I)^{-1}(DU^T)UDV^T\beta \\
&= V(D^2 + \lambda I)^{-1}D^2V^T\beta
\end{aligned}$$

In this formulation, we need to invert the diagonal matrix $(D^2+\lambda I)^{-1}$ instead of $(X_c^T X_c+\lambda I)^{-1}$, that can be ill-conditioned or really large, improving then the computational efficiency.

• Show that $Var(\beta^*_{\text{OLS}}) \geq Var(\beta^*_{\text{Ridge}})$:

Using Covariance properties:

$$\begin{aligned}
\text{Given: } \hat{\beta}_{\text{ridge}} &= (X_c^T X_c + \lambda I)^{-1} X_c^T y_c \\
\text{Variance: } \text{Var}(\hat{\beta}_{\text{ridge}}) &= \text{Var}((X_c^T X_c + \lambda I)^{-1} X_c^T y_c) \\
&= (X_c^T X_c + \lambda I)^{-1} \text{Var}(X_c^T y_c)(X_c^T X_c + \lambda I)^{-1} \\
&= (X_c^T X_c + \lambda I)^{-1} \sigma^2 (X_c^T X_c)(X_c^T X_c + \lambda I)^{-1}
\end{aligned}$$

It's easy to see that for all positive λ, the OLS term will be bigger than the ridge, and if $\lambda = 0$, the terms are equal.

• When λ increases what happens to the bias and to the variance ? To evaluate this cases, let's recall the SVD decomposition for the Ridge estimator.

First, for the Variance:

$$\begin{aligned}
\text{Var}(\hat{\beta}_{\text{ridge}}) &= V((D^2 + \lambda I)^{-1} D^2 V^T \beta) \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{d_i^2}{(d_i^2 + \lambda)^2} v_i (v_i^T \beta) \right)
\end{aligned}$$

And for the bias:

$$\sum_{i=1}^{n} \left(\frac{d_i^2}{(d_i^2 + \lambda)^2} v_i (v_i^T \beta) \right) - \beta$$

For inference, if λ increases the absolute value of the bias becomes bigger, while the variance becomes smaller. In the limit case, If $\lambda \to \infty$, $Var(\hat{\beta}_{\text{ridge}}) \to 0$.

• Show that $\hat{\beta}_{\text{ridge}} = \hat{\beta}_{\text{OLS}}(1 + \lambda)$ when $x_c^T x_c = I_d$:

$\beta_{\text{ridge}} = (I + \lambda I)^{-1} X_c^T y_c$ and $\beta_{\text{OLS}} = X_c^T y_c$, So: $\beta_{\text{ridge}} \frac{\beta_{\text{OLS}}}{1+\lambda}$.

3 Elastic Net

Equation 2 is a strictly convex function, we can use Fermat's rule to find its minima:

$$\partial f = 2X_c^T\left(Y_c - X_c\beta\right) + 2\lambda_2\beta + \lambda_1 \begin{cases} \{-1\} & ,\beta < 0 \\ \{1\} & ,\beta > 0 \\ [-1, 1] & ,\beta = 0 \end{cases}$$

$$-2X_c^T\left(Y_c - X_c\beta\right) + 2\lambda_2\beta \pm \lambda_1 = 0$$
$$-2X_c^T Y_c + 2X_c^T X_c\beta + 2\lambda_2\beta \pm \lambda_1 = 0$$

If $X_c^T X_c = I$, we have that $\beta_{\text{OLS}} = X_c^T y_c$, and we have:

$$\beta_{\text{OLS}} = -2\beta(1+\lambda_2) \pm \lambda_1 \implies \beta_{\text{El.NET}} = \frac{\beta_{\text{OLS}} \pm \frac{\lambda_1}{2}}{(1+\lambda_2)}$$