Machine Learning

Ridge Regression

Irving Gómez Méndez

August-December, 2021

Correlated Features

The interpretation of the linear regression model depends on the assumption that the features are not strongly correlated. When this assumption is violated we might face numerical and statistical problems. A numerical problem is when the matrix $\mathbf{X}^T\mathbf{X}$ is singular or nearly singular and the statistical problem arises from getting parameters' estimators with a lot of uncertainty, making them unreliable.

Ridge Estimator

One way to solve the presence of colinearity is to detect a set of correlated variables and eliminate them from the analysis. If we want t o keep all the features, one option is ridge regression. Assume the usual model

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \quad \text{with } \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$$

The ridge estimator is obtained solving

$$(\mathbf{X}^T\mathbf{X} + \lambda I)\beta = \mathbf{X}^T\mathbf{Y}$$

where $\lambda \geq 0$ is the ridge parameter. Thus

$$\hat{\beta}_R = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{Y}$$

Equivalent Expressions

Since $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, then $\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})\hat{\beta}$ and the ridge estimator might be written as

$$\hat{\beta}_R = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^T\mathbf{X})\hat{\beta} \equiv \mathbf{Z}_\lambda\hat{\beta}$$

Note that

$$\begin{aligned}\mathbf{Z}_\lambda(I + \lambda(\mathbf{X}^T\mathbf{X})^{-1}) &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^T\mathbf{X})(I + \lambda(\mathbf{X}^T\mathbf{X})^{-1}) \\ &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^T\mathbf{X} + \lambda I) \\ &= I,\end{aligned}$$

hence

$$\mathbf{Z}_\lambda = \left(I + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1}$$

On the other hand, note that

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{X} + \lambda I)\left(I - \lambda(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\right) \\
&= \mathbf{X}^T\mathbf{X} + \lambda I - \lambda(\mathbf{X}^T\mathbf{X} + \lambda I)(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \\
&= \mathbf{X}^T\mathbf{X} + \lambda I - \lambda I \\
&= \mathbf{X}^T\mathbf{X},
\end{aligned}$$

while

$$\begin{aligned}
(\mathbf{X}^T\mathbf{X} + \lambda I)\mathbf{Z}_\lambda &= (\mathbf{X}^T\mathbf{X} + \lambda I)(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}(\mathbf{X}^T\mathbf{X}) \\
&= \mathbf{X}^T\mathbf{X},
\end{aligned}$$

hence

$$\mathbf{Z}_\lambda = \left(I - \lambda(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\right)$$

- $\hat{\beta}_R = \mathbf{Z}_\lambda \hat{\beta}$.
- $\mathbf{Z}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{X}$.
- $\mathbf{Z}_\lambda = \left(I + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1} \Rightarrow$ if $\lambda = 0, \mathbf{Z}_\lambda = I$ and $\hat{\beta}_R = \hat{\beta}$.
- $\mathbf{Z}_\lambda = \left(I - \lambda(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\right) \Rightarrow \lim_{\lambda\to\infty} \mathbf{Z}_\lambda = \mathbf{0}$ and $\lim_{\lambda\to\infty} \hat{\beta}_R = \mathbf{0}$.

Estimation of the Response

Let $\mathbf{X} = UDV^T$ the SVD decomposition of $\mathbf{X}$, $D = \text{diag}(d_1, \ldots, d_p)$ and assume $d_1 \geq \cdots \geq d_p$. Consider the prediction of the response vector

$$
\begin{aligned}
\hat{\mathbf{Y}}_R &= \mathbf{X}\hat{\beta}_R \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{Y} \\
&= UDV^T \left(VDU^T \cancelto{I}{UDV^T} + \lambda VV^T \right)^{-1} VDU^T\mathbf{Y} \\
&= UDV^T(V^T)^{-1}\left(D^2 + \lambda I\right)^{-1} V^{-1}VDU^T\mathbf{Y} \\
&= UD(D^2 + \lambda I)^{-1}DU^T\mathbf{Y} \\
&\equiv S\mathbf{Y}
\end{aligned}
$$

Degrees of Freedom

The trace of S is called the effective number of parameters or the degrees of freedom of the estimator.

$$\begin{aligned} tr(S) &= tr(UD(D^2+\lambda I)^{-1}DU^T) \\ &= tr(D(D^2+\lambda I)^{-1}D) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda} \\ &\equiv df(\hat{\beta}_R). \end{aligned}$$

Note that, if $\lambda = 0$, $df(\hat{\beta}_R) = p$.

Trace of ridge

The trace of ridge is a graph of $\hat{\beta}_R$ against $df(\hat{\beta}_R)$ or λ.

Bias of the ridge estimator

Since $\hat{\beta}_R = \mathbf{Z}_\lambda \hat{\beta}$, then $\mathbb{E}[\hat{\beta}_R] = \mathbf{Z}_\lambda \beta$ and

$$\begin{aligned} \text{bias}(\hat{\beta}_R) &= (\mathbf{Z}_\lambda - I)\beta \\ &= -\lambda(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\beta, \end{aligned}$$

thus

$$\text{bias}^2(\hat{\beta}_R) = \beta^T \lambda^2 (\mathbf{X}^T\mathbf{X} + \lambda I)^{-2}\beta$$

Note that if $\lambda = 0$, $\text{bias}^2(\hat{\beta}_R) = 0$
and $\lim_{\lambda\to\infty} \text{bias}^2(\hat{\beta}_R) = \beta^T\beta = \|\beta\|_2^2$.

Variance of the ridge estimator

On the other, using the expression

$$\hat{\beta}_R = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{Y},$$

we obtain

$$\begin{aligned}\mathbb{V}(\hat{\beta}_R) &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbb{V}(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} \\ &= \sigma^2\mathbf{Z}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{Z}_\lambda^T,\end{aligned}$$

so

$$\begin{aligned}tr(\mathbb{V}(\hat{\beta}_R)) &= \sigma^2 tr\left((\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-2}\right) \\ &= \sigma^2\sum_{j=1}^{p}\frac{d_j^2}{(d_j^2 + \lambda)^2}\end{aligned}$$

Note that if $\lambda = 0$, $tr(\mathbb{V}(\hat{\beta}_R)) = \sigma^2\sum_{j=1}^{p}\frac{1}{d_j^2}$ and

$\lim_{\lambda\to\infty} tr(\mathbb{V}(\hat{\beta}_R)) = 0$.

MSE of the ridge estimator

Let y be a random vector with mean μ and variance Σ, then

$$\mathbb{E}[y^T A y] = \mu^T A \mu + tr(A\Sigma).$$

Using this formula, we can calculate the MSE of the ridge estimator as

$$\begin{aligned} MSE(\hat{\beta}_R) &= \mathbb{E}\left[\left(\hat{\beta}_R - \beta\right)^T \left(\hat{\beta}_R - \beta\right)\right] \\ &= tr\left(\mathbb{V}\left(\hat{\beta}_R - \beta\right)\right) + \mathbb{E}\left[\left(\hat{\beta}_R - \beta\right)\right]^T \mathbb{E}\left[\left(\hat{\beta}_R - \beta\right)\right] \\ &= \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2} + \beta^T \lambda^2 (\mathbf{X}^T\mathbf{X} + \lambda I)^{-2} \beta \end{aligned}$$

It has been proved[1] that, if $\beta^T\beta$ is bounded, exists $\lambda > 0$ such that

$$MSE(\hat{\beta}_R) < MSE(\hat{\beta}),$$

later, it was proved[2] that

$$MSE(\hat{\beta}_R) < MSE(\hat{\beta})$$

for any $\lambda \leq 2\frac{\sigma^2}{\|\beta\|_2^2}$.
In practice, however, it is common to tune up λ through cross validation.

---

[1] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[2] Chris M Theobald. "Generalizations of mean square error applied to ridge regression". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.1 (1974), pp. 103–106.