Ridge regression Part-3

How to apply Gradient Descent in Ridge regression:

As we know that

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

we can write in vector form →

$$L = (XW - Y)^T (XW - Y) + \lambda \|W\|^2$$
$$= (XW - Y)^T (XW - Y) + \lambda W^T W$$

where,

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad W = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \rightarrow (n+1)$$

as we know that in GD, we take update for each weights.

$$w_0 = w_0 - \eta \frac{\partial L}{\partial w_0}, \quad w_1 = w_1 - \eta \frac{\partial L}{\partial w_1} \; \cdots\cdots \; w_n = w_n - \eta \frac{\partial L}{\partial w_n}.$$

⇓ collectively in vector form.

$$W_{new} = W_{old} - \eta \boxed{\frac{\Delta L}{\Delta W}} \rightarrow \text{gradient} \rightarrow \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

adding for convenience

$$L = \frac{1}{2}(XW - Y)^T (XW - Y) + \frac{1}{2} \lambda W^T W$$
$$= \frac{1}{2}(W^T X^T - Y^T)(XW - Y) + \frac{1}{2} \lambda W^T W$$
$$= \frac{1}{2}\Big[W^T X^T X W \underbrace{- W^T X^T Y - Y^T W X}_{\text{equal term}} + Y^T Y\Big] + \frac{1}{2} \lambda W^T W$$
$$= \frac{1}{2}\left[W^T X^T X W - 2 Y^T W X + Y^T Y\right] + \frac{1}{2} \lambda W^T W$$

$$\frac{\partial L}{\partial W} = \frac{1}{\cancel{2}}\left[\cancel{2} X^T X W - \cancel{2} Y^T X\right] + \frac{1}{\cancel{2}} \cancel{2} \lambda W$$

$$\frac{\partial L}{\partial W} = X^T X W - Y^T X + \lambda W$$

$$W = \begin{array}{c} \quad w_0 \;\; w_1 \qquad\quad w_n \\ [0, 1, \cdots\cdots !] \end{array}$$

for epochs

$$\boxed{W = W - \eta \frac{\partial L}{\partial W}}$$