Gradient Descent for Ridge Regression

$$L = (Xw - y)^T (Xw - y) + \lambda w^T w$$

$$L = \frac{1}{2}[w^T X^T - y^T](Xw - y)^2 + \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}\left[w^T X^T Xw - w^T X^T y - y^T Xw + y^T y \right]$$

$$+ \frac{1}{2}\lambda w^T w$$

$$= \frac{1}{2}\left[w^T X^T Xw - 2 y^T Xw + y^T y \right] + \frac{1}{2}\lambda w^T w$$

$$\frac{\partial L}{\partial w} = \frac{1}{2}\left[2X^T Xw - 2X^T y \right] + \frac{1}{2} 2\lambda w$$

$$\boxed{\frac{\partial L}{\partial w} = X^T Xw - X^T y + \lambda w}$$

$$w_{old} = w$$

$$\boxed{w_{new} = w_{old} - \eta \frac{\partial L}{\partial w}}$$

key points.

1) when we will increase $\lambda \uparrow$ all the coef. will start to shrink but never be zero

2) Coef having larger value will shrink fast as we increase the λ as compared to Coef having lower value.

3) Bias Variance trade-off

$\lambda \downarrow$

Bias ↓ overfit Variance ↑

$\lambda \uparrow$

Bias ↑ underfiting Variance ↓.

4) Impact on loss function.

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

As we increase the λ Loss function moves towards the origin

5) why Called Ridge.