Ridge Regression (L2)

Saturday 13 April 2024 7:03 PM

2D-data

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

$$= \sum_{i=1}^{n} (y_i - m x_i - b)^2 + \lambda m^2$$

$\frac{\partial L}{\partial b}$

$$L = \sum_{i=1}^{n} (y_i - m x_i - b)^2 + \lambda m^2$$

$$\frac{\partial L}{\partial b} = -2 \sum_{i=1}^{n} (y_i - m x_i - b) = 0$$

$$\sum_{i=1}^{n} (y_i - m x_i - b) = 0$$

$$\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0$$

$$\bar{y} - m\bar{x} - b = 0$$

$$-b = -\bar{y} + m\bar{x}$$

$$\boxed{b = \bar{y} - m\bar{x}}$$

$\frac{\partial L}{\partial m}$

$$L = \sum_{i=1}^{n} (y_i - m x_i - b)^2 + \lambda (m^2)$$

$$= \sum_{i=1}^{n} (y_i - m x_i - \bar{y} + m\bar{x})^2 + \lambda (m^2)$$

$$= \sum_{i=1}^{n} \left[(y_i - \bar{y}) + m(\bar{x} - x_i) \right] + \lambda m^2$$

$$\frac{\partial L}{\partial m} = 2 \sum_{i=1}^{n} (\bar{x} - x_i) \left[(y_i - \bar{y}) + m(\bar{x} - x_i) \right] + 2\lambda m = 0$$

$$2 \left[\sum_{i=1}^{n} \left[(\bar{x} - x_i)(y_i - \bar{y}) + m(\bar{x} - x_i)^2 \right] + \lambda m \right] = 0$$

$$\sum_{i=1}^{n} (\bar{x} - x_i)(y_i - \bar{y}) + m \sum_{i=1}^{n} (\bar{x} - x_i)^2 + \lambda m = 0$$

$$m \left[\sum_{i=1}^{n} (\bar{x} - x_i)^2 + \lambda \right] = -\sum_{i=1}^{n} (\bar{x} - x_i)(y_i - \bar{y})$$

$$\boxed{m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (\bar{x} - x_i)^2 + \lambda}}$$

SLR vs RIDGE

$$\boxed{m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} > \boxed{m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (\bar{x} - x_i)^2 + \lambda}}$$

$\lambda \rightarrow$ hyperparameter $[0, \infty)$

$\boxed{\uparrow \lambda \downarrow m}$

$$\boxed{b = \bar{y} - m\bar{x}} < \boxed{b = \bar{y} - m\bar{x}}$$

nD-data

x_1	x_2	x_3 ... x_m	y
x_{11}	x_{12}	x_{13} ... x_{1m}	y_1
x_{21}	x_{22}	x_{23} ... x_{2m}	y_2
x_{n1}	x_{n2}	x_{n3} ... x_{nm}	y_n

$$\boxed{\hat{y}_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm}}$$

$$\hat{y} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}_{n \times (m+1)}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}_{(m+1) \times 1}}_{\beta} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

① $\boxed{\hat{y} = X\beta}$

② $\boxed{E = (y - \hat{y})^T (y - \hat{y}) + \lambda \|\beta\|^2}$ $\rightarrow \lambda (\beta_0^2 + \beta_1^2 + \dots + \beta_n^2)$

$$= (y^T - \hat{y}^T)(y - \hat{y}) + \lambda \beta^T \beta$$

$$= y^T y - \underbrace{y^T \hat{y} - \hat{y}^T y} + \hat{y}^T \hat{y} + \lambda \beta^T \beta$$

$\rightarrow (\hat{y}^T y)^T = y^T (\hat{y}^T)^T = y^T \hat{y}$

let prove that $\hat{y}^T y$ is a symmetrix matrix

$\underbrace{\hat{y}^T_{1 \times n} y_{n \times 1}} = [\ ]_{1 \times 1} \rightarrow$ Scalar $\therefore \hat{y}^T y$ is a symmetrix matrix

③ $\boxed{E(\beta) = y^T y - 2\hat{y}^T y + \hat{y}^T \hat{y} + \lambda \beta^T \beta}$

$\hookrightarrow$ find such value for β matrix for which $E(\beta)$ is minimum

$$\boxed{\frac{dE}{d\beta} = 0}$$

$$E = y^T y - 2\hat{y}^T y + \hat{y}^T \hat{y} + \lambda \beta^T \beta \qquad (\hat{y} = X\beta)$$

$$= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta + \lambda \beta^T \beta$$

$\rightarrow$ let $X^T X = A$, given A is symmetric

$(X^T X)^T = X^T (X^T)^T = X^T X$

$$\frac{dE}{d\beta} = -\cancel{2} X^T y + \cancel{2} A\beta + \cancel{2} \lambda \beta = 0$$

$$-X^T y + A\beta + \lambda \beta = 0$$

$$A\beta + \lambda \beta = X^T y$$

$$\beta^T A + \beta^T \lambda = y^T X$$

$$\beta^T (A + \lambda) = y^T X$$

$$\beta^T (A + \lambda)(A + \lambda)^{-1} = y^T X (A + \lambda)^{-1}$$

$$\beta^T = y^T X (A + \lambda)^{-1}$$

$$\beta = [y^T X (A + \lambda)^{-1}]^T$$

$$\beta = [\underbrace{(A + \lambda)^{-1}}]^T X^T y$$

$\rightarrow$ need to prove $(A + \lambda)^{-1}$ is symmetric matrix i.e. $[(A + \lambda)^{-1}]^T = (A + \lambda)^{-1}$

$$= (A + \lambda)^{-1} X^T y$$

we know A is a symmetric matrix so $A + \lambda$ is also a symmetrix matrix

$$\boxed{\beta = (X^T X + \lambda I)^{-1} X^T y}$$

$I \rightarrow (m+1) \times (m+1)$

from sklearn.linear_model import Ridge

reg = Ridge(alpha, solver='cholesky')

$$\begin{bmatrix} 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & 1 & \cdots \end{bmatrix}_{(m+1) \times (m+1)}$$

let $A + \lambda = C$

$$CC^{-1} = I$$

$$[C^{-1}]^T C = I$$

$$[C^{-1}]^T C C^{-1} = I C^{-1}$$

$$[C^{-1}]^T I = I C^{-1}$$

$$[C^{-1}]^T = C^{-1} \qquad \therefore [(A + \lambda)^{-1}]^T = (A + \lambda)^{-1}$$

With Gradient Descent

$$\beta_m = \beta_{m\,old} - \eta \frac{\partial L}{\partial \beta_m}$$

Assume only one feature, with two records.

x_1	y
x_{11}	y_1
x_{21}	y_2

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \beta_1^2$$

$$= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \lambda \beta_1^2$$

$$= (y_1 - \beta_0 - \beta_1 x_{11})^2 + (y_2 - \beta_0 - \beta_1 x_{21})^2 + \lambda \beta_1^2$$

How?

$$\boxed{\frac{dL}{dw} = X^T X w - X^T y + \lambda w}$$

$$\frac{\partial L}{\partial \beta_0} = (-2)(y_1 - \hat{y}_1) + (-2)(y_2 - \hat{y}_2)$$

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

$$\frac{\partial L}{\partial \beta_1} = -2x_{11}(y_1 - \hat{y}_1) - 2x_{21}(y_2 - \hat{y}_2) + 2\lambda \beta_1$$

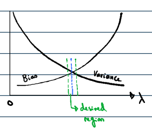
$$= -2 \sum_{i=1}^{n} x_{i1} (\hat{y}_i - \hat{y}_i) + 2\lambda \beta_1$$

Points to Remember

1. as $\lambda \uparrow$, coefficients
   - approaches 0
   - shrink
   - very close to 0 But Never 0 $\hookrightarrow$ very important
   - converge toward 0

2. Higher coefficient values are impacted more in comparison to lower coefficients values.

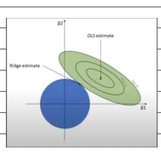
3. Bias Variance Tradeoff: When we apply regularization we aim to $\uparrow$ bias $\downarrow$ variance i.e. generalize our model. This is subjective to the hyperparameter λ.



class `sklearn.linear_model.Ridge`(*alpha=1.0, *, fit_intercept=True, copy_X=True, max_iter=None, tol=0.0001, solver='auto', positive=False, random_state=None*) [source]

4. Effect on Loss function:

   $\rightarrow$ As $\lambda \uparrow$
   - loss function tends to shift towards the origin
   - loss function is translated upwards
   - it also shrinks



5. Why is it called Ridge — [study Hard constraint Ridge Regression]

6. Apply Ridge when coef >= 2 & when we don't wan't to remove any features

*7. Can deal with multicolinearity

1. Can deal with multicollinearity