Two interpretations of Linear Regression

Liang Liang

A probabilistic interpretation

- $\hat{y} = w^T x$, a linear model
- $y = \hat{y} + \varepsilon$
- ε is random noise (something the model can not explain)
- Assume ε follows a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, then the PDF is

$$p(y_n|x_n) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\left(y_n - w^T x_n\right)^2}{2\sigma^2}\right)$$

- Assume i.i.d., the negative log likelihood (NLL) loss is

$$NLL(w) = -\frac{1}{N} log\left(\prod_{n=1}^{N} p(y_n|x_n)\right)$$
$$= -log\frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2}\left[\frac{1}{N}\sum_{n=1}^{N}\left(y_n - w^T x_n\right)^2\right]$$

- Thus, under the assumption of Gaussian noise and i.i.d., MSE = NLL

A geometric interpretation

- Define $\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N]^T$, a vector of predicted target values

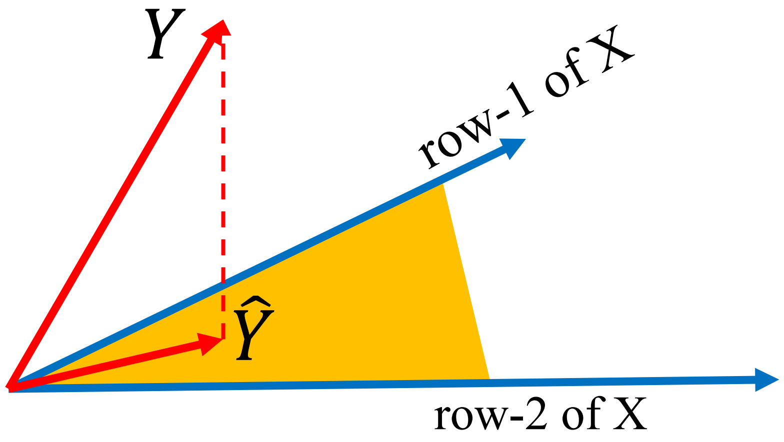
  Then, $\hat{Y} = X^T w$ because $\hat{y}_n = w^T x_n$, where $w = \left(XX^T\right)^{-1} XY$

- Define residual (a vector) to be $\hat{Y} - Y$

  $$\hat{Y} - Y = X^T\left(XX^T\right)^{-1} XY - Y = \left(X^T\left(XX^T\right)^{-1} X - I\right) Y$$

  $$X(\hat{Y} - Y) = X\left(X^T\left(XX^T\right)^{-1} X - I\right) Y = 0$$

- $\hat{Y}$ is the orthogonal projection of Y onto the space spanned by the rows of X



$$X = [x_1, x_2, \dots, x_N]$$
$$Y = [y_1, y_2, \dots, y_N]^T$$
$$\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N]^T$$