Simple regression

Important equations

hypothesis = $h_\theta(x) = \theta_0 + \theta_1 x$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

gradient descent update = $\theta_j := \theta_j - \alpha \frac{dJ}{d\theta_j}$

$$\frac{dJ}{\partial\theta_0} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

$$\frac{\partial J}{\partial\theta_1} = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x^{(i)}$$

multiple linear regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \; ---- \; \theta_n x_n$$

Vector form $h_\theta(x) = X\theta$

Performance Metrics

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$\hat{y}_i \Rightarrow$ real value of y from data
$\bar{y} \rightarrow$ mean of real dat

$$R^2_{adj} = 1 - (1 - R^2)\frac{N-1}{N-P-1}$$

N = number of samples
p = number of featwres

Error metrics

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

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Overfitting & underfitting

overfitting → Low bais, high variance
undefitting → high bais, low variance

Data split → 70% train 30% test

Simple regression (OLS)

$$y = X\beta + \epsilon \text{ error}$$

$$J(\beta) = (y - X\beta)^T (y - X\beta)$$

$$\frac{\partial J}{d\beta} = -2X^T y + 2X^T X \beta$$

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

Simple linear

$$\beta_1 = \frac{\Sigma(y_i - \bar{y})}{\Sigma(x_i - \bar{x})}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$