LINEAR REGRESSION

SIMPLE LINEAR REGRESSION:

used for continous data

Given some data, can you pass a line through the data that you can then use to predict unseen data.

what defines a line?

intercept and slope.

prediction

$$\hat{y} = \theta_0 + \theta_1 x \qquad [y = c + mx]$$

How do we find the slope and intercept that give us the best predictions?

Assume we're working under L_2 loss
$[L_2 = (y_i - \hat{y}_i)^2]$

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

We can find the minimum θ_0 and θ_1 by setting the gradient of the objective function to 0.

We need to test for second derivative > 0 but we'll ignore that for now.

$$\frac{\partial}{\partial \theta_0}(\text{MSE}) = \frac{\partial}{\partial \theta}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \theta}(y_i - \theta_0 - \theta_1 x_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \theta_0 - \theta_1 x_i)(-1)$$

$$= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)$$

$$\frac{\partial (\text{MSE})}{\partial \theta_0} = 0 \Rightarrow -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i) = 0$$

$$\Downarrow$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

$e = y_i - \hat{y}_i$

sum of residuals for a linear model with intercept is 0

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\theta_0 - \sum_{i=1}^{n}\theta_1 x_i = 0$$

$$\sum_{i=1}^{n} \theta_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \theta_1 x_i$$

$$n\theta_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \theta_1 x_i$$

$$\theta_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} \theta_1 x_i\right)$$

→ optimum

$$\hat{\theta}_0 = \bar{y} - \frac{1}{n}\left(\sum_{i=1}^{n} \theta_1 x_i\right)$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \qquad \text{(1)}$$

Similarly for θ_1,

$$\frac{\partial(\text{MSE})}{\partial\theta_1} = \frac{\partial}{\partial\theta_1}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_1}(y_i - \theta_0 - \theta_1 x_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)(2)(-x_i)$$

$$= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)(x_i)$$

$$\frac{\partial(\text{MSE})}{\partial\theta_1} = 0 \Rightarrow -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)(x_i) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i) = 0$$

And with a bit of algebra we get the following:

$$\hat{\theta}_1 = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)\right)\frac{\sqrt{(1/n)\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\boxed{\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}\ ; \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

CONSTANT MODEL:

$$\hat{y} = \theta_0$$

↳ so, what is the optimal θ_0?

→ Assume our objective function is MSE:

↓ Following a similar approach to above:

$$\frac{\partial(MSE)}{\partial\theta_0} = \frac{\partial}{\partial\theta}\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(2)(y_i - \theta_0)(-1)$$

$$= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0)$$

Setting derivative to 0:

$\hat{\theta}_0 = \bar{y}_i$ → more sensitive to outliers

→ unique solution

mean absolute error.

→ Now, if we do the same thing with MAE:

$$\frac{\partial(MAE)}{\partial\theta} = \frac{\partial}{\partial\theta}\left[\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\underbrace{\frac{\partial}{\partial\theta}|y_i - \theta_0|}$$

$$\frac{\partial|y_i - \theta_0|}{\partial\theta} = \begin{cases} 1 & \text{when } \theta_0 > y_i \\ -1 & \text{when } y_i > \theta_0 \end{cases}$$

$$= \frac{1}{n}\left[\sum_{y_i > \theta_0}(-1) \quad \sum_{\theta_0 > y_i}(1)\right]$$

Setting the derivative to 0

$$\sum_{y_i > \theta_0}(1) = \sum_{\theta_0 > y_i}(1) \implies \hat{\theta}_0 = \text{median}(y)$$

→ more robust to outliers.

↳ can be a line → uniqueness not guaranteed.

ORDINARY LEAST SQUARES:

↳ what if we have multiple features:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_p x_p$$

↳ we can formulate this in terms of matrix multiplication

$$\underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_p \\ 1 & | & | & & | \\ \vdots & | & | & & | \\ 1 & | & | & & | \end{bmatrix}}_{n \times (p+1)} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}}_{(p+1) \times 1}$$

$n \times (p+1)$ ⇓ Design Matrix

$\hat{Y} = X\theta$ ⇒ How do we get the most optimal weights & biases $\hat{\theta}$?

$$\hat{Y} = \underbrace{\theta_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \theta_1 \begin{bmatrix} x_1 \\ | \end{bmatrix} + \theta_2 \begin{bmatrix} x_2 \\ | \end{bmatrix} + \dots + \theta_p \begin{bmatrix} x_p \\ | \end{bmatrix}}$$

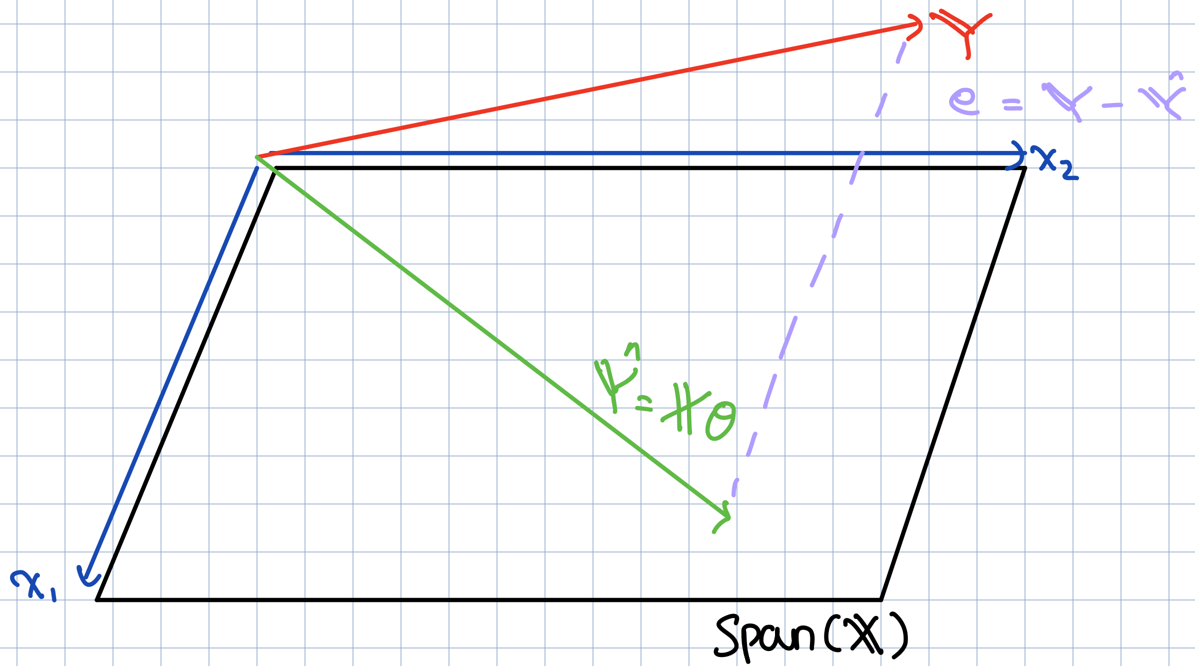
Notice that this is a linear combination of the columns of X.

↳ Predictions are in the span of X → column space.

However, our true Y can lie anywhere

↓

see diagram



→ The vector in the span(X) that is closest to Y is the orthogonal projection of Y onto span (X)

↓

We must choose e to be a vector that is perpendicular to every vector in the column space.

↓

A vector is orthogonal to the span of a matrix ⟺ it is orthogonal to every column.

$$M^T\vec{v} = \vec{0}$$

$$e = Y - X\hat{\theta}$$ → optimal θ that makes e orthogonal to the column space.

$$X^T(Y - X\hat{\theta}) = \vec{0}$$

$$X^TY - X^TX\hat{\theta} = \vec{0}$$

$$X^TX\hat{\theta} = X^TY$$

$$\boxed{\hat{\theta} = (X^TX)^{-1}X^TY}$$

Assumption:
(X^TX) is invertible
⇓
X is full rank
⇓
rank = # of linearly independent columns
rank = # of columns

Therefore, we require that $d << n$ → d: number of features; n: number of training pts.