Linear Regression-Normal Equation

Dr.Vinod P.

Name of institution

September 13, 2020

Normal Equation

$$\theta = (X^T.X)^{-1}(X^T.y) \tag{1}$$

Normal Equation

$$\theta = (X^T.X)^{-1}(X^T.y) \tag{1}$$

- θ : parameters
- X : input feature value of each instance
- Y :Output value of each instance

Hypothesis function

$$h_\theta = \theta_0 x_0 + \theta_1 x_1 + \cdot + \theta_n x_n$$

- n : number of features in the dataset.
- $x_0 = 1$ (for vector multiplication)

Hypothesis function

$$h_\theta = \theta_0 x_0 + \theta_1 x_1 + \cdot + \theta_n x_n$$

- n : number of features in the dataset.
- $x_0 = 1$ (for vector multiplication)

Dot product between θ and X can written as :

$$\boxed{h_\theta = \theta_0^T X}$$

Cost function

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

- $x^{(i)}$ = ith training example
- m : number of training instances
- n : number of features
- $y^{(i)}$ = the expected outcome for ith instance

Cost function to Vector Form

$$\begin{bmatrix} h_\theta x^{(0)} \\ h_\theta x^{(1)} \\ \vdots \\ h_\theta x^{(m)} \end{bmatrix} - \begin{bmatrix} y^{(0)} \\ y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$\because \boxed{h_\theta(x) = \theta^T X}$

$$\begin{bmatrix} \theta^T x^{(0)} \\ \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(m)} \end{bmatrix} - \begin{bmatrix} y^{(0)} \\ y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$$
\begin{bmatrix}
\theta_0 x_0^{(0)} + \theta_1 x_1^{(0)} + \cdots + \theta_n x_n^{(0)} \\
\theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \cdots + \theta_n x_n^{(1)} \\
\vdots \\
\theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + \cdots + \theta_n x_n^{(m)}
\end{bmatrix}
-
\begin{bmatrix}
y^{(0)} \\
y^{(1)} \\
\vdots \\
y^{(m)}
\end{bmatrix}
$$

$$\begin{bmatrix} \theta_0 x_0^{(0)} + \theta_1 x_1^{(0)} + \cdots + \theta_n x_n^{(0)} \\ \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \cdots + \theta_n x_n^{(1)} \\ \vdots \\ \theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + \cdots + \theta_n x_n^{(m)} \end{bmatrix} - y$$

$\boxed{\text{x}_j^i\text{: } j^{th} \text{ feature of } i^{th} \text{ training example}}$

$$\begin{bmatrix} \theta_0 x_0^{(0)} + \theta_1 x_1^{(0)} + \cdots + \theta_n x_n^{(0)} \\ \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \cdots + \theta_n x_n^{(1)} \\ \vdots \\ \theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + \cdots + \theta_n x_n^{(m)} \end{bmatrix} - y$$

$$\boxed{\text{x}_j^i\text{: } j^{th} \text{ feature of } i^{th} \text{ training example}}$$

We'll define the "design matrix" X (uppercase X) as a matrix of m rows, in which each row is the i-th sample (the vector $x^{(i)}$)

$$X\theta - y$$

But in our cost function there is a square, to get the squared values, multiply the matrix with it's transpose.

$$J(\theta) = \frac{1}{2m}\left((X\theta - y)^T(X\theta - y)\right)$$

But in our cost function there is a square, to get the squared values, multiply the matrix with it's transpose.

$$J(\theta) = \frac{1}{2m}\Big((X\theta - y)^T(X\theta - y)\Big)$$

Ignore the $\frac{1}{2m}$ since it's not going to make any difference in derivation

$$J(\theta) = \Big((X\theta)^T - y^T)(X\theta - y)\Big)$$

But in our cost function there is a square, to get the squared values, multiply the matrix with it's transpose.

$$J(\theta) = \frac{1}{2m}\Big((X\theta - y)^T(X\theta - y)\Big)$$

Ignore the $\frac{1}{2m}$ since it's not going to make any difference in derivation

$$J(\theta) = \Big((X\theta)^T - y^T)(X\theta - y)\Big)$$

J$(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y$

But in our cost function there is a square, to get the squared values, multiply the matrix with it's transpose.

$$J(\theta) = \frac{1}{2m}\Big((X\theta - y)^T(X\theta - y)\Big)$$

Ignore the $\frac{1}{2m}$ since it's not going to make any difference in derivation

$$J(\theta) = \Big((X\theta)^T - y^T)(X\theta - y)\Big)$$

$\mathrm{J}(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y$
$\mathrm{J}(\theta) = (X\theta)^T X\theta - \color{red}{(X\theta)^T y} - \color{red}{y^T(X\theta)} + y^T y$

But in our cost function there is a square, to get the squared values, multiply the matrix with it's transpose.

$$J(\theta) = \frac{1}{2m}\Big((X\theta - y)^T(X\theta - y)\Big)$$

Ignore the $\frac{1}{2m}$ since it's not going to make any difference in derivation

$$J(\theta) = \Big((X\theta)^T - y^T)(X\theta - y)\Big)$$

J$(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y$
J$(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y$
J$(\theta) = (X\theta)^T X\theta - 2(X\theta)^T y + y^T y$

Taking partial derivatives w.r.t. to θ and equating the result to 0

$$J(\theta) = (X\theta)^T X\theta - 2(X\theta)^T y + y^T y$$

$$\frac{\partial}{\partial\theta}\left((X\theta)^T X\theta - 2(X\theta)^T y\right) = 0$$

Taking partial derivatives w.r.t. to θ and equating the result to 0

$$J(\theta) = (X\theta)^T X\theta - 2(X\theta)^T y + y^T y$$

$$\frac{\partial}{\partial\theta}\left((X\theta)^T X\theta - 2(X\theta)^T y\right) = 0$$

$$\mathrm{B}\theta = 2\left[\begin{bmatrix} x_{11} + x_{12} + \cdots + x_{1n} \\ x_{21} + x_{22} + \cdots + x_{2n} \\ \vdots \\ x_{m1} + x_{m2} + \cdots + x_{mn} \end{bmatrix}\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}\right]^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$B\theta = 2\begin{bmatrix}\begin{bmatrix} x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n \\ x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n \\ \vdots \\ x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n \end{bmatrix}\end{bmatrix}^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$B\theta = 2\left[\begin{bmatrix} x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n \\ x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n \\ \vdots \\ x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n \end{bmatrix}\right]^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$= 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$$\vdots$$

$$2(x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$$

$$\frac{\partial}{\partial \theta_1} = 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$$\vdots$$

$$2(\mathrm{x}_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$$

$$= 2(\mathrm{x}_{11}y_1 + x_{21}y_2 + \cdots + x_{m1}y_m)$$

$$\frac{\partial}{\partial \theta_1} = 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$$\vdots$$

$$2(x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$$
$$= 2(x_{11}y_1 + x_{21}y_2 + \cdots + x_{m1}y_m)$$

$$\frac{\partial}{\partial \theta_2} = 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$$\vdots$$

$$2(x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$$
$$= 2(x_{12}y_1 + x_{22}y_2 + \cdots + x_{m2}y_m)$$

$$\frac{\partial}{\partial \theta_n} = 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$$\vdots$$

$$2(x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$$
$$= 2(x_{1n}y_1 + x_{2n}y_2 + \cdots + x_{mn}y_m)$$

$$\frac{\partial}{\partial \theta_n} = 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$\vdots$

$2(x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$
$= 2(x_{1n}y_1 + x_{2n}y_2 + \cdots + x_{mn}y_m)$

$$\frac{\partial B}{\partial \theta} = 2X^T.y$$

$$\frac{\partial}{\partial \theta_n} = 2(x_{11}\theta_1 + x_{12}\theta_2 + \cdots + x_{1n}\theta_n) + 2(x_{21}\theta_1 + x_{22}\theta_2 + \cdots + x_{2n}\theta_n)$$

$$\vdots$$

$$2(x_{m1}\theta_1 + x_{m2}\theta_2 + \cdots + x_{mn}\theta_n)$$
$$= 2(x_{1n}y_1 + x_{2n}y_2 + \cdots + x_{mn}y_m)$$

$$\frac{\partial B}{\partial \theta} = 2X^T.y$$

$$\frac{\partial}{\partial \theta}\left((X\theta)^T X\theta - 2(X\theta)^T y\right) = 0$$

$$A(\theta) = (X\theta)^T(X\theta) = \theta^T X^T \theta$$

$$\begin{bmatrix}\theta_0 \cdots \theta_n\end{bmatrix} \begin{bmatrix} x_{10} \cdots x_{m0} \\ \vdots \ddots \vdots \\ x_{1n} \cdots x_{mn} \end{bmatrix} \begin{bmatrix} x_{10} \cdots x_{1n} \\ \vdots \ddots \vdots \\ x_{m0} \cdots x_{mn} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$A(\theta) = (X\theta)^T(X\theta) = \theta^T X^T \theta$$

$$\begin{bmatrix}\theta_0 \cdots \theta_n\end{bmatrix} \begin{bmatrix} x_{10} \cdots x_{m0} \\ \vdots \ddots \vdots \\ x_{1n} \cdots x_{mn} \end{bmatrix} \begin{bmatrix} x_{10} \cdots x_{1n} \\ \vdots \ddots \vdots \\ x_{m0} \cdots x_{mn} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$= [(\theta_0 x_{10} + \theta_1 x_{11} + \cdots + \theta_n x_{1n}) \cdots (\theta_0 x_{m0} + \theta_1 x_{m1} + \cdots + \theta_n x_{mn})]$$

$$\begin{bmatrix} (\theta_0 x_{10} + \theta_1 x 11 + \cdots + \theta_n x_{1n}) \\ (\theta_0 x_{20} + \theta_1 x 11 + \cdots + \theta_n x_{1n}) \\ \vdots \\ (\theta_0 x_{m0} + \theta_1 x m1 + \cdots + \theta_n x_{mn}) \end{bmatrix}$$

$$A(\theta) = \begin{matrix} (\theta_0 x_{10} + \theta_1 x11 + \cdots + \theta_n x_{1n})^2 + \\ (\theta_0 x_{20} + \theta_1 x11 + \cdots + \theta_n x_{1n})^2 + \\ \vdots \\ (\theta_0 x_{m0} + \theta_1 xm1 + \cdots + \theta_n x_{mn})^2 \end{matrix}$$

$$A(\theta) = \begin{array}{c} (\theta_0 x_{10} + \theta_1 x11 + \cdots + \theta_n x_{1n})^2 + \\ (\theta_0 x_{20} + \theta_1 x11 + \cdots + \theta_n x_{1n})^2 + \\ \vdots \\ (\theta_0 x_{m0} + \theta_1 xm1 + \cdots + \theta_n x_{mn})^2 \end{array}$$

Taking partial derivative w.r.t θ

$$\frac{\partial}{\partial \theta_0} = 2x_{10}(\theta_0 x_{10} + \theta_1 x11 + \cdots + \theta_n x_{1n}) +$$
$$\vdots$$
$$+2x_{m0}(\theta_0 x_{m0} + \theta_1 xm1 + \cdots + \theta_n x_{mn})$$

$$\frac{\partial}{\partial \theta_1} = 2x_{11}(\theta_0 x_{10} + \theta_1 x11 + \cdots + \theta_n x_{1n}) +$$
$$\vdots$$
$$+2x_{m1}(\theta_0 x_{m0} + \theta_1 xm1 + \cdots + \theta_n x_{mn})$$
$$\vdots$$

$$\frac{\partial}{\partial \theta_n} = 2x_{1n}(\theta_0 x_{10} + \theta_1 x11 + \cdots + \theta_n x_{1n}) +$$
$$\vdots$$
$$+2x_{mn}(\theta_0 x_{m0} + \theta_1 xm1 + \cdots + \theta_n x_{mn})$$

$$\frac{\partial A}{\partial \theta} = 2X^T X\theta$$

$$2X^T X\theta = 2X^T y$$

$$X^T X\theta = X^T y$$

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

Thank You