DS 312: Project 2: Question 1:

Rachel Newman
09/16/24

Question 1: Theoretical Foundations of Linear Regression:

Consider the linear equation $y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$. Matrix form of above linear regression: $y = Xw$, where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

1. Loss Function: Write the Mean Squared Error (MSE) as the loss function for this linear model in matrix form.

starting with original model ($t = w_0 + w_1 x$) in vector form:

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad x_n = \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

$$f(x_n; w_0, w_1) = w^T x_n = w_0 + w_1 x_n$$

squared loss function can be expressed as

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (y_n - w^T x_n)^2$$

can express this average loss as a function of vectors and matrices

$$\mathcal{L} = \frac{1}{N} (y - Xw)^T (y - Xw)$$

performing matrix multiplication of X and w results in a vector:

$$Xw = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} + \cdots + w_n x_{1n} \\ w_0 + w_1 x_{11} + w_2 x_{22} + \cdots + w_n x_{2n} \\ \vdots \\ w_0 + w_1 x_{n1} + w_2 x_{n2} + \cdots + w_n x_{nn} \end{bmatrix}$$

Subtracting this from y gives

$$y - Xw = \begin{bmatrix} y_1 - w_0 - w_1 x_{11} - w_2 x_{12} - \cdots - w_n x_{1n} \\ y_2 - w_0 - w_1 x_{21} - w_2 x_{22} - \cdots - w_n x_{2n} \\ \vdots \\ y_n - w_0 - w_1 x_{n1} - w_2 x_{n2} - \cdots - w_n x_{nn} \end{bmatrix}$$

$$(Xw - y)^T (Xw - y) = (w_0 + w_1 x_{11} + \cdots w_n x_{11} - y_1)^2 + \cdots + (w_0 + w_1 x_{n1} + \cdots + w_n x_{nn})^2$$

$$(Xw - y)^T (Xw - y) = \sum_{n=1}^{N} (w_0 + w_1 x_n + \cdots + w_n x_n - y_n)^2$$

$$(Xw - y)^T (Xw - y) = \sum_{n}^{N} (y_n - f(x_n; w_0, w_1, \cdots w_n))^2$$

Loss function can be written as

$$\boxed{\mathcal{L} = \frac{1}{N} (y - Xw)^T (y - Xw)}$$

$$\mathcal{L} = \frac{1}{N} (Xw - y)^T (Xw - y)$$

$$\mathcal{L} = ((Xw)^T - y^T)(Xw - y)$$

$$\mathcal{L} = \frac{1}{N} (Xw)^T (Xw) - \frac{1}{N} y^T X w - \frac{1}{N} (Xw)^T y + \frac{1}{N} y^T y$$

$$\mathcal{L} = \frac{1}{N} w^T X^T X w - \frac{2}{N} w^T X^T y + \frac{1}{N} y^T y$$

$$\boxed{\mathcal{L} = \frac{1}{N} (w^T X^T X w - 2 w^T X^T y + y^T y)}$$

2. Gradient Descent: Show the gradient descent steps to minimize the loss function with respect to the weights/coefficients w.

$$\frac{\partial \mathcal{L}}{\partial w} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_0} \\ \frac{\partial \mathcal{L}}{\partial w_1} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial w_n} \end{bmatrix}$$

useful identities when differentiating with respect to a vector

$f(w)$	$\frac{\partial f}{\partial w}$
$w^T x$	x
$x^T w$	x
$w^T w$	$2w$
$w^T C w$	$2Cw$

From these identities, and equating derivatives to 0, the following is obtained

$$\mathcal{L} = \frac{1}{N} w^T X^T X w - \frac{2}{N} w^T X^T y + \frac{1}{N} y^T y$$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{2}{N} X^T X w - \frac{2}{N} X^T y + 0$$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{2}{N} X^T X w - \frac{2}{N} X^T y = 0$$

$$\frac{2}{N} X^T X w = \frac{2}{N} X^T y$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial w} = X^T X w = X^T y}$$

3. Derive Closed-Form Equations: Derive the following equation for w:

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

To obtain $\hat{w}$, an optimum value of w, the equation $\frac{\partial \mathcal{L}}{\partial w}$ must be rearranged. It is important to note that the inverse of $X^T X$ is denoted by $(X^T X)^{-1}$.

Suppose $I_{n \times n}$ is an identity matrix

from the definition of an identity matrix

$$Iw = w$$

$$(X^T X)^{-1} (X^T X) I w = (X^T X)^{-1} X^T y$$

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

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