Homework 1

Bryan Pitsker Data Mining 9/15/23

① Let $\vec{X} = [1, x_1, x_2, \cdots, x_D]^T = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$

- So we have D features for each x vector.

Let $W = [w_0, w_1, w_2, \cdots, w_D]^T = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$

- And we have D weights for each of the D independent features

$y_* = w^T \vec{x}_* = [w_0, w_1, \cdots, w_D] \cdot \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix} =$

↑ Prediction for y-output

$y_* = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D$

Now if we let

$$X = \begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vdots \\ \vec{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_D \\ 1 & x_1 & x_2 & \cdots & x_D \\ & & \vdots & & \\ 1 & x_1 & x_2 & \cdots & x_D \end{bmatrix}$$

Note

"$\vec{x}_1, \vec{x}_2, \cdots \vec{x}_N$ are all individual vectors that we initialized above in the first line"

* So, we have D features/columns in our matrix, but there are N total rows, or observances, of each feature in the matrix.

$$X_W = \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_D \\ 1 & x_1 & x_2 & \cdots & x_D \\ & & \vdots & & \\ x_1 & & x_2 & \cdots & x_D \end{bmatrix} \cdot \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}$$

$$N \cdot D \quad \cdot \quad D \cdot 1$$

$$N \cdot 1$$

$$X_W = \begin{bmatrix} w_0 + x_1 w_1 + x_2 w_2 + \cdots + x_D w_D \\ w_0 + x_1 w_1 + x_2 w_2 + \cdots + x_D w_D \\ \vdots \\ w_0 + x_1 w_1 + x_2 w_2 + \cdots + x_D w_D \end{bmatrix}$$

N rows, 1 column

- Now if the actual outputs y are:

$$y = [y_1, y_2, y_3, \cdots, y_N]^T = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Then this means that

$$X_W - y = \begin{bmatrix} (w_0 + x_1 w_1 + \cdots + x_D w_D) - y_1 \\ (w_0 + x_1 w_1 + \cdots + x_D w_D) - y_2 \\ \vdots \\ (w_0 + x_1 w_1 + \cdots + x_D w_D) - y_N \end{bmatrix}$$

Thus, the L2-norm squared is: $||Xw-y||^2$

$$\hookrightarrow ||Xw-y||^2 =$$

$$= \sqrt{\left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_1\right]^2+\left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_2\right]^2 + \cdots + \left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_N\right]^2}^2$$

The square root and the squared cancel out, so all we are left with is:

$$||Xw-y||^2 = \left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_1\right]^2 + \left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_2\right]^2 + \cdots + \left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_N\right]^2$$

- Therefore $\frac{1}{N}\cdot||Xw-y||^2 =$

$$= \frac{1}{N}\cdot\left[\left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_1\right]^2+\left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_2\right]^2 + \cdots + \left[(w_0+x_1w_1+\cdots+x_Dw_D)-y_N\right]^2\right]$$

BUT, this is the exact same thing as writing

$$\Rightarrow \frac{1}{N} \cdot \sum_{n=1}^{N} \left(w^T \vec{x}_n - y_n\right)^2 =$$

↖ reminder: X is the vector here, NOT an observance or a point

$$= \frac{1}{N} \cdot \left[\left(w^T \vec{x}_1 - y_1\right)^2 + \left(w^T \vec{x}_2 - y_2\right)^2 + \cdots + \left(w^T \vec{x}_N - y_N\right)^2 \right]$$

Note: $w^T \cdot \vec{x}_*$ we showed before to be

$$w^T \cdot \vec{x}_* = w_0 + w_1 x_1 + \cdots + w_D x_D$$

So, → the equation at the very top becomes

$$= \frac{1}{N} \cdot \left[\begin{array}{l} \left[(w_0 + w_1 x_1 + \cdots + w_D x_D) - y_1\right]^2 \\ + \left[(w_0 + w_1 x_1 + \cdots + w_D x_D) - y_2\right]^2 + \cdots \\ \cdots + \left[(w_0 + w_1 x_1 + \cdots + w_D x_D) - y_N\right]^2 \end{array} \right]$$

which is what we got before!

Hence, we proved that the Mean Squared Error:

$$E(w) = \frac{1}{N} \cdot \sum_{n=1}^{N} \left(w^T x_n - y_n\right)^2 = \frac{1}{N} \cdot \|Xw - y\|^2$$

↑ matrix X