ASSIGNMENT 6

Nayan Man Singh Pradhan

Excercise 1. (Gradient descent on paper)

Consider the following training data:

$$\mathcal{T} = \left\{ \left((4,1)^\top, 2\right), \left((2,8)^\top, -14\right), \left((1,0)^\top, 1\right), \left((3,2)^\top, -1\right) \right\}.$$

We want to compute the predictor from linear regression by least squares. However, instead of using Theorem 4.1, we use batch gradient descent to compute the coefficient vector $\hat{\beta}$.
Choose the initial guess $\beta^{(0)} = (1,1,1)^\top$ and calculate the first two steps of batch gradient descent with learning rate $\eta = 0.1$.

(4 Points)

Given,

$$\beta^0 = (1,1,1)^T$$
$$\eta = 0.1$$
$$X = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 2 & 8 \\ 1 & 1 & 0 \\ 1 & 3 & 2 \end{pmatrix}, \quad X^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix}$$
$$y = \begin{pmatrix} 2 \\ -14 \\ 1 \\ -1 \end{pmatrix}$$

We know,

$$\beta^{n+1} = \beta^n + \eta 2 X^T (y - X\beta^n)$$

Solving,

$$\beta^1 = \beta^0 + \eta 2 X^T (y - X\beta^0)$$
$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0.1 \times 2 \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix} \cdot \left[\begin{pmatrix} 2 \\ -14 \\ 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 & 4 & 1 \\ 1 & 2 & 8 \\ 1 & 1 & 0 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right]$$
$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0.2 \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix} \cdot \left[\begin{pmatrix} 2 \\ -14 \\ 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 6 \\ 11 \\ 2 \\ 6 \end{pmatrix} \right]$$
$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0.2 \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} -4 \\ -25 \\ -1 \\ -7 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0.2 \times \begin{pmatrix} -37 \\ -88 \\ -218 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} -7.4 \\ -17.6 \\ -43.6 \end{pmatrix}$$

$$= \begin{pmatrix} -6.4 \\ -16.6 \\ -42.6 \end{pmatrix}$$

Again,

$$\beta^2 = \beta^1 + \eta\, 2X^T (y - X\beta^1)$$

$$= \begin{pmatrix} -6.4 \\ -16.6 \\ -42.6 \end{pmatrix} + 0.2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix} \left[\begin{pmatrix} 2 \\ -14 \\ 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 & 4 & 1 \\ 1 & 2 & 8 \\ 1 & 1 & 0 \\ 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} -6.4 \\ -16.6 \\ -42.6 \end{pmatrix} \right]$$

$$= \begin{pmatrix} -6.4 \\ -16.6 \\ -42.6 \end{pmatrix} + 0.2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix} \left[\begin{pmatrix} 2 \\ -14 \\ 1 \\ -1 \end{pmatrix} - \begin{pmatrix} -115.4 \\ -380.4 \\ -23 \\ -141.4 \end{pmatrix} \right]$$

$$= \begin{pmatrix} -6.4 \\ -16.6 \\ -42.6 \end{pmatrix} + 0.2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix} \begin{pmatrix} 117.4 \\ 366.4 \\ 24 \\ 140.4 \end{pmatrix}$$

$$= \begin{pmatrix} -6.4 \\ -16.6 \\ -42.6 \end{pmatrix} + \begin{pmatrix} 129.64 \\ 329.52 \\ 665.88 \end{pmatrix}$$

$$= \begin{pmatrix} 123.24 \\ 312.92 \\ 623.28 \end{pmatrix}$$

Excercise 2. (Gradient descent variations for other models)

In this task, we consider again the training data

$$\mathcal{T} = \left\{ \left((4,1)^\top, 2\right), \left((2,8)^\top, -14\right), \left((1,0)^\top, 1\right), \left((3,2)^\top, -1\right) \right\} .$$

However, this time we do not use the linear model. Instead we use a quadratic model

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2$$

and stick to the L_2 loss.

a) Derive the gradient $\nabla_\beta L_2(y_i, f(\mathbf{x}_i))$ of the loss with respect to the coefficient vector.

b) Use the initial guess $\beta^{(0)} = \mathbf{0}$ and the learning rate $\eta = 0.1$ and perform the first two steps of stochastic gradient descent. Recall that in stochastic gradient descent, we at some point pick random samples from the training set. In order to get a unique solution in this task, we assume that a random selection of samples from the training set would give you the training samples in their original order.

c) Use the initial guess $\beta^{(0)} = \mathbf{0}$ and the learning rate $\eta = 0.1$ and perform the first two steps of mini-batch gradient descent with a batch size of $N_b = 2$. Use the same approach for the randomization part as in the previous sub-task.

(4+2+2 Points)

a) $(y_i - f(x_i)) = \begin{pmatrix} [2 - (\beta_0 + \beta_1 4 + \beta_2 1 + \beta_3 4 + \beta_4 16 + \beta_5 1)]^2 \\ [-14 - (\beta_0 + \beta_1 2 + \beta_2 8 + \beta_3 16 + \beta_4 4 + \beta_5 64)]^2 \\ [1 - (\beta_0 + \beta_1 1 + \beta_2 0 + \beta_3 0 + \beta_4 1 + \beta_5 \cdot 0)]^2 \\ [-1 - (\beta_0 + \beta_1 3 + \beta_2 2 + \beta_3 6 + \beta_4 9 + \beta_5 \cdot 4)]^2 \end{pmatrix}$

$L_2(y_i, f(x_i))$

$= \begin{pmatrix} 2 - \beta_0 - 4\beta_1 - \beta_2 - 4\beta_3 - 16\beta_4 - \beta_5 \\ -14 - \beta_0 - 2\beta_1 - 8\beta_2 - 16\beta_3 - 4\beta_4 - 64\beta_5 \\ 1 - \beta_0 - \beta_1 + 0 + 0 - \beta_4 + 0 \\ -1 - \beta_0 - 3\beta_1 - 2\beta_2 - 6\beta_3 - 9\beta_4 + 4\beta_5 \end{pmatrix}$

$L_2 = (y_i - f(x_i))^\top (y_i - f(x_i)) = \begin{pmatrix} (2 - \beta_0 - 4\beta_1 - \beta_2 - 4\beta_3 - 16\beta_4 - \beta_5)^2 \\ (-14 - \beta_0 - 2\beta_1 - 8\beta_2 - 16\beta_3 - 4\beta_4 - 64\beta_5)^2 \\ (1 - \beta_0 - \beta_1 - \beta_4)^2 \\ (-1 - \beta_0 - 3\beta_1 - 2\beta_2 - 6\beta_3 - 9\beta_4 + 4\beta_5)^2 \end{pmatrix}$

$$\nabla_B L_2(y_i, f(x_i)) = \begin{pmatrix} -1 & -4 & -1 & -4 & -16 & -1 \\ -1 & -2 & -8 & -16 & -4 & -64 \\ -1 & -1 & 0 & 0 & -1 & 0 \\ -1 & -3 & -2 & -6 & -9 & 4 \end{pmatrix}$$

(i couldn't solve it, but I spent quite alot of time understanding / trying!)

$$(y - X\beta) = \begin{pmatrix} 2 - (\beta_0 + \beta_1 4 + \beta_2 1 + \beta_3 4 + \beta_4 16 + \beta_5 1) \\ -14 - (\beta_0 + \beta_1 2 + \beta_2 8 + \beta_3 16 + \beta_4 4 + \beta_5 64) \\ 1 - (\beta_0 + \beta_1 1 + \beta_2 0 + \beta_3 0 + \beta_4 1 + \beta_5 \cdot 0) \\ -1 - (\beta_0 + \beta_1 3 + \beta_2 2 + \beta_3 6 + \beta_4 9 + \beta_5 \cdot 4) \end{pmatrix}$$

$$= \begin{pmatrix} 2 - \beta_0 - 4\beta_1 - \beta_2 - 4\beta_3 - 16\beta_4 - \beta_5 \\ -14 - \beta_0 - 2\beta_1 - 8\beta_2 - 16\beta_3 - 4\beta_4 - 64\beta_5 \\ 1 - \beta_0 - \beta_1 + 0 + 0 - \beta_4 + 0 \\ -1 - \beta_0 - 3\beta_1 - 2\beta_2 - 6\beta_3 - 9\beta_4 + 4\beta_5 \end{pmatrix}$$

$$\nabla_\beta L_2(y_i, f(x_i)) = 2X^T \cdot (y - X\beta)$$

$$= 2 \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 1 & 8 & 0 & 2 \end{pmatrix}$$

a) $L_2(y_i, f(x_i)) = \begin{pmatrix} [2 - (\beta_0 + \beta_1 4 + \beta_2 1 + \beta_3 4 + \beta_4 16 + \beta_5 1)] \\ [-14 - (\beta_0 + \beta_1 2 + \beta_2 8 + \beta_3 16 + \beta_4 4 + \beta_5 64)]^2 \\ [1 - (\beta_0 + \beta_1 1 + \beta_2 0 + \beta_3 0 + \beta_4 1 + \beta_5 \cdot 0)]^2 \\ [-1 - (\beta_0 + \beta_1 3 + \beta_2 2 + \beta_3 6 + \beta_4 9 + \beta_5 \cdot 4)]^2 \end{pmatrix}$

$= 4 - 2 \cdot 2 \cdot (\beta_0 + 4\beta_1 + \beta_2 + 4\beta_3 + 16\beta_4 + \beta_5) + (\beta_0 + 4\beta_1 + \beta_2 + 4\beta_3 + 16\beta_4 + \beta_5)^2 + (-14)^2 - 2(-14)(\beta_0 + 2\beta_1 + 8\beta_2 + 16\beta_3 + 4\beta_4 + 64\beta_5) + (\beta_0 + 2\beta_1 + 8\beta_2 + 16\beta_3 + 4\beta_4 + 64\beta_5)^2 + 1^2 - 2 \cdot 1(\beta_0 + 1\beta_1 + \beta_4) + (\beta_0 + 1\beta_1 + \beta_4)^2 + (-1)^2 - 2(-1)(\beta_0 + 3\beta_1 + 2\beta_2 + 6\beta_3 + 9\beta_4 + 4\beta_5) + (\beta_0 + 3\beta_1 + 2\beta_2 + 6\beta_3 + 9\beta_4 + 4\beta_5)^2$

$\nabla_\beta L_2(y_i, f(x_i)) = \begin{pmatrix} 0 - 4(1) + 2(\beta_0 + 4\beta_1 + \beta_2 + 4\beta_3 + 16\beta_4 + \beta_5) \cdot 1 + 0 + 28 + \\ -2x \end{pmatrix}$

b) Given, $\beta^0 = 0$
$\eta = 0.1$

1st Iteration

$$\beta^1 = \beta^0 + \eta 2X^T(y - X\beta^0)$$
$$= 0 + 0.1 \times 2x$$