Andrea Domuro Casanova
309354
Group 1: Tuesday Tutorial

EXERCISE SHEET 4

1. Linear Regression with Gradient Descent

$$X = \begin{pmatrix} 1 & 1.5 & 2 \\ 1 & 3 & 2.5 \\ 1 & 4.5 & 3 \end{pmatrix} \qquad y = \begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix}$$

$$\mathcal{L}(X,\beta,y) = \sum_{i=1}^{3} (\beta^T x_i - y_i)^2$$

a) $\mathcal{L}(X,\beta,y) = (y - X\beta)^T (y - X\beta)$ $\rightarrow$ the sum of the loss function is the same as this formula

$$\frac{\partial \mathcal{L}(X,\beta,y)}{\partial \beta} = -2X^T(y - X\beta)$$

As our goal is to minimize the loss function, we can $\frac{\partial \mathcal{L}(X,\beta,y)}{\partial \beta} = 0$, and find the β that makes loss function to be 0.

$$-2X^T(y - X\beta) = 0$$

$$-2X^T y + 2X^T X\beta = 0 \Rightarrow \cancel{2}X^T X\beta = \cancel{2}X^T y$$

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

$$(X^T X)^{-1} = \left(\begin{pmatrix} 1 & 1 & 1 \\ 1.5 & 3 & 4.5 \\ 2 & 2.5 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1.5 & 2 \\ 1 & 3 & 2.5 \\ 1 & 4.5 & 3 \end{pmatrix} \right)^{-1} = \begin{pmatrix} 3 & 9 & 7.5 \\ 9 & 31.5 & 24 \\ 7.5 & 24 & 19.25 \end{pmatrix}^{-1} \Rightarrow \text{not invertible.}$$

When $X^T X$ is not invertible, then $X^T X\beta = X^T y$ does not have a uniqe solution. There will be infinite number of solutions β^* that makes the loss function to be minimized.

b) Sometimes analytical solutions gives you a good solution or an approximation of the correct solution, but by using machine learning to learn the solution, the learning model will ~~fied~~ search for different solutions until finds the most optimized one,

~~And also, with the same time it the solution~~

Another reason is that for example in that case, when finding β, using a learning model can be computationally cheaper than with closed-form, when it comes to large data. The calculations can be distributed across multiple processors.

c) $\beta = (1,1,1)^T$

First iteration:

$$\beta^{(1)} = \beta^{(0)} - \mu \nabla \mathcal{L}(X,\beta,y) \qquad \nabla \mathcal{L} = -2X^T(y - X\beta)$$

$$\beta^{(1)} = \begin{pmatrix}1\\1\\1\end{pmatrix} - 0.1\left(-2 \cdot \begin{pmatrix}1 & 1 & 1\\ 1.5 & 3 & 4.5\\ 2 & 2.5 & 3\end{pmatrix} \cdot \left[\begin{pmatrix}10\\15.5\\21\end{pmatrix} - \begin{pmatrix}1 & 1.5 & 2\\ 1 & 3 & 2.5\\ 1 & 4.5 & 3\end{pmatrix}\begin{pmatrix}1\\1\\1\end{pmatrix}\right]\right) =$$

$$= \begin{pmatrix}1\\1\\1\end{pmatrix} - 0.1 \cdot \left(-2 \begin{pmatrix}1 & 1 & 1\\ 1.5 & 3 & 4.5\\ 2 & 2.5 & 3\end{pmatrix} \cdot \begin{bmatrix}5.5\\9\\12.5\end{bmatrix}\right) = \begin{pmatrix}1\\1\\1\end{pmatrix} - 0.1 \cdot \left(-2\begin{pmatrix}27\\91.5\\71\end{pmatrix}\right) = \begin{pmatrix}1\\1\\1\end{pmatrix} + \begin{pmatrix}5.4\\18.3\\14.2\end{pmatrix} = \begin{pmatrix}6.4\\19.3\\15.2\end{pmatrix}$$

$$\beta^{(1)} = \begin{pmatrix}6.4\\19.3\\15.2\end{pmatrix}$$

$$\hat{y}_i = \sum_{i=1}^{n} \beta_i x_i$$

$$MSE = \frac{1}{3}\sum_{i=1}^{3}(y_i - \hat{y}_i)^2 = \frac{1}{3}\Big((10 - (6.4\times1 + 1.5\times19.3 + 2\cdot15.2)) + (15.5 - (6.4\times1 + 19.3\times3 + 15.2\cdot2.5)) +$$

$$+ (21 - (6.5\times1 + 19.3\times4.5 + 15.2\cdot 3))\Big)^2 = \frac{1}{3}\left(-55.75 - 86.8 - 117.95\right)^2 = \underline{22620.1}$$

$$\hookrightarrow RMSE = \underline{150.4}$$

$$\mathcal{L}(X,\beta,y) = (y - X\beta)^T(y - X\beta) = (5.5,\ 9,\ 12.5)\begin{pmatrix}5.5\\9\\12.5\end{pmatrix} = \underline{267.5}$$

$\downarrow$

when $\beta = (1,1,1)^T$

Second iteration

$$\beta^{(2)} = \beta^{(1)} - \mu \cdot \nabla \mathcal{L}(X, \beta, y)$$

$$\beta^{(2)} = \begin{pmatrix} 6.4 \\ 19.3 \\ 15.2 \end{pmatrix} - 0.1 \left[-2 \begin{pmatrix} 1 & 1 & 1 \\ 1.5 & 3 & 4.5 \\ 2 & 2.5 & 3 \end{pmatrix} \cdot \left[\begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix} - \begin{pmatrix} 1 & 1.5 & 2 \\ 1 & 3 & 2.5 \\ 1 & 4.5 & 3 \end{pmatrix} \begin{pmatrix} 6.4 \\ 19.3 \\ 15.2 \end{pmatrix} \right] \right] =$$

$$= \begin{pmatrix} 6.4 \\ 19.3 \\ 15.2 \end{pmatrix} - 0.1 \left(-2 \begin{pmatrix} 1 & 1 & 1 \\ 1.5 & 3 & 4.5 \\ 2 & 2.5 & 3 \end{pmatrix} \cdot \begin{pmatrix} -55.75 \\ -86.8 \\ -117.85 \end{pmatrix} \right) = \begin{pmatrix} 6.4 \\ 19.3 \\ 15.2 \end{pmatrix} - 0.1 \left(-2 \begin{pmatrix} -260.4 \\ -874.35 \\ -682.05 \end{pmatrix} \right) =$$

$$= \begin{pmatrix} 6.4 \\ 19.3 \\ 15.2 \end{pmatrix} - \begin{pmatrix} 52.08 \\ 174.87 \\ 136.41 \end{pmatrix} = \begin{pmatrix} -45.68 \\ -155.57 \\ -212.21 \end{pmatrix}$$

$$MSE = \frac{1}{3} \left((10 - ((-45.68) 1 + (-155.7) \cdot 1.5 + (-212.21) \cdot 2)) + (15.5 - ((-45.68) \cdot 1 + (-155.57) \cdot 3 + (-212.21) \cdot 2.5)) + \right.$$

$$\left. + (21 - ((-45.68) \cdot 1 + (-155.57) \cdot 4.5 + (-212.21) \cdot 3)) \right)^2 = \frac{1}{3} \left(\overset{521.65}{\cancel{773.68}} + \overset{815.415}{\cancel{1058.415}} + \overset{1109.375}{\cancel{1403.375}} \right)^2 =$$

$$1995022.891$$

$= \cancel{[\text{illegible}]}$ RMSE = ~~1833~~.3 1412.45

$$\mathcal{L}(X, \beta^{(1)}, y) = (-55.75, -86.8, -117.85) \begin{pmatrix} -55.75 \\ -86.8 \\ -117.85 \end{pmatrix} = 24530.92$$

↳ when $\beta = (6.4, 19.3, 15.2)^T$

$$\mathcal{L}(X, \beta^{(2)}, y) = \left[\begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix} - \begin{pmatrix} 1 & 1.5 & 2 \\ 1 & 3 & 2.5 \\ 1 & 4.5 & 3 \end{pmatrix} \begin{pmatrix} -45.68 \\ -155.57 \\ -212.21 \end{pmatrix} \right]^T \cdot \left[\begin{pmatrix} 10 \\ 15.5 \\ 21 \end{pmatrix} - \begin{pmatrix} 1 & 1.5 & 2 \\ 1 & 3 & 2.5 \\ 1 & 4.5 & 3 \end{pmatrix} \begin{pmatrix} -45.68 \\ -155.57 \\ -212.21 \end{pmatrix} \right] =$$

$$= (531.455, 830.915, 1130.375) \begin{pmatrix} 531.455 \\ 830.915 \\ 1130.375 \end{pmatrix} = 2250611.79475$$

2. Linear Regression with Stochastic Gradient Descent & Adagrad

a) Normal gradient descent used the exact gradient $\nabla f(x)$ for updating parameters, it means that it ~~used~~ uses $\frac{\delta \mathcal{L}(X,\beta,y)}{\delta \beta}$, being $\mathcal{L}(X,\beta,y)$, the sum of all $(\beta^T x - y)^2$, $\left(\sum_i (\beta^T x_i - y_i)^2\right)$, while stochastic gradient descent only uses one subset of the data. For example, it could use $\frac{\delta \mathcal{L}(x_1,\beta,y_1)}{\delta \beta}$, being $\mathcal{L}(x_1,\beta,y_1) = (\beta^T x_1 - y_1)^2$, and for each iteration, one different subset is taken. Every β update will be consisting of the old β, the learning rate and the $\mathcal{L}(x_i,\beta,y_i)$ of the current ~~&~~ subset.

4.1

(b) First epoch

$$g(\beta^{(0)}) = -2x_i^T(y_i - x_i\beta) = -2\begin{pmatrix}1\\1.5\\2\end{pmatrix}\cdot\left(10 - (1\ \ 1.5\ \ 2)\begin{pmatrix}1\\1\\1\end{pmatrix}\right) = -2\begin{pmatrix}1\\1.5\\2\end{pmatrix}\cancel{8} = \begin{pmatrix}-16\\-24\\-32\end{pmatrix}$$

5.5

$$\Delta x = -g(\beta^{(0)}) = \begin{pmatrix}16\\24\\32\end{pmatrix}$$

$$\beta^{(1)} = \beta^{(0)} + \mu\begin{pmatrix}16\\24\\32\end{pmatrix} = \begin{pmatrix}1\\1\\1\end{pmatrix} + 0.1\begin{pmatrix}16\\24\\32\end{pmatrix} = \begin{pmatrix}2.6\\3.4\\4.2\end{pmatrix}$$

4.2

$$\text{error} = (10 - (2.6\times1 + 3.4\times1.5 + 4.2\times2))^2 = (-6.1)^2 = 37.21$$

$$\text{loss function: } \left(10 - (1\ \ 1.5\ \ 2)\begin{pmatrix}2.6\\3.4\\4.2\end{pmatrix}\right)^T\left(10 - (1\ \ 1.5\ \ 2)\begin{pmatrix}2.6\\3.4\\4.2\end{pmatrix}\right) \cancel{(-16.1)^2} = (-6.1)^2 = 37.21$$

Second epoch →

$$g(\beta^{(1)}) = -2\begin{pmatrix}1\\3\\2.5\end{pmatrix}\cdot\left(15.5 - (1\ \ 3\ \ 2.5)\begin{pmatrix}2.6\\3.4\\4.2\end{pmatrix}\right) = -2\begin{pmatrix}1\\3\\2.5\end{pmatrix}(-7.8) = \begin{pmatrix}15.6\\46.8\\39\end{pmatrix}$$

$$\Delta x = -g(\beta^{(1)}) = -\begin{pmatrix}15.6\\46.8\\39\end{pmatrix}$$

$$\beta^{(2)} = \beta^{(1)} - \mu\begin{pmatrix}15.6\\46.8\\39\end{pmatrix} = \begin{pmatrix}2.6\\3.4\\4.2\end{pmatrix} - \begin{pmatrix}1.56\\4.68\\3.9\end{pmatrix} = \begin{pmatrix}1.04\\-1.28\\0.3\end{pmatrix}$$

$$\text{error} = (15.5 - (1.04\cdot1 + 1.28\cdot3 + 0.3\times2.5))^2 = 17.55^2 = 308.0$$

$$\text{loss function} = \left(15.5 - (1\ \ 3\ \ 2.5)\begin{pmatrix}1.04\\-1.28\\0.3\end{pmatrix}\right)^T\left(15.5 - (1\ \ 3\ \ 2.5)\begin{pmatrix}1.04\\-1.28\\0.3\end{pmatrix}\right) = \underline{\underline{308}}$$

c) First epoch:

$$g(\beta^{(0)}) = \begin{pmatrix} -16 \\ -24 \\ -32 \end{pmatrix}$$

G = 0
G_1 = G + g*g

mu = lambda / (sqrt(G) + epsilon)

5.1

$$\Delta x = \begin{pmatrix} 16 \\ 24 \\ 32 \end{pmatrix}$$

$$\mu_{Adagrad} = \frac{\mu}{\sqrt{G}} = \frac{0.1}{\sqrt{1}} = 0.1$$

$$\beta^{(1)} = \beta^{(0)} + \mu \cdot \begin{pmatrix} 16 \\ 24 \\ 32 \end{pmatrix} = \begin{pmatrix} 2.6 \\ 3.4 \\ 4.2 \end{pmatrix}$$

The first epoch is the same as the first epoch from part b)

error = 37.21

loss function = 37.21

Second epoch:

$$g(\beta^{(1)}) = \begin{pmatrix} 15.6 \\ 46.8 \\ 39 \end{pmatrix}$$

G = G_1 + g*g

$$\Delta x = -\begin{pmatrix} 15.6 \\ 46.8 \\ 39 \end{pmatrix}$$

$$\mu_{Adagrad} = \frac{0.1}{\begin{pmatrix} \sqrt{15.6^2} \\ \sqrt{46.8^2} \\ \sqrt{39^2} \end{pmatrix}} = \begin{bmatrix} \frac{0.1}{\sqrt{15.6}} \\ \frac{0.1}{\sqrt{46.8}} \\ \frac{0.1}{\sqrt{39}} \end{bmatrix}$$

$$\beta^{(2)} = \beta^{(1)} - \mu_{Adagrad} \begin{pmatrix} 15.6 \\ 46.8 \\ 3.9 \end{pmatrix} = \begin{pmatrix} 2.6 \\ 3.4 \\ 4.2 \end{pmatrix} - \begin{bmatrix} \frac{0.1}{3.95} \\ \frac{0.1}{6.84} \\ \frac{0.1}{6.24} \end{bmatrix} \odot \begin{pmatrix} 15.6 \\ 46.8 \\ 3.9 \end{pmatrix} =$$

$$= \begin{pmatrix} 2.6 \\ 3.4 \\ 4.2 \end{pmatrix} - \begin{pmatrix} 0.39 \\ 0.68 \\ 0.0625 \end{pmatrix} = \begin{pmatrix} 2.21 \\ 2.72 \\ 4.14 \end{pmatrix}$$

$$\text{error} = (15.5 - (2.21 \times 1 + 2.72 \cdot 3 + 4.14 \times 2.5))^2 = 27.25$$

loss function = 27.25

Adagrad helped in minimizing the error and the loss function comparing to the second iteration of stochastic gradient without Adagrad.

Index der Kommentare

4.1 should be 5.5

4.2 equations are correct but earlier mistake affect all the results

5.1 Note : in adagrad you do 3 iteration inside every epoch (similar to SGD) but divide by square root of the sum of gradients