

## sheet 4 solution

Ahmed Ashraf Mohamed, 2103134

### Question 1

Consider a perceptron model consisting of two input neurons, denoted by  $x_1$  and  $x_2$ , each with two weights  $w_{11}$  and  $w_{21}$ , and a bias term  $b_1$ . These inputs feed into a single output neuron with one weight  $w_1$  and a bias term  $b_2$ . The output of the perceptron can be written as:

$$y = f(w_{11}x_1 + w_{21}x_2 + b_1)w_1 + b_2$$

where  $f$  is the activation function.

Assume that  $f$  is a linear activation function, i.e.,  $f(x) = x$ .

In this case, we can simplify the above equation as:

$$\begin{aligned} y &= w_{11}x_1w_1 + w_{21}x_2w_1 + b_1w_1 + b_2 \\ &= w_1(w_{11}x_1 + w_{21}x_2 + b_1) + b_2 \end{aligned}$$

We can see that the output of the perceptron is a linear function of the inputs, which means that the perceptron is limited to representing only linear decision boundaries. This is known as the "linear collapse" problem, where the perceptron can only learn linearly separable patterns.

To overcome this limitation, we can introduce a non-linear activation function, such as the sigmoid function, which is defined as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Using the sigmoid activation function, the output of the perceptron can be written as:

$$\begin{aligned} y &= f(w_{11}x_1 + w_{21}x_2 + b_1)w_1 + b_2 \\ &= \frac{1}{1 + e^{-(w_{11}x_1 + w_{21}x_2 + b_1)}}w_1 + b_2 \end{aligned}$$

Now, we can see that the output of the perceptron is a non-linear function of the inputs, which allows the perceptron to represent non-linear decision boundaries. This is the benefit of using a non-linear activation function in a perceptron model.

In summary, the linear collapse problem arises when a perceptron model with linear activation function is unable to represent non-linear decision boundaries. By introducing a non-linear activation function, we can overcome this limitation and allow the perceptron to represent more complex decision boundaries.

## Question 2

Student	Highschool GPA	First-year college GPA
1	2.5	2.8
2	3.0	3.5
3	3.2	3.3
4	3.8	3.9

- Compute the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of the high school GPAs and first-year college GPAs.:

	$\mu$	$\sigma$
Highschool GPA	3.125	0.465698
First-year college GPA	3.375	3.5

- Compute the slope and intercept of the least-squares regression of first-year college GPA

$$\text{Intercept} = B_0 = \bar{y} - B_1 \bar{x}$$

$$\text{slope} = B_1 = \frac{\sum xy - \frac{1}{n} \sum x \sum y}{\sum x^2 - \frac{1}{n} (\sum x)^2}$$

$$n = 4$$

$$\sum x^2 = 39.93$$

$$\sum x = 12.5$$

$$(\sum x)^2 = 156.25$$

$$\sum y = 13.5$$

$$\sum xy = 42.88$$

$$B_1 = \frac{42.88 - \frac{12.5 \cdot 13.5}{4}}{39.93 - \frac{(12.5)^2}{4}} \approx 0.798$$

$$B_0 = \frac{13.5}{4} - 0.798 \times \frac{12.5}{4} \approx 0.8804$$

- Use the regression equation to predict the first-year college GPA of a student who had a high school GPA of 3.5.

$$\hat{y} = B_0 + B_1 \cdot x = 0.8804 + 0.798 \times 3.5 \approx 3.67$$

### Question 3

Houses	Price	size
1	120	2.0
2	140	2.5
3	130	2.2
4	150	2.7

- Compute the sample mean and sample standard deviation of the house prices and sizes.

	$\mu$	$\sigma$
House Price	135	11.18
House size	2.35	0.26925

- Compute the slope and intercept of the least-squares regression line of house price on size

$x$  represents the size,  $y$  represents the price.

$$\text{intercept} = B_0 = \bar{y} - B_1 \bar{x}$$

$$\text{slope} = B_1 = \frac{\sum xy - \frac{1}{n} \sum x \sum y}{\sum x^2 - \frac{1}{n} (\sum x)^2}$$

$$\sum x^2 = 22.38$$

$$\sum x = 9.4$$

$$(\sum x)^2 = 88.36$$

$$\sum y = 540$$

$$\sum xy = 1281$$

$$n = 4$$

$$B_1 = \frac{1281 - 1269}{22.38 - 22.09} \approx 41.379$$

$$B_0 = \frac{540}{4} - 41.379 \cdot \frac{9.4}{4} \approx 37.759$$

- Use the regression equation to predict the price of a house that has a size of 3.0 thousand square feet.

$$\hat{y} = B_0 + B_1 \cdot x = 37.759 + 41.379 \times 3.0 \approx 161.89$$

## Question 4

Compute the logistic regression coefficients for the model  $y = b_0 + b_1 x_1 + b_2 x_2$ .

initially we assume that  $w_0 = w_1 = w_2 = 0.5$  and use the sigmoid as the activation function where the threshold is 0.5

$$f(x) = \frac{1}{1 + e^{-x}}$$

$$\text{act}(f(x)) = \begin{cases} 1 & \text{if } f(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

$$Z_1 = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

$$= 0.5 + (0.5) \cdot 45 + (0.5) \cdot 0 = 23$$

$$\hat{y} = \text{act}(f(23)) = 1$$

$$\text{error} = y - \hat{y} = 1 - 1 = 0$$


---

$$\begin{aligned}
Z_2 &= b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \\
&= 0.5 + (0.5) \times 52 + (0.5) \times 1 = 27 \\
\hat{y} &= \text{act}(f(27)) = 1 \\
\text{error} &= y - \hat{y} = 0 - 1 = -1 \\
b_{\text{new}} &= b_{\text{old}} + (\text{error} \cdot x_i) \\
b_0 &= 0.5 - 1 = -0.5 \\
b_1 &= 0.5 + (-1 \times 52) = -51.5 \\
b_2 &= 0.5 + (-1 \times 1) = -0.5
\end{aligned}$$


---

$$\begin{aligned}
Z_3 &= b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \\
&= -0.5 + (-51.5) \times 27 + (-0.5) \times 1 = -1391.5 \\
\hat{y} &= \text{act}(f(-1391.5)) = 0 \\
\text{error} &= y - \hat{y} = 1 - 0 = 1 \\
b_{\text{new}} &= b_{\text{old}} + (\text{error} \cdot x_i) \\
b_0 &= -0.5 + 1 = 0.5 \\
b_1 &= -51.5 + (1 \times 27) = -24.5 \\
b_2 &= -0.5 + (1 \times 1) = 0.5
\end{aligned}$$


---

$$\begin{aligned}
Z_4 &= b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \\
&= 0.5 + (-24.5) \times 31 + (0.5) \times 0 = -759 \\
\hat{y} &= \text{act}(f(759)) = 0 \\
\text{error} &= y - \hat{y} = 0 - 0 = 0
\end{aligned}$$


---

$$\begin{aligned}
Z_5 &= b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \\
&= 0.5 + (-24.5) \times 48 + (0.5) \times 1 = -1175 \\
\hat{y} &= \text{act}(f(-1175)) = 0 \\
\text{error} &= y - \hat{y} = 1 - 0 = 1 \\
b_{\text{new}} &= b_{\text{old}} + (\text{error} \cdot x_i) \\
b_0 &= 0.5 + 1 = 1.5 \\
b_1 &= -24.5 + (1 \times 48) = 23.5 \\
b_2 &= 0.5 + (1 \times 1) = 1.5
\end{aligned}$$

- Use the logistic regression equation to predict the probability that a 35-

year-old non-smoker will have a positive outcome.

$$\begin{aligned}y &= 1.5 + 23.5 \cdot x_1 + 1.5 \cdot x_2 \\&= 1.5 + 23.5 \times 35 + 1.5 \times 0 = 824 \\ \hat{y} &= act(f(x)) = act(f(824)) = 1\end{aligned}$$

## Question 5

- Write down the expression for the overall loss in terms of  $B_0$  and  $B_1$

$$MSE = \frac{1}{n} \sum (y - (B_1 \cdot x + B_0))^2$$

- Derive the partial derivatives of the loss with respect to each coefficient

$$\begin{aligned}\frac{\delta MSE}{\delta B_0} &= \frac{1}{n} \sum (-2 \cdot (y - (B_1 \cdot x + B_0))) \\ \frac{\delta MSE}{\delta B_1} &= \frac{1}{n} \sum (-2 \cdot (y - (B_1 \cdot x + B_0)) \cdot x)\end{aligned}$$

- Explain the meaning of these partial derivatives in the context of training the logistic regression model and how they can be used to update the values of the coefficients during training:

Partial derivatives is used to get the rate of change of loss function with respect of weight and the update the rule as follow

$$\begin{aligned}B_{0,new} &= B_{0,old} - \alpha \cdot \frac{\delta}{\delta B_0} \text{error} \\ B_{1,new} &= B_{1,old} - \alpha \cdot \frac{\delta}{\delta B_1} \text{error}\end{aligned}$$

## Question 6

$$L(\theta) = y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

Write down the expression for the overall loss in terms of  $B_0$  and  $B_1$ :

$$L(\theta) = \sum y \log(\alpha_1 \cdot x + B_0) + (1 - y) \log(1 - B_1 \cdot x + B_0)$$

Derive the partial derivatives of the loss with respect to each coefficient:

$$\begin{aligned}\frac{\delta L(\theta)}{\delta B_1} &= \sum y \cdot \frac{x}{\ln(10) \cdot B_1 \cdot x + B_0} + (1 - y) \frac{-x}{\ln(10) \cdot (1 - B_1 \cdot x + B_0)} \\ \frac{\delta L(\theta)}{\delta B_0} &= \sum y \cdot \frac{1}{\ln(10) \cdot \alpha_1 \cdot x + B_0} + (1 - y) \frac{-1}{\ln(10) \cdot (1 - B_1 \cdot x + B_0)}\end{aligned}$$