Natural Language Processing

Tutorial 6 (Week 9): ML & DL

Question 1

Imagine you're working on a basic sentiment analysis task. You have collected a small dataset where each data point consists of the number of positive words in a movie review and the corresponding rating given by the reviewer on a scale of 1 to 10. Now use a simple **linear regression model** to **predict the movie rating based on the number of positive words**. Your model should be defined as: $y = wx + b$, where x = number of positive words.

Question 1

Use the following data to derive

(1) The optimal weight w and bias b using the least squares method.

Number of Positive Words	Movie Rating
1	3
3	6
5	8
6	9

Solution 1 (1)

$$J(w,b) = \sum_i (y_i - \hat{y}_i)^2$$
$$= \sum_i (y_i - (wx_i + b))^2$$

This is convex, compute partial derivatives and equate to 0

$$\frac{\partial J}{\partial w} = -2\sum_i x_i(y_i - (wx_i + b)) = 0$$
$$\frac{\partial J}{\partial b} = -2\sum_i (y_i - (wx_i + b)) = 0$$

Solution 1 (1)

This is convex, compute partial derivatives and equate to 0

$$\frac{\partial J}{\partial w} = -2\sum_i x_i(y_i - (wx_i + b)) = 0$$

$$\frac{\partial J}{\partial b} = -2\sum_i (y_i - (wx_i + b)) = 0$$

Rearrange the above equations to get

$$\sum_i x_i y_i = w\sum_i x_i^2 + b\sum_i x_i$$

$$\sum_i y_i = w\sum_i x_i + nb$$

Solution 1 (1)

Rearrange the above equations to get

$$\sum_i x_i y_i = w \sum_i x_i^2 + b \sum_i x_i$$

$$\sum_i y_i = w \sum_i x_i + nb$$

Solving this system of 2 equations over 2 variables, we get

$$w = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} \qquad b = \frac{\sum_i y_i - w \sum_i x_i}{n}$$

Solution 1 (1)

Solving this system of 2 equations over 2 variables, we get

$$w = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \qquad b = \frac{\sum_i y_i - w\sum_i x_i}{n}$$

Now plug in the values from the table

$n = 4$

Number of Positive Words (x_i)	Movie Rating (y_i)
1	3
3	6
5	8
6	9

$$w = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - (\sum_i x_i)^2}$$

$$b = \frac{\sum_i y_i - w\sum_i x_i}{n}$$

$$\boxed{w = \frac{4(13+36+58+69)-(1+3+5+6)(3+6+8+9)}{4(1+9+25+36)-(1+3+5+6)^2} \approx 1.2}$$

$$\boxed{b = \frac{(3+6+8+9)-1.2(1+3+5+6)}{4} = 2}$$

Now plug in the values from the table

$$n = 4$$

Number of Positive Words (x_i)	Movie Rating (y_i)
1	3
3	6
5	8
6	9

Question 1

Use the following data to derive

(2) Once you have the parameters, what would be the predicted rating for a review with 4 positive words?

Number of Positive Words	Movie Rating
1	3
3	6
5	8
6	9

Solution 1 (2)

$$x = 4, w = 1.2, b = 2, y = ?$$

Number of Positive Words	Movie Rating
1	3
3	6
5	8
6	9

Solution 1 (2)

$$x = 4, w = 1.2, b = 2, y = ?$$

$$\boxed{y = wx + b = 1.2 \times 4 + 2 \approx 7}$$

Number of Positive Words	Movie Rating
1	3
3	6
5	8
6	9

Question 2

You're trying to predict whether a student will pass or fail an upcoming exam based on the number of hours they've studied. You have the following data:

Hours studied (x)	Pass (1) or Fail (0)
1	0
2	0
4	1
5	1

(1) Write the logistic regression function given the input x.

Solution 2 (1)

- The logistic function can be written as the sigmoid function:

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad z = wx + b$$

Question 2

You're trying to predict whether a student will pass or fail an upcoming exam based on the number of hours they've studied. You have the following data:

Hours studied (x)	Pass (1) or Fail (0)
1	0
2	0
4	1
5	1

(2) Compute the logistic loss for the first data point (x=1, y=0) given that w=1 and b=−3.

Solution 2 (2)

- Given w=1, b=-3, we can have $z = wx + b = 1 \times 1 - 3 = -2$
- The predicted output becomes: $\hat{y} = \sigma(-2) = \frac{1}{1+e^2}$
- According to the logistic loss function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

- Plug in the values to obtain

$$L(0, \hat{y}) = -\log(1 - \hat{y}) = -\log \frac{e^2}{1+e^2}$$

Question 2

You're trying to predict whether a student will pass or fail an upcoming exam based on the number of hours they've studied. You have the following data:

Hours studied (x)	Pass (1) or Fail (0)
1	0
2	0
4	1
5	1

(3) Derive the gradient of the logistic loss with respect to w and b. Then compute the gradient for the first data point using the same w and b values.

Solution 2 (3)

Enumerate the entire forward process

$$z = wx + b$$

$$\hat{y} = \sigma(z)$$

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

Use chain rule

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w}$$

Solution 2 (3)

- According to the logistic loss function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

- The gradient wrt $\hat{y}$

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$$

Solution 2 (3)

- The gradient wrt $\hat{y}$ $\quad \frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$

- The gradient wrt z

$$\hat{y} = \frac{1}{1+e^{-z}} \qquad \frac{\partial \hat{y}}{\partial z} = -\frac{-1 \times e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

Solution 2 (3)

- The gradient wrt $\hat{y}$ $$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}$$
- The gradient wrt z

$$\hat{y} = \frac{1}{1+e^{-z}} \qquad \frac{\partial \hat{y}}{\partial z} = -\frac{-1 \times e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}$$

- The gradient wrt w $$\frac{\partial z}{\partial w} = x$$

Solution 2 (3)

- Applying chain rule

$$\boxed{\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}} \quad \boxed{\frac{\partial \hat{y}}{\partial z} = -\frac{-1 \times e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2}} \quad \boxed{\frac{\partial z}{\partial w} = x}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w} = \left(-\frac{y}{\hat{y}} + \frac{(1-y)}{1-\hat{y}}\right)\frac{e^{-z}}{(1+e^{-z})^2}x$$

$$= -\frac{ye^{-z}}{1+e^{-z}}x + \frac{1-y}{1+e^{-z}}x = \frac{-yx(1+e^{-z})+x}{1+e^{-z}}$$

$$= -yx + \frac{x}{1+e^{-z}} = (\hat{y} - y)x$$

Solution 2 (3)

- The gradient wrt w is $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w} = (\hat{y} - y)x$

- Similarly, the gradient wrt b is

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial b} = (\hat{y} - y)$$

Solution 2 (3)

- The gradient for the first data point (x=1, y=0) is

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w} = (\hat{y} - y)x = (\hat{y} - 0) \times 1 = \frac{1}{1+e^2}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial b} = (\hat{y} - y) = \frac{1}{1+e^2}$$

Coding Practice

- Linear Regression: https://colab.research.google.com/drive/12QpBf7x_Jt6-zypN4OrUFFHXz1u6CmYe?usp=sharing
- Logistic Regression: https://colab.research.google.com/drive/1nTrYW5dUu6WO9cx7SGEvP9oX7qRbsGJk?usp=sharing