HW3 Written Part

Due 5pm Friday Feb 21, 2020

What is your name?

Problem 1

Suppose I fit a neural network model with the following structure:

- Input layer has 2 features
- One hidden layer with 2 units and a linear activation function
- Output layer has 1 unit and a sigmoid activation function

Show that this model is equivalent to a logistic regression model in the sense that the activation in the output layer could be written as

$$a_1^{(i)[2]} = \frac{\exp(b^* + w_{11}^* x_1^{(i)} + w_{12}^* x_2^{(i)})}{1 + \exp(b^* + w_{11}^* x_1^{(i)} + w_{12}^* x_2^{(i)})}$$

for some parameters b^*, w_{11}^*, and w_{12}^* that are combinations of the biases and weights in all units of the full neural network model. Your answer should give exact formulas for how to calculate b^*, w_{11}^*, and w_{12}^* in terms of the neural network parameters $b_1^{[1]}$, $w_{11}^{[1]}$, $w_{12}^{[1]}$, $b_2^{[1]}$, $w_{21}^{[1]}$, $w_{22}^{[1]}$, $b_1^{[2]}$, $w_{11}^{[2]}$, and $w_{12}^{[2]}$. Comment briefly (1 sentence) on why it is necessary to use non-linear activation functions in neural network models.

$$a_1^{(i)[1]} = b_1^{[1]} + w_{11}^{[1]} x_1^{(i)} + w_{12}^{[1]} x_2^{(i)}$$

$$a_2^{(i)[1]} = b_2^{[1]} + w_{21}^{[1]} x_1^{(i)} + w_{22}^{[1]} x_2^{(i)}$$

$$a_1^{(i)[2]} = \text{sigmoid}\left(b_1^{[2]} + w_{11}^{[2]} a_1^{(i)[1]} + w_{12}^{[2]} a_1^{(i)[1]}\right)$$

$$= \text{sigmoid}\left(b_1^{[2]} + w_{11}^{[2]}\left(b_1^{[1]} + w_{11}^{[1]} x_1^{(i)} + w_{12}^{[1]} x_2^{(i)}\right) + w_{12}^{[2]}\left(b_2^{[1]} + w_{21}^{[1]} x_1^{(i)} + w_{22}^{[1]} x_2^{(i)}\right)\right)$$

$$= \text{sigmoid}\left(b_1^{[2]} + b_1^{[1]} w_{11}^{[2]} + b_2^{[1]} w_{12}^{[2]} + \left(w_{11}^{[2]} w_{11}^{[1]} + w_{12}^{[2]} \cdot w_{21}^{[1]}\right) x_1^{(i)} + \left(w_{11}^{[2]} w_{12}^{[1]} + w_{12}^{[2]} w_{22}^{[1]}\right) x_2^{(i)}\right)$$

$$= \text{sigmoid}\left(b^* + w_{11}^* x_1^{(i)} + w_{12}^* x_2^{(i)}\right)$$

where $b^* = b_1^{[2]} + b_1^{[1]} w_{11}^{[2]} + b_2^{[1]} w_{12}^{[2]}$, $w_{11}^* = w_{11}^{[2]} w_{11}^{[1]} + w_{12}^{[2]} \cdot w_{21}^{[1]}$,

and $w_{12}^* = w_{11}^{[2]} w_{12}^{[1]} + w_{12}^{[2]} w_{22}^{[1]}$.

A hidden layer with only a linear activation doesn't add any thing to the model in terms of representative capacity. We need non-linear activations to get a more flexible model.

Problem 2

Suppose I am working on a classification problem where the response has three classes and I have two input features. I will use a neural network model with the following structure:

- Input layer has 2 features
- One hidden layer has 2 units and a relu activation
- Output layer has 3 units and a softmax activation

My full data set has 100 observations in it.

(a) Give the shapes of each of the following quantities. Use the convention that each observation is in a column of X and each feature is in a row of X. For example, if I am predicting whether an animal is a bird, a cat, or a dog using its weight and its height, the weights and height for the first animal in my data set would be in the first column of X. Also, suppose we are using a one-hot encoding for the response, so if the first animal in my data set is a dog then I will have $y^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

- X (2, 100)
- y (3, 100)
- $z^{[1]}$ (2, 100) ← my answers were wrong. ☺
- $a^{[1]}$ (2, 100) ←
- $b^{[1]}$ (2, 1)
- $w^{[1]}$ (2, 2)
- $z^{[2]}$ (3, 100)
- $a^{[2]}$ (3, 100)
- $b^{[2]}$ (3, 1)
- $w^{[2]}$ (2, 3)
- $\frac{\partial J(b,w)}{\partial a^{[2]}}$ (3, 100)
- $\frac{\partial J(b,w)}{\partial z^{[2]}}$ (3, 100)
- $\frac{\partial J(b,w)}{\partial b^{[2]}}$ (3, 1)
- $\frac{\partial J(b,w)}{\partial w^{[2]}}$ (2, 3)
- $\frac{\partial J(b,w)}{\partial a^{[1]}}$ (2, 100)
- $\frac{\partial J(b,w)}{\partial z^{[1]}}$ (2, 100)
- $\frac{\partial J(b,w)}{\partial b^{[1]}}$ (2, 1)
- $\frac{\partial J(b,w)}{\partial w^{[1]}}$ (2, 2)

(b) In the backpropagation algorithm, why do we calculate $\frac{\partial J(b,w)}{\partial a^{[2]}}$ before we calculate $\frac{\partial J(b,w)}{\partial z^{[2]}}$? Your answer should involve a formula for how $a^{[2]}$ is calculated and an application of the chain rule.

$a^{[2]} = g^{[2]}(z^{[2]})$ and $z^{[2]}$

Since $a^{[2]}$ depends on b and w, we could write

$J(b,w) = J(a^{[2]}(z^{[2]}(b,w)))$.

Then $\frac{\partial J}{\partial z^{[2]}} = \frac{\partial J}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}}$ from the chain rule.

We must calculate $\frac{\partial J}{\partial a^{[2]}}$ before $\frac{\partial J}{\partial z^{[2]}}$ because $z^{[2]}$ is used in the calculation of $a^{[2]}$ so the chain rule says the calculation of $\frac{\partial J}{\partial z^{[2]}}$ will involve $\frac{\partial J}{\partial a^{[2]}}$.

(c) In the backpropagation algorithm, why do we calculate $\frac{\partial J(b,w)}{\partial z^{[2]}}$ before we calculate $\frac{\partial J(b,w)}{\partial b^{[2]}}$? Your answer should involve a formula for how $z^{[2]}$ is calculated and an application of the chain rule.

$z^{[2]} = b^{[2]} + (w^{[2]})^T a^{[1]}$

Therefore ~~$J(b,w)$~~ we could write $J(b,w) = J(z^{[2]}(b,w))$,

and so by the chain rule

$$\frac{\partial J}{\partial b^{[2]}} = \frac{\partial J}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}}$$

Because $b^{[2]}$ is used in the calculation of $z^{[2]}$, the chain rule says $\frac{\partial J}{\partial z^{[2]}}$ will be used in the calculation of $\frac{\partial J}{\partial b^{[2]}}$, so we need to calculate it first.

Problem 3

Suppose I am working on a classification problem where the response has two classes (say dog and cat) and I have one input feature. In the model statements below, I'm suppressing as much notation as possible.

Our first option for this task is a logistic regression model where $Y^{(i)}$ is encoded as 0 for a dog or 1 for a cat:

$$Y^{(i)} \sim \text{Bernoulli}(a^{(i)})$$
$$a_1^{(i)} = \frac{\exp(b + w_1 x^{(i)})}{1 + \exp(b + w_1 x^{(i)})}$$

However, a reasonable person might also formulate this as a multinomial regression problem using a one-hot encoding of $Y^{(i)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ for a dog or $Y^{(i)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ for a cat:

$$Y^{(i)} \sim \text{Categorical}(a_1^{(i)}, a_2^{(i)})$$
$$a_1^{(i)} = \frac{\exp(b_1 + w_1 x^{(i)})}{\exp(b_1 + w_1 x^{(i)}) + \exp(b_2 + w_2 x^{(i)})}$$
$$a_2^{(i)} = \frac{\exp(b_2 + w_2 x^{(i)})}{\exp(b_1 + w_1 x^{(i)}) + \exp(b_2 + w_2 x^{(i)})}$$

Note that by convention the numbering of classes is 0 and 1 in the logistic regression model, but 1 and 2 in the multinomial regression model. So class 1 in the logistic regression model refers to the same thing as class 2 in the multinomial regression model. This is awkward but I think it'll be more confusing if we change the standard notation...

Suppose these models will be estimated by gradient descent, and the parameter values for the two models are intialized so that for any value of x the initial estimated probability of being a cat from the logistic model is equal to the initial estimated probability of being a cat from the multinomial regression model.

(a) Write down the formulas for the updates to b and w for the logistic regression model in terms of $a^{(i)}$, $y^{(i)}$, and $x^{(i)}$, $i = 1, \ldots, m$. Note that you don't need to calculate the value of $a^{(i)}$ explicitly in terms of b and w.

$$\frac{\partial J}{\partial z} = \begin{bmatrix} a^{(1)} - y^{(1)} & a^{(2)} - y^{(2)} & \cdots & a^{(m)} - y^{(m)} \end{bmatrix}$$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J^{(i)}}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J^{(i)}}{\partial z^{(i)}} \cdot \frac{\partial z^{(i)}}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J^{(i)}}{\partial z^{(i)}} \cdot 1 = \frac{1}{m} \sum_{i=1}^{m} \left(a^{(i)} - y^{(i)} \right)$$

$$\frac{\partial J}{\partial w} = \frac{1}{m} X \cdot \left(\frac{\partial J}{\partial z} \right)^T = \frac{1}{m} \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(m)} \end{bmatrix} \begin{bmatrix} a^{(1)} - y^{(1)} \\ a^{(2)} - y^{(2)} \\ \vdots \\ a^{(m)} - y^{(m)} \end{bmatrix} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \left(a^{(i)} - y^{(i)} \right)$$

$$b^{new} = b^{old} - \alpha \cdot \frac{\partial J}{\partial b} = b^{old} - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(a^{(i)} - y^{(i)} \right)$$

$$w^{new} = w^{old} - \alpha \cdot \frac{\partial J}{\partial w} = w^{old} - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} x^{(i)} \left(a^{(i)} - y^{(i)} \right)$$

where α is the learning rate

If you want to do derivations "from scratch":

$$J(b,w) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(a^{(i)}) + (1-y^{(i)})\log(1-a^{(i)})\right]$$

$$\frac{\partial J(b,w)}{\partial a^{(i)}} = \frac{-1}{m}\left[\frac{y^{(i)}}{a^{(i)}} + \frac{1-y^{(i)}}{1-a^{(i)}}(-1)\right]$$

$$a^{(i)} = \frac{\exp(z^{(i)})}{1+\exp(z^{(i)})}, \quad 1-a^{(i)} = \frac{1}{1+\exp(z^{(i)})}$$

$$\frac{\partial a^{(i)}}{\partial z^{(i)}} = \frac{\exp(z^{(i)})\left(1+\exp(z^{(i)})\right) - \exp(z^{(i)})\exp(z^{(i)})}{\left(1+\exp(z^{(i)})\right)^2}$$

$$= \frac{\exp(z^{(i)})}{\left(1+\exp(z^{(i)})\right)^2} = a^{(i)}\cdot(1-a^{(i)})$$

$$\frac{\partial J(b,w)}{\partial z^{(i)}} = \frac{\partial J(b,w)}{\partial a^{(i)}}\cdot\frac{\partial a^{(i)}}{\partial z^{(i)}}$$

$$= -\frac{1}{m}\left[\frac{y^{(i)}}{a^{(i)}} + \frac{1-y^{(i)}}{1-a^{(i)}}(-1)\right]\cdot a^{(i)}(1-a^{(i)})$$

$$= \frac{-1}{m}\left[y^{(i)} - y^{(i)}a^{(i)} - a^{(i)} + y^{(i)}a^{(i)}\right]$$

$$= \frac{1}{m}\left[a^{(i)} - y^{(i)}\right]$$

(b) Suppose we have two observations with feature, response, and output layer activation values for a logistic regression model as given in the table below. The current parameter values are $b = 1$ and $w = -1$. Find the updated parameter values after one step using a learning rate of $\alpha = 0.1$.

$x^{(i)}$	$y^{(i)}$	$a_1^{(i)}$
1	1	0.5
2	0	0.269

i. Find $\frac{\partial J(b,w)}{\partial z^{[1]}}$. (First, think about what its shape should be.) Shape (1, 2)

$$\frac{\partial J}{\partial z} = [a^{(1)} - y^{(1)} \quad a^{(2)} - y^{(2)}] = [0.5 - 1 \quad 0.269 - 0]$$
$$= [-0.5 \quad 0.269]$$

ii. Find $\frac{\partial J(b,w)}{\partial b}$. (First, think about what its shape should be.) Shape (1, 1)

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial J^{(i)}}{\partial z^{(i)}} = \frac{1}{m}\sum_{i=1}^{m}(a^{(i)} - y^{(i)}) = \frac{1}{2}(-0.5 + 0.269) = -0.1155$$

iii. Find $\frac{\partial J(b,w)}{\partial w}$. (First, think about what its shape should be.)

$$\frac{\partial J}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}(a^{(i)} - y^{(i)}) = \frac{1}{2}\{1 \cdot (-0.5) + 2 \cdot 0.269\} = 0.019$$

iv. Find the updated values of b and w from one gradient descent update step using a learning rate of $\alpha = 0.01$.

$$b^{new} = b^{old} - \alpha \cdot \frac{\partial J}{\partial b} = 1 - 0.01 \cdot (-0.1155) = 1.001155$$

$$w^{new} = w^{old} - \alpha \cdot \frac{\partial J}{\partial w} = -1 - 0.01 \cdot 0.019 = -1.00019$$

(c) Write down the formulas for the updates to b and w for the multinomial regression model in terms of $a^{(i)}$, $y^{(i)}$, and $x^{(i)}$, $i = 1, \ldots, m$. Note that you don't need to calculate the value of $a^{(i)}$ explicitly in terms of b and w.

$$\frac{\partial J}{\partial Z} = \begin{bmatrix} a^{(1)} - y^{(1)} & a^{(2)} - y^{(2)} & \cdots & a^{(m)} - y^{(m)} \end{bmatrix} \leftarrow \text{note this is of shape } (2, m)$$

$$\frac{\partial J}{\partial b} = \cdots = \frac{1}{m} \sum_{i=1}^{m} \left(a^{(i)} - y^{(i)} \right) \leftarrow \text{shape } (2, 1)$$

$$\frac{\partial J}{\partial w} = \cdots = \frac{1}{m} X \cdot \left(\frac{\partial J}{\partial Z} \right)^T = \frac{1}{m} \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(m)} \end{bmatrix} \begin{bmatrix} (a^{(1)} - y^{(1)})^T \\ (a^{(2)} - y^{(2)})^T \\ \vdots \\ (a^{(m)} - y^{(m)})^T \end{bmatrix}$$

$$= \frac{1}{m} \sum_{i=1}^{m} x^{(i)} (a^{(i)} - y^{(i)})^T \leftarrow \text{shape } (1, 2) \text{ since each } (a^{(i)} - y^{(i)}) \text{ has shape } (2, 1). \text{ this matches the shape of } w.$$

$$b^{new} = b^{old} - \alpha \cdot \frac{\partial J}{\partial b} = b^{old} - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} (a^{(i)} - y^{(i)})$$

$$w^{new} = w^{old} - \alpha \cdot \frac{\partial J}{\partial w} = w^{old} - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} x^{(i)} (a^{(i)} - y^{(i)})^T$$

(d) Suppose I have two observations with feature, response, and output layer activation values for a multinomial regression model as given in the table below. My current parameter values are $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $w = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$. Find the updated parameter values after one step using a learning rate of $\alpha = 0.1$.

Should be $w = [0 \ -1]$

$x^{(i)}$	$y^{(i)}$	$a_1^{(i)}$	$a_2^{(i)}$
1	1	0.5	0.5
2	0	0.731	0.269

i. Find $\frac{\partial J(b,w)}{\partial z^{[1]}}$. (First, think about what its shape should be.) Shape $(2, m) = (2, 2)$

$$\frac{\partial J}{\partial z} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 0.731 \\ 0.269 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5 & -0.269 \\ -0.5 & 0.269 \end{bmatrix}$$

ii. Find $\frac{\partial J(b,w)}{\partial b}$. (First, think about what its shape should be.) Shape $(2, 1)$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} (a^{(i)} - y^{(i)}) = \frac{1}{2} \left(\begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} + \begin{bmatrix} -0.269 \\ 0.269 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 0.1155 \\ -0.1155 \end{bmatrix}$$

iii. Find $\frac{\partial J(b,w)}{\partial w}$. (First, think about what its shape should be.) Shape $(1, 2)$

$$\frac{\partial J}{\partial w} = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} (a^{(i)} - y^{(i)})^T$$

$$= \frac{1}{2} \left(1 \cdot [0.5 \quad -0.5] + 2 \cdot [-0.269 \quad 0.269] \right)$$

$$= [-0.019 \quad 0.019]$$

iv. Find the updated values of b and w from one gradient descent update step using a learning rate of $\alpha = 0.01$.

$$b^{new} = b^{old} - \alpha \cdot \frac{\partial J}{\partial b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 0.01 \begin{bmatrix} 0.1155 \\ -0.1155 \end{bmatrix} = \begin{bmatrix} -0.001155 \\ 1.001155 \end{bmatrix}$$

$$w^{new} = w^{old} - \alpha \cdot \frac{\partial J}{\partial w} = [0 \ -1] - 0.01 \cdot [0.019 \quad 0.019]$$

$$= [0.00019 \quad -1.00019]$$

(e) Based on your answers to parts (a) and (c), argue that if the logistic model and multinomial model currently provide the same probability that each animal is a cat (so that $a_1^{(i)}$ in the logistic regression model is equal to $a_2^{(i)}$ in the multinomial regression model for all observations i), the updates to b and w in the logistic regression model will be the same as the updates to b_2 and w_2 in the multinomial regression model.

We basically need to show that the relevant components of the gradient vectors are the same.

From part (c) the derivative of J with respect to b in the multinomial regression model is

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}\left(a^{(i)} - y^{(i)}\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\begin{bmatrix} a_1^{(i)} \\ a_2^{(i)} \end{bmatrix} - \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m}(a_1^{(i)} - y_1^{(i)}) \\ \frac{1}{m}\sum_{i=1}^{m}(a_2^{(i)} - y_2^{(i)}) \end{bmatrix}$$

If for every observation i, $a_2^{(i)}$ in this model is the same as $a_1^{(i)}$ in the logistic regression model then this matches the derivative $\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m}(a_1^{(i)} - y^{(i)})$ from the logistic regression model in part (a).

From part (c) the derivative of J with respect to w in the multinomial regression model is

$$\frac{\partial J}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}\left(a^{(i)} - y^{(i)}\right)^T = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}\left(\begin{bmatrix} a_1^{(i)} - y_1^{(i)} & a_2^{(i)} - y_2^{(i)} \end{bmatrix}\right)$$

$$= \begin{bmatrix} \frac{1}{m}\sum_{i=1}^{m} x^{(i)}(a_1^{(i)} - y_1^{(i)}) & \frac{1}{m}\sum_{i=1}^{m} x^{(i)}(a_2^{(i)} - y_2^{(i)}) \end{bmatrix}$$

Again, if $a_2^{(i)}$ in this model is equal to $a^{(i)}$ in the logistic regression model for every $i = 1, \ldots, m$ then the second term in this gradient vector is equal to the expression $\frac{\partial J}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}(a^{(i)} - y^{(i)})$ from the logistic regression model in part (a).