

Ling 572 HW9

Daniel Campos dacampos@uw.edu

03/13/2019

1 Q1

1.1 What does $f'(x)$ intend to measure?

The derivative of a function measures the rate of change of a function relative to the change in the argument (in this case x).

1.2 Let $h(x) = f(g(x))$

$$h'(x) = f'(g(x)) \cdot g'(x)$$

1.3 Let $h(x) = f(x)g(x)$

$$h'(x) = f'(x)g(x) + f(x)g'(x)$$

1.4 Let $f(x) = a^x$ where $a > 0$

$$f'(x) = a^x \log(a)$$

1.5 Let $f(x) = x^{10} - 2x^8 \frac{4}{x^2} + 10$

$$f'(x) = 10x^9 - 16x^7 + \frac{8}{x^3}$$

2 Q2

The logistic function is $f(x) = \frac{1}{1+e^{-x}}$. The tanh function is $g(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

2.1 Prove that $f'(x) = f(x)(1 - f(x))$

$$1. f'(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$2. f'(x) = \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2}$$

$$3. f'(x) = \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right) \text{ which means } f'(x) = f(x)(1 - f(x))$$

2.2 Prove that $g'(x) = 1 - g^2(x)$

$$1. \tanh(x) = \frac{\sinh(x)}{\cosh(x)}$$

$$2. g'(x) = \frac{df}{dx} \frac{\sinh(x)}{\cosh(x)}$$

$$3. g'(x) = \frac{\cosh^2(x) - \sinh^2(x)}{\cosh^2(x)}$$

4. $g'(x) = \frac{\cosh^2(x)}{\cosh^2(x)} - \sinh^2(x)\cosh^2(x)$
5. $g'(x) = 1 - \sinh^2(x)\cosh^2(x)$
6. $g'(x) = 1 - \tanh^2(x)$

2.3 Prove that $g(x) = 2f(2x) - 1$

1. $f(2x) = \frac{1}{1+e^{-2x}}$
2. $2f(2x) = \frac{2}{1+e^{-2x}}$
3. $2f(2x) - 1 = \frac{2}{1+e^{-2x}} - 1$
3. $2f(2x) - 1 = \frac{2}{1+e^{-2x}} - \frac{1+e^{-2x}}{1+e^{-2x}}$
4. $2f(2x) - 1 = \frac{2-e^{-2x}}{1+e^{-2x}}$
5. $2f(2x) - 1 = \frac{(e^x-1)(e^x+1)}{1+e^{2x}}$
6. $2f(2x) - 1 = \frac{e^{2x}-1}{1+e^{2x}}$
7. $2f(2x) - 1 = \frac{e^{2x}-1}{e^{2x}+1} = \tanh(x) = g(x)$

3 Q3

3.1 What is f'_x trying to measure?

A partial derivative is trying to measure the change in a function based on a variable assuming all other variables in the function remain constant. In other words, our derivative is representing the effect of a variable on the equation when no other variables are affecting the equation.

3.2 $f(x, y) = x^3 + 3x^2y + y^3 + 2x.$

$$f'_x = 3x^2 + 6xy + 2$$

$$f'_y = 3(x^2 + y^2)$$

3.3 $z = \sum_{i=1}^n w_i x_i.$

$$\frac{dz}{dw_i} = \sum i = 1^n x_i$$

3.4 $f(z) = \frac{1}{1+e^{-z}}$ and $z = \sum_{i=1}^n w_i x_i.$

$$\frac{df}{dz} = f(z) * \frac{e^{-z}}{1+e^{-z}}$$

$$\frac{df}{dw_i} = f(z) * \frac{e^{-\sum_{i=1}^n x_i}}{1+e^{-\sum_{i=1}^n x_i}}$$

3.5 $E(z) = \frac{1}{2}(t - f(z))^2$, $f(z) = \frac{1}{1+e^{-z}}$ and $z = \sum_{i=1}^n w_i x_i.$

$$\frac{dE}{dw_i} = -(t - f(z)) * f(z) * \frac{e^{-\sum_{i=1}^n x_i}}{1+e^{-\sum_{i=1}^n x_i}}$$

4 Q4 Softmax Function

4.1 Where in NNs is the softmax function used and why?

Softmax is used to normalize the outputs of NN to interval (0,1) and to make all components to add up to 1 so that they can be interpreted as regular probabilities. This tends to be implemented as the final step of a NN to get a probability distribution over all possible classes/predictions.

4.2 x is [1, 2, 3, -1, -4, 0], what is the value of $\text{softmax}(x)$

[0.08607859048507978, 0.23398586833496002, 0.6360395340111326, 0.011649470423906664, 0.0005799929804444501, 0.03166654376447658]

5 Q5: FNN

5.1 How many connections (i.e., weights) are there in this network?

$connections = \sum_{i=1}^{m-1} n_i * n_{i+1}$ since there are $m-1$ layers of connections.

5.2 Given the input x , what is the formula for calculating the output of the first hidden layer?

$$Y_h = g(M_k * X)$$

5.3 Given the input x , what is the formula for calculating the output of the output layer?

$Y_o = g(M_m g(M_{m-1} * \dots * g(M_k * X)))$ where m is total depth of network

6 Q6 MNIST NNs

6.1 What's the loss function used in the digit recognition task?

The loss function is Mean squared error(MSE) or $\frac{1}{2n} \sum x ||y(x) - a||^2$.

6.2 Why do they choose to minimize this function instead of maximizing classification accuracy?

They choose to minimize this since the classification accuracy is not a smooth function that can be connected back to the weights and biases in the network. It is difficult to optimize small changes in weights and biases since they are unlikely to affect changes in training accuracy but will likely affect the MSE.

6.3 In gradient descent, what's the formula for updating the weight matrix (or vector)? Why is that a good formula?

$v \rightarrow v' = v - n * \nabla C$ is a good formula because it is easy to calculate and the gradient vector, ∇C and it gives us a way of repeatedly changing the position of v in order to find a minimum of the function C .

6.4 What are the main ideas and benefits of stochastic gradient descent?

The main idea with stochastic gradient descent (SGD) is to estimate the gradient ∇C on a small sample of randomly chosen training inputs. By doing this average over this sample we can get a good estimate of the true gradient and speed up gradient descent. SGD is useful since it allows an efficient and accurate way of directing a NN to a global minimum is a measurable, stepable function without computing the gradient of the entire dataset.

6.5 What is a training epoch?

A training epoch represents all the mini-batches that needed to be run until the network has exhausted all training input. Once this happens people usually report a current loss function and accuracy and start another epoch.

6.6 Let T be the size of the training data, m be the size of mini-batch, and your training process contains E training epoches. How many times is each weight in the NN updated?

$$\text{updates} = E * \frac{T}{m}$$

6.7 How can one choose the learning rate?

The learning rate needs to be explored during hyperparameter tuning. There are various optimization to choose learning rate such as ADAM but in general many selections of learning rates come from the type of task being modeled and tinkering.

6.8 What's the risk if the rate is too big?

If the learning rate is too big then the network may never learn the function properly since a high learning rate can effectively skip over maximums

6.9 What's the risk if the rate is too small?

If the learning rate is too small then the network will train extremely slowly and it may never leave a local minimum.

7 Q7 MNSIT NN in practice

Table 1: Results on digit recognition

Expt id	# of hidden neurons	epoch #	mini batch size	learning rate	accuracy
1	30	30	10	3.0	0.9461
2	10	30	10	3.0	0.9172
3	30	30	10	0.5	0.9403
4	30	30	10	10	0.9457
5	30	30	100	3.0	0.9302