Name: ____________________

Our problems multiply

2. (10 points) We will consider a neural network with a slightly unusual structure. Let the input x be $d \times 1$ and let the weights be represented as k $1 \times d$ vectors, $W^{(1)}, \ldots, W^{(k)}$. Then the final output is

$$\hat{y} = \prod_{i=1}^{k} \sigma(W^{(i)}x) = \sigma(W^{(1)}x) \times \quad \cdots \quad \times \sigma(W^{(k)}x) \ .$$

Define $a^{(j)} = \sigma(W^{(j)}x)$.

(a) What is $\partial L(\hat{y}, y)/\partial a^{(j)}$ for some j? Since we have not specified the loss function, you can express your answer in terms of $\partial L(\hat{y}, y)/\partial \hat{y}$.

Solution:

$$\frac{\partial L(\hat{y}, y)}{\partial \hat{y}} \prod_{i \neq j} \sigma(W^{(i)}x)$$

(b) What are the dimensions of $\partial a^{(j)}/\partial W^{(j)}$?

Solution: Because $a^{(j)}$ is a scalar, they are the same as for $W^{(j)}$, which is $1 \times d$.

(c) What is $\partial a^{(j)}/\partial W^{(j)}$? (Recall that $d\sigma(v)/dv = \sigma(v)(1 - \sigma(v))$.)

Solution:

$$a^{(j)}(1 - a^{(j)})x^T$$

(d) What would the form of a stochastic gradient descent update rule be for $W^{(j)}$? Express your answer in terms of $\partial L(\hat{y}, y)/\partial a^{(j)}$ and $\partial a^{(j)}/\partial W^{(j)}$. Use η for the step size.

Solution:

$$W^{(j)} = W^{(j)} - \eta \frac{\partial L(\hat{y}, y)}{\partial a^{(j)}} \frac{\partial a^{(j)}}{\partial W^{(j)}}$$