CS 5242
Prof. You's Neural Networks and Deep Learning
August 19, 2022

Homework for Lecture 1

For all gradient calculations, please show the necessary steps of derivation. By default, x means scalar, $\boldsymbol{x}$ means vector [n×1], $\mathbf{X}$ means matrix.

Question 1. Solve for the gradients of the following functions.

1) Sigmoid function

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

2) Softmax

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, 1 \leq i \leq n$$

3) Softplus activation

$$f(x) = \frac{1}{\beta} \cdot \ln(1 + e^{\beta x})$$

Question 2. Solve for the gradients of the following functions and do a shape check (Lecture 1 slide 51, by tracing the shape of intermediate results, make sure that the shape of the final result is correct).

1)

$$f(\boldsymbol{x}) = \boldsymbol{x}^T(\mathbf{A}\boldsymbol{x} + \boldsymbol{z})$$

2) L2 loss

$$L(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{x} - y)^2$$

3) L2 loss (multiple examples)

$$L(\boldsymbol{w}) = \frac{1}{2m}\|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

Question 3. Solve for the gradients $\frac{\partial L}{\partial \mathbf{W}}$ and do a shape check. (Hint: to avoid calculating the gradient of vector-to-matrix, try vectorization introduced in class for $\mathbf{W}$, or calculate the gradient for each position one by one.)

$$\boldsymbol{z} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b}$$
$$L = \|\boldsymbol{z} - \boldsymbol{y}\|^2$$

Question 4. Consider a linear regression without intercept $y = xw, x \in \mathbb{R}, w \in \mathbb{R}$. L2 loss and gradient descent are used. Initial $w = 0$ and learning rate is α. Suppose we only have one example $x = 1, y = 100$ (which is not a setting in reality and we use this toy example for ease of computation).

1) Show how gradient descent works for $\alpha = 0.5, 1.5, 2.5$.
2) Give the condition of α that gradient descent starts oscillating around the optimal position. Give the condition of α that gradient descent can converge. (For the stopping criteria, we can stop when the distance between w and the optimal point is smaller than 1, or any reasonable stop criteria that you propose).
3) Try to prove your statement in 2). (Hint: consider $|100 - w_t|$)