DSA5105: Homework 3

Due: 05/11/2022

[20 points] Consider a 1 layer fully connected neural network given by

$$f(x) = \sum_{i=1}^{n} v_i \sigma(w_i^T x)$$

where $x, w_i \in \mathbb{R}^d$, $v_i \in \mathbb{R}$, and $\sigma(z) = \max(z, 0)$ is the ReLU activation function.

(a) (2 points) We say that a function $g : \mathbb{R}^d \to \mathbb{R}$ is positively homogeneous if $g(\lambda x) = \lambda g(x)$ for all $x \in \mathbb{R}^d$ and $\lambda > 0$.

Show that f is positively homogeneous.

(b) (3 points) Explain why it is not possible to approximate the oracle function $f^*(x) = e^x$ using the neural network f. Is there a way to solve this issue?

(c) (3 points) We train the neural network on some dataset $\mathcal{D} = \{(x_j, y_j), j = 1 \dots N\}$ using the mean squared error. The empirical risk is given by

$$R_{emp}(\theta) = \frac{1}{2N} \sum_{j=1}^{N} (f(x_j) - y_j)^2$$

where $\theta = (w_1, \dots, w_n, v_1, \dots, v_n)$.

Show that

$$\frac{\partial R_{emp}}{\partial w_i} = \frac{v_i}{N} \sum_{j=1}^{N} (f(x_j) - y_j) \sigma'(w_i^T x_j) x_j$$

where σ' is the derivative of σ.

(d) (3 points) We say that the i^{th} neuron is activated for input x when $w_i^T x > 0$. Conclude from the previous question that the gradient $\frac{\partial R_{emp}}{\partial w_i}$ is a weighted average of the datapoints x_j for which the i^{th} neuron is activated (weights can be positive or negative).

(e) (3 points) Conclude that if $w_i^T x_j \leq 0$ for all $j \in \{1, \dots, N\}$, the weight vector w_i will not be updated by the next gradient descent step.

(f) (6 points) In practice, we do not have to worry about the situation where the i^{th} neuron is deactivated for all datapoints x_j. To see why, let us take a look at what happens when the inputs x_j are iid Gaussian variables.
We suppose that the inputs $(x_j)_{1\ leq j \leq N}$ are iid sampled from a multivariate standard Gaussian variable with mean 0 and covariance matrix I. In other words, denoting $x_j = (x_{j,k})_{1 \leq k \leq d}$, we assume that the coordinates $x_{j,k}$ are iid random variables with distribution $\mathcal{N}(0, 1)$ for all i and k.

i (2 points) Given i, show that $w_i^T x_j$ is a Gaussian random variable with mean 0 and variance $\|w_i\|^2$.

ii (4 points) We define the event A by

$$A = \cap_{j=1}^{N} A_j$$

where $A_j = \{w_i^T x_j \leq 0\}$ for all $j \in \{1, \ldots, N\}$.

What is $\mathbb{P}(A)$? Give an interpretation to this result.

End of paper