CIS662 Quiz 4 (11/1/2023)

[Closed-book, open-notes, no communication. Answer each question in the space provided.]

1. Comment on the expected performance of the following variation of the perceptron training algorithm: start with a very small value for the learning rate (e.g., 0.000001), and double the learning rate in each iteration (epoch)?

Learning will be initially very slow, but in later iterations the connection weights will change drastically at each step, so that the algorithm may not converge.

2. Suppose, for a two-class classification problem, the best possible accuracy is 90% using a 1-hidden layer neural network with a specific architecture (e.g., 5-10-1 with sigmoid node functions). What are the reasons why an execution of the backpropagation algorithm may not succeed in reaching 90% accuracy?

Backpropagation attempts to minimize MSE, which may not result in maximizing accuracy.
Also, gradient descent may stop at local optima, and is not guaranteed to result in the best possible solution.

3. If all nodes in a 1-hidden layer neural network computed linear functions (i.e., no sigmoids, no tanh, no ReLUs, etc.), what are the consequences?

Network outputs will be linear combinations of the inputs, hence the network cannot solve a problem requiring nonlinear functions (e.g., the XOR problem for 2-class classification).

4. What is the vanishing gradient problem?

The chain rule may result in computing a value for a gradient (of the loss function with respect to an early layer connection weight) which is almost 0, since many numbers are multiplied together, some of which may be very small. Hence the weight update for such a connection weight is almost 0, so that learning does not succeed.

CIS662 Quiz 4 (11/1/2023)

[Closed-book, open-notes, no communication. Answer each question in the space provided.]

1. In a given 2-class problem, the best possible accuracy obtainable using a perceptron is 90%. Why may the perceptron learning algorithm not result in 90% accuracy?

If the data is not linearly separable, the perceptron learning algorithm does not terminate, and there are no guarantees that the best possible result is obtained. If its execution is stopped at any point, the accuracy value obtained at that point may not be the best possible one, whether or not we have may reached 90% accuracy in a previous step of the algorithm.

2. What is the problem of “exploding gradients”?

The chain rule may result in computing a value for a gradient (of the loss function with respect to an early layer connection weight) which is very large, since many numbers are multiplied together, some of which may be large. Hence the weight update for such a connection weight is very large, so that learning diverges, and the weight values as well as the MSE (or other loss function value) fluctuate widely.

3. Comment on the expected performance of the following variation of backpropagation: start with a very small value for the learning rate (e.g., 0.0001), and increase it by 10% in each iteration (epoch) of backpropagation?

Learning will be initially very slow, but in later iterations the connection weights will change drastically at each step, so that the algorithm may not converge.

4. In backpropagation, why are the changes in the “later” layers of a feedforward network (closer to the output nodes) computed before the changes in the “earlier” layers (closer to the input nodes)?

The derivative of the MSE with respect to an earlier layer weight depend on (and can be computed easily from) the values computed for the derivatives of the MSE with respect to later layer weights.