CS-419m: Quiz 3

Sep 28, 2018. 10:20–10:55 AM

Roll: _____

Name: _____

1. In class we discussed how to express the XOR function as a feed-forward neural network. Likewise, show how you can express the AND function using the smallest number of layers. Does a regular linear function suffice in this case? ..3

2. Design a feedforward network for outputing a y for an input x of one-dimension where y is real-valued and is defined as

$$f(x) = \begin{cases} 0 & : x < -1 \\ x & : -1 \leq x < 1 \\ 1 & : x \geq 1 \end{cases}$$

   Your answer should clearly specify the input, the number of units in each hidden layer, the output layer and the W and b for each hidden layer. Assume that the activation function is RELU. ..4

   First note that $f(x) = 1/2(RELU(x+1) - RELU(x-1))$

   The input layer is just x.

   The first hidden layer will have two units with $W^1 = [1,1]^T, b^1 = [1,-1]^T$

   The second one will just be a linear layer with $W^2 = [1,-1], b^2 = [0]$

3. Let W_{jk} denote the weight connecting jth output of layer $l-1$ to kth unit of layer l in a 2 hidden-layer feed-forward network. Note W_{jk} is the same for both hidden layers. Let the dimension of the input and the number of units in each hidden layer be the same and denote it by m. The last layer is a linear layer trained with square loss. Denote the last layer parameters as W^o_j. The activation unit at each hidden layer is a ReLU. For an example $(\mathbf{x}, y)$, let h^l_k denote the kth output of the l-th layer and o denote the output from the network. Write the expression for the gradient of loss on this example wrt W_{12}. Your answer should be in the full expanded form and not in recursive form like in Backpropagation. [Hint: the space below is sufficient for the correct answer. Do not ramble.] ..4
   $-(y-o)\sum_{j=1}^{m} W^o_{j1}\delta(h^2_j > 0)W_{2j}\delta(h^1_2 > 0)x_1 + -(y-o)W^o_{21}\delta(h^2_1 > 0)h^1_1$

4. Let us assume that your training data is obtained from two distributions for each of the two classes $P(\mathbf{x}|y=1)$ and $P(\mathbf{x}|y=2)$. What would you choose for these distribution so that a standard feed forward neural network will require lots of parameters to classify, but one of the classification methods we studied earlier can provide a good fit with few parameters. ..4 A Gaussian distribution with unequal variance across the two classes. This will give rise to a quadratic decision boundary which to be mimiced via NN will require many layers and hidden units much like for decision trees.

5. State two ill-effect of initializing parameters of a feed forward neural network with all zeros during training. ..3 The symmetries of the hidden layers are not broken. So, all hidden units might take the same parameter value.

   The ReLU will be saturated and gradient information will not flow below the ReLU layer.

Total: 18