EECS545 Lecture 11 Quiz Solutions

1. Assume we have a fully-connected neural network with 1 hidden layer with ReLU activations ($h(a) = \max(0, a)$) for binary classification. Which of the following statements are true about the behavior of the network? (**Choose all options that apply**)

   (a) The total training time is always the fastest for the smallest possible batch size since each gradient step takes less time.

   (b) This model will have a non-linear decision boundary.

   (c) Adding more layer sometimes perform worse than shallow networks.

   (d) Multiplying all the weights and biases in the network by a factor of 10 after training the network will not change its classification accuracy.

   Solution: (b),(c),(d).
   (a): Although each gradient update is faster for a small batch size, the number of updates is larger. Also, the train time varies depending on the dataset and the environment.
   (b): True, as ReLU activation brings nonlinearity.
   (c): True. Deep networks trained with backpropagation (without any sort of unsupervised pretraining) sometimes perform worse than shallow networks due to overfitting (See Slide 67)
   (d): The logit values may change, but the predictions remain the same.

2. For the sigmoid activation function and the ReLU activation function, which of the following are true in general? (**Choose all options that apply**)

   (a) Both activation functions are monotonically non-decreasing

   (b) Both functions have a monotonic first derivative

   (c) Compared to the sigmoid, the ReLU is more computationally expensive

   (d) The first derivative of ReLU is quadratic.

   (e) The first derivative of ReLU is always zero.

   Solution: (a).
   (a) True. Simply graph the activation functions
   (b) False. Sigmoid has non-monotonic derivative $\sigma(x)(1 - \sigma(x))$
   (c) and (d) and (e) False. ReLU is simpler as all positives have derivative 1 and all negatives have 0.

3. (True/False) Logistic regression can be viewed as a single-layer neural network (no hidden layer) without any non-linear activation before applying softmax in the output layer.

Solution: True. Both of them will learn $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$.

4. (True/False) When initializing weights in a fully connected Neural Network, we should set the weight to 0 in order to preserve symmetry across all neurons.

Solution: False. We should not set it to 0 to help hidden units do not get the same gradients from the beginning.

5. (True/False) Any multi-layer neural network with linear activation functions for all hidden layers can be represented as a neural network without any hidden layer.

Solution: True. If linear activation functions are used for all the hidden units, output from hidden units will be written as a linear combination of input features.