Name: ______________________________ Student I.D. #: ______________________

EE 599 Deep Learning, Quiz 1

Wednesday, Feb. 20, 2019

50 minutes

Closed Book; Open notes

Calculators O.K.

No connected devices (laptops, phones, tablets, etc.)

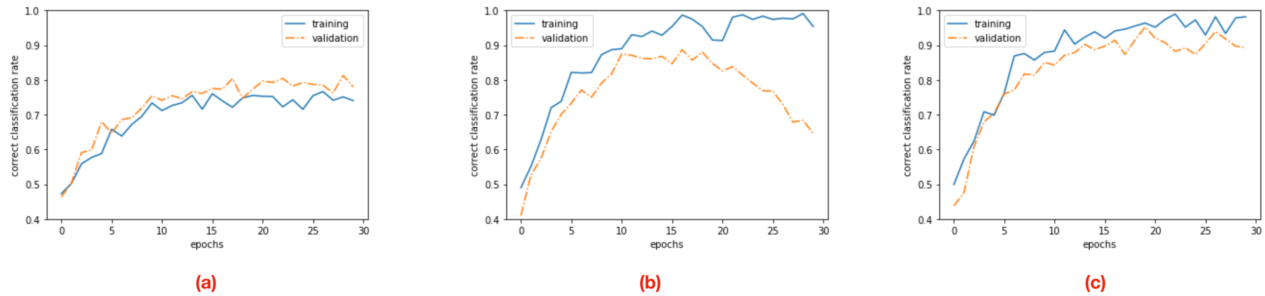

Figure 1: Three training results for the same dataset and ML task.

1. Figure 1 shows three the learning curves for the same dataset and the same binary classification task. Label each as "under-fitting", "over-fitting", or "desired behavior".

   (a)

   (b)

   (c)

2. L1 regularization may be viewed as an a-priori distribution on the weights of the type (circle one):

   (a) Poisson

   (b) Gaussian

   (c) Uniform

   (d) Laplacian

3. A typical split for training/validation/test is (circle one):

   (a) 80/10/10

   (b) 33/33/33

   (c) 20/40/40

   (d) 95/2.5/2.5

4. An MLP has two input nodes, one hidden layer, and two outputs. The activation for the hidden layer is ReLu. The output layer is linear (*i.e.,* identity activation). The two sets of weights and biases are given by

$$\mathbf{W}_1 = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{W}_2 = \begin{bmatrix} 2 & 2 \\ 3 & -3 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 0 \\ -4 \end{bmatrix}$$

What is the output activation when the input is $\mathbf{x} = [\ +1\ \ -1\ ]^{\mathrm{t}}$?

$\mathbf{y}^{(2)} =$

---

5. "Neural Nets are lazy" means (circle one):
   (a) they will favor variance in the bias-variance trade-off
   (b) they train fast because they do not need to find a global minimum of the loss function
   (c) they will find the easiest way to classify data, even if the method is an artifact of the data coverage
   (d) they often will not learn properly, even after many epochs

6. Suppose an MLP has a linear output layer (identity activation) and uses an L1 cost function. If the true label is $\mathbf{y}$ and the final layer activation is $\mathbf{a}^{(L)}$, specify how the back-prop recursion will be initialized.

$\boldsymbol{\delta}^{(L)} =$

---

7. Give the name for each of the activation functions in Figure 2:
   (a)
   (b)
   (c)
   (d)
   (e)

8. The main difference between the PCA and LDA methods for reducing dimensionality is (circle one):
   (a) PCA uses the SVD factorization while LDA is based on a QR decomposition
   (b) PCA is more numerically stable than LDA due to regularization
   (c) PCA does not use the class labels while LDA takes those into account
   (d) LDA is another name for PCA, so there is no difference

9. Circle all of the following techniques that can be viewed as regularization
   (a) Adding an L2 penalty function to the loss

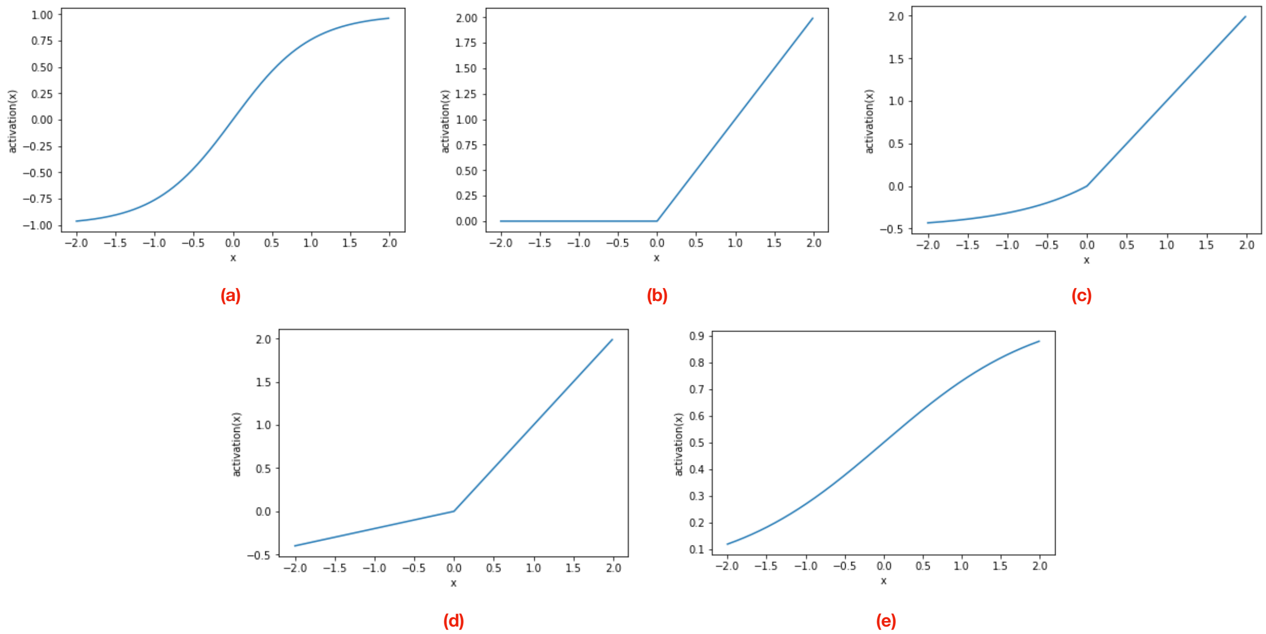

Figure 2: Example activation functions.

(b) Using tanh(·) normalization

(c) Early stopping

(d) Using the adam optimizer

10. When using drop-out, if a node in a neural network has 0.6 probability of drop-out, then this is accounted for in the model used for inference by (circle one):

    (a) using the node 60% of the time in inference

    (b) scaling the weights associated with this node by 0.6

    (c) scaling the weights associated with this node by 1/0.6

    (d) no accounting is needed, you just use the trained network as if drop-out was not applied during training

11. Consider the case when the LMS algorithm is used for on-line linear regression of y_n against x_n with 1 tap – *i.e.*,

$$\hat{y}_n = w_0 x_n$$

Below is data for y_n and x_n, provide the LMS updates to w_0 by filling in the table below. Assume that $x_{-1} = 0$, use $\eta = 0.5$, and take the initial value of w_0 to be 0 as shown.

n :	0	1	2	3	4
w :	0				
x_n :	1	-1	1	-2	-2
y_n :	2	-2	2	-4	-4
$\hat{y}_n$:					