Deep Learning CS583 fall 2021
Final Exam (Section A)

December 14, 2021

Instructor: Jia Xu

Student name: _____

Student ID: _____

Student email address: _____

- **Read these instructions carefully**
- Fill-in your personal info, as indicated above.
- There are ten questions. Each question worths the same (1 point).
- Both computer-typed and hand-writing in the very clear form are accepted.
- This is an open-book test.
- You should work on the exam only by yourself.
- Submit your PDF/Doc/Pages **by 12:00 December 14th** on Canvas under Final exam.

good luck!

1 Question

- After training a neural network, you observe a large gap between the training accuracy (100%) and the test accuracy (42%). Which of the following methods is commonly used to reduce this gap?

  (i) Generative Adversarial Networks

  (ii) Dropout

  (iii) Sigmoid activation

  (iv) RMSprop optimizer

- Consider a Generative Adversarial Network (GAN) which successfully produces images of apples. Which of the following propositions is false?

  (i) The generator aims to learn the distribution of apple images.

  (ii) The discriminator can be used to classify images as apple versus non-apple.

  (iii) After training the GAN, the discriminator loss eventually reaches a constant value.

  (iv) The generator can produce unseen images of apples.

- Applying back-propagation to train a neural network is guaranteed to find the global optimal.

  A. TRUE. B. FALSE.

- Regardless of the choice of the activation function, it makes the network function as a linear mapping from inputs to outputs to set all weights close to zero in a neural network.

  A. TRUE. B. FALSE.

- A multi-layer feedforward network with linear activation functions is more expressive than a single-layer feedforward network with linear activation functions.

  A. TRUE. B. FALSE.

2 Question

- A neuron with 4 inputs has the weight vector $w = [1, 2, 3, 4]^T$ and a bias $\theta = 0$ (zero). The activation function is linear, where the constant of proportionality equals 2 – that is, the activation function is given by $f(net) = 2 \times net$. If the input vector is $x = [4, 8, 5, 6]^T$ then the output of the neuron will be

  A. 1.

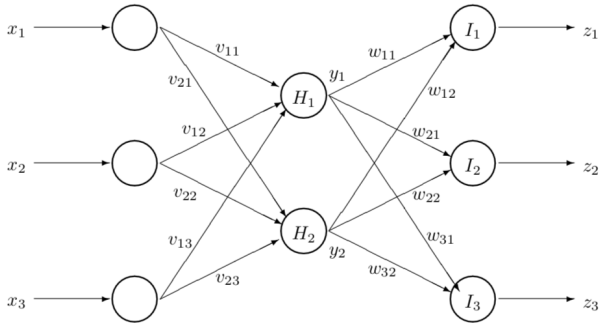
  B. 56.

  C. 59.

  D. 112.

  E. 118.

- A training pattern, consisting of an input vector $x = [x1, x2, x3]^T$ and desired outputs $t = [t1, t2, t3]^T$, is presented to the following neural network. What is the usual sequence of events for training the network using the back-propagation algorithm?



  A. (1) calculate $z_k = f(I_k)$, (2) update W_{kj}, (3) calculate $y_j = f(H_j)$, (4) update v_{jv}.

  B. (1) calculate $y_j = f(H_j)$, (2) update v_{ji}, (3) calculate $z_k = f(I_k)$, (4) update w_{kj}.

  C. (1) calculate $y_j = f(H_j)$, (2) calculate $z_k = f(I_k)$, (3) update v_{ji}, (4) update w_{kj}.

  D. (1) calcualte $y_j = f(H_j)$, (2) calculate $z_k = f(I_k)$, (3) update w_{kj}, (4) update v_{ji}.

3 Question

- The learning rate is a function of the number of training steps t. Why is this function important?

  A. It is not important – the actual value of the learning rate will not affect the performance of the system.

  B. The learning rate is decreased by 1% every learning step so that the learning will "stabilize" and the weights will eventually reach a steady state.

  C. The learning rate depends on the neighborhood of the winning neuron – the neighbors of the winning unit are adapted by a smaller amount so that the network learns a topological mapping.

  D. The learning rate is increased by 1% every learning step so that the robot can improve its performance over time.

- Which of the following statements is the best description of overfitting?

  A. The network becomes "specialized" and learns the training set too well.

  B. The network can predict the correct outputs for test examples that lie outside the range of the training examples.

  C. The network does not contain enough adjustable parameters (e.g., hidden units) to find a good approximation to the unknown function, which generated the training data.

4 Question

- How does splitting a dataset into train, dev and test sets help identify overfitting?
- You want to solve a classification task. You first train your network on 20 samples. Training converges, but the training loss is very high. You then decide to train this network on 10,000 examples. Is your approach to fixing the problem correct? If yes, explain the most likely results of training with 10,000 examples. If not, give a solution to this problem.

5 Question

- A convolutional neural network has an input image of 28×28, and the filter is 3×3. The stride is 1. What is the output dimension of the 12th layer?

- What is the max pooling result of a 2×2 pool size on the below input?

20	55	101	102
32	11	103	104
55	55	10	20
55	55	30	40

- What is the average pooling result of a 2×2 pool size on the above input?

6 Question

- In a deep neural network or a recurrent neural network, we can get vanishing or exploding gradients because the backward pass of back-propagation is linear, even for a network where all hidden units are logistic. Explain in what sense the backward pass is linear.

7 Question

Consider the following linear auto-encoder with 1 input and 1 output: $\tilde{x} = w_2 w_1 x$, trained with the squared reconstruction error:

$L(W) = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{2}(x^i - w_2 w_1 x^i)^2$

The scalar training samples have variance 1.

(a) What is the set of solutions (with 0 loss)?

(b) Does the loss have a saddle point? Where?

8 Question

A neural network has been encrypted on a device. You can access neither its architecture nor the values of its parameters. Is it possible to create an adversarial example to attack this network? Explain why.

9 Question

You want to perform a regression task with the following dataset: $x^{(i)} \in R$ and $y^{(i)} \in R$, $i = 1, \cdots, m$ are the the ith example and output in the dataset, respectively. Denote the prediction for example i by $f(x^{(i)})$. Remember that for a given loss L we minimize the following cost function

$J = \frac{1}{m} \sum_{i=1}^{m} L(f(x^{(i)}), y^{(i)})$.

In this part we are deciding between using loss 1 and loss 2, given by:

$L_1(f(x^{(i)}), y^{(i)}) = |y^{(i)} - f(x^{(i)})|$,

$L_2(f(x^{(i)}), y^{(i)}) = (y^{(i)} - f(x^{(i)})^2$.

(a) Draw $L_1(x, 0)$ and $L_2(x, 0)$ versus $x \in R$ on the same plot.

(b) An outlier is a datapoint which is very different from other datapoints of the same class. Based on your plots, which method do you think works better when there is a large number of outliers in your dataset? Hint: Contributions of outliers to gradient calculations should be as small as possible.

(c) "Using L_1 loss enforces sparsity on the weights of the network." Do you agree with this statement? Why/Why not?

10 Question

You want to perform a classification task. You are hesitant between two choices: Approach A and Approach B. The only difference between these two approaches is the loss function that is minimized.

Assume that $x^{(i)} \in R$ and $y^{(i)} \in \{1, -1\}$, $i = 1, \ldots, m$ are the ith example and output label in the dataset, respectively. $f(x^{(i)})$ denotes the output of the classifier for the ith example. Recall that for a given loss L you minimize the following cost function:

$J = \frac{1}{m} \sum_{i=1}^{m} L(f(x^{(i)}), y^{(i)})$.

As we mentioned, the only difference between approach A and approach B is the choice

$$L_A(f(x^{(i)}), y^{(i)}) = max\{0, 1 - y^{(i)} f(x^{(i)})\}, \tag{1}$$

$$L_B(f(x^{(i)}), y^{(i)}) = \log_2^{(1+exp(-y^{(i)} f(x^{(i)})))}. \tag{2}$$

(i) Rewrite L_B in terms of the sigmoid function.

(ii) You are given an example with $y^{(i)} = -1$. What value of $f(x^{(i)})$ will minimize L_B?

(iii) You are given an example with $y^{(i)} = -1$. What is the greatest value of $f(x^{(i)})$ that will minimize L_A?