University at Buffalo
Department of Compute
and Engineering
School of Engineering and Applied Sciences

Announcements

- PA5 due **today!**
- **Any issues grading on anything except PA5 or the Final must be completed before TOMORROW**
- There's a study guide on Piazza. Don't go beyond it.
- **TODAY: Review Jeopardy!**
    - Same as last time, except more bonus points and fewer questions ☺

Finals Format

- **Next Tuesday, 7:15PM** (expect the exam should take roughly 1.5 hours, max)
- Format
  - 12 MC (2 points each, some harder than others)
  - 4 Short Answer (10 points each)
  - 1 bonus point
  - Grading out of 50
- **Bring nothing but yourself and pen(cil)s and your UBID CARD!!!!!**
- **There will be assigned seating, you will learn your assigned seat on 5/16 (Piazza)**
- If you talk, I will simply pick up your paper and ask you to leave the room
- We will likely have video recordings set up (sorry, I know, I know)
- Work shown where requested, otherwise no credit will be given

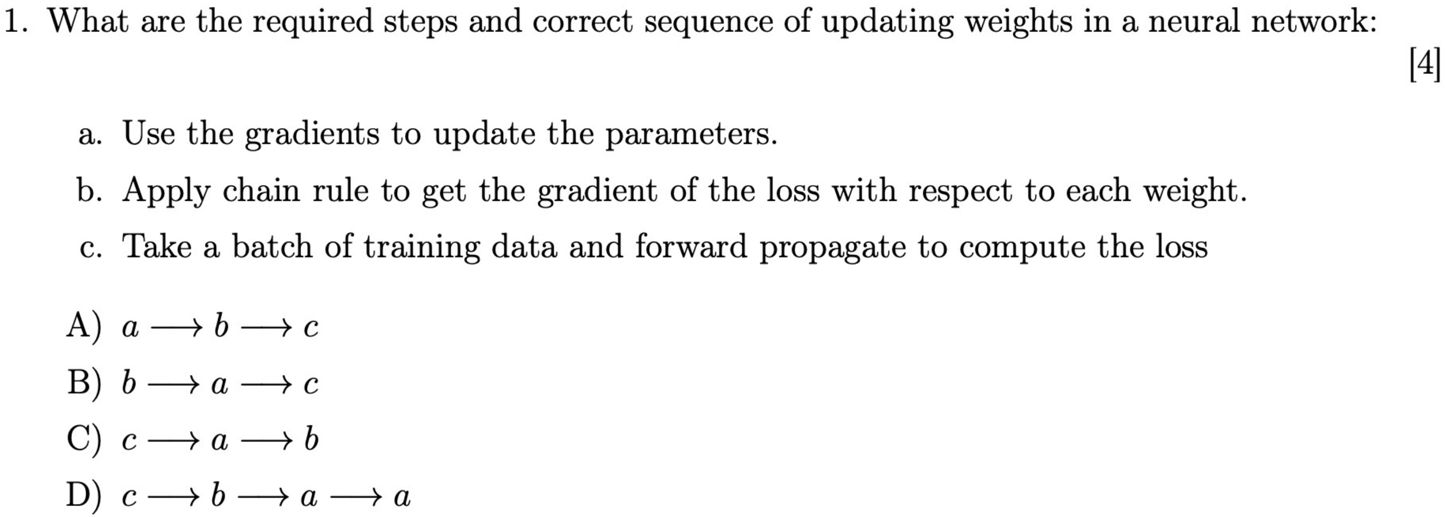
Plan for today

- High fives
- ~~Review Quiz 12~~
- Kahoot Jeopardy
  - **Prizes (points on midterm)**
    - (1) Everyone, as long as at least one team gets every question right
    - (3) Top 5 teams overall (for team members in attendance)
    - (3) Top 3 undergrad teams

First

- Thank you all!
- I hope you learned some stuff
- I learned a lot from you all
- My thoughts
  - Second half smoother than the first
  - Will try for more Kahoot/interaction
  - Probably reorder stuff quite a bit
  - Some PAs better than others (thanks TAs ☺)
- **Do your course eval if so inclined!**

Review: Quiz 12



1. What are the required steps and correct sequence of updating weights in a neural network: [4]

   a. Use the gradients to update the parameters.

   b. Apply chain rule to get the gradient of the loss with respect to each weight.

   c. Take a batch of training data and forward propagate to compute the loss

   A) $a \longrightarrow b \longrightarrow c$

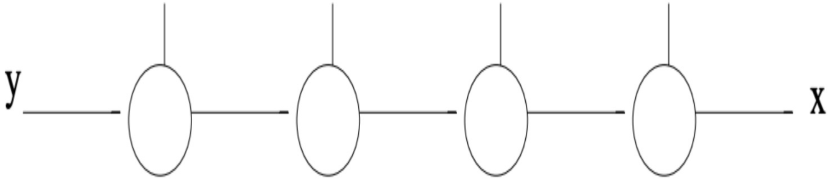
   B) $b \longrightarrow a \longrightarrow c$

   C) $c \longrightarrow a \longrightarrow b$

   D) $c \longrightarrow b \longrightarrow a \longrightarrow a$

Review: Quiz 12

2. Suppose we are trying to train the following chain like neural network with back-propagation. Assume that the activation functions are sigmoid functions and that all the weights are initially set to 1 and all the biases are set to -0.5. As shown in Fig.2, giving such a network an input x = 0.5 causes all the outputs of the units to become 0.5. Now, given a single input x = 0.5 and corresponding target output y = 1, what can you say about the order of magnitude of the gradient updates for weights in this network?



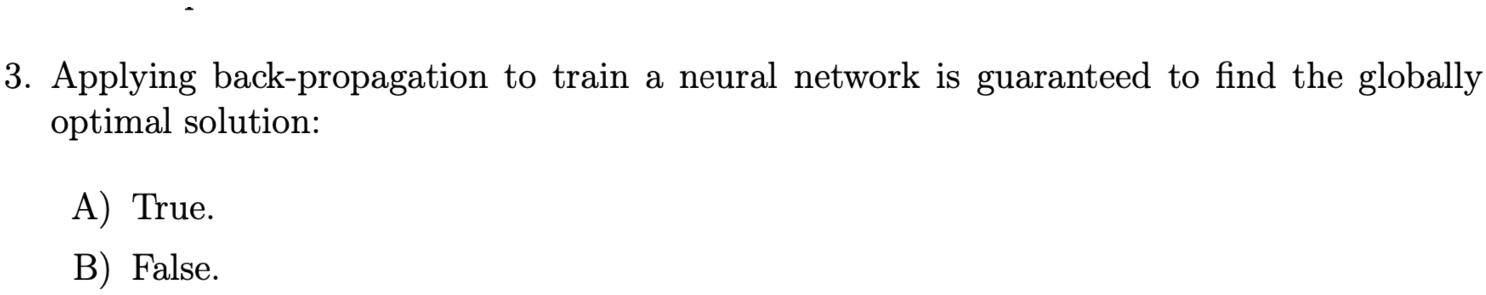
A) The magnitude of the weight updates decreases exponentially as we move away from the output unit.

B) The magnitude of the weight updates increases exponentially as we move away from the output unit.

C) The magnitude of the weight updates increases slowly as we move away from the output unit.

D) The magnitude of the weight updates decreases slowly as we move away from the output unit.

Review: Quiz 12

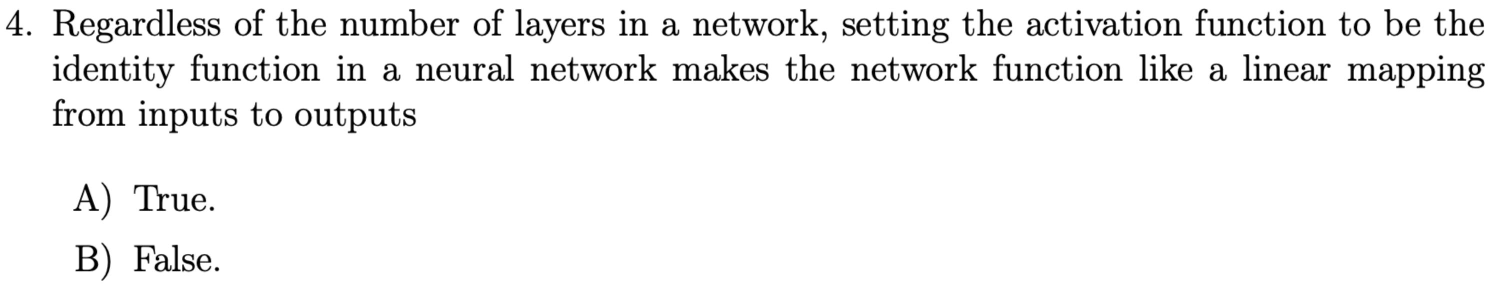


3. Applying back-propagation to train a neural network is guaranteed to find the globally optimal solution:

A) True.

B) False.

Review: Quiz 12



4. Regardless of the number of layers in a network, setting the activation function to be the identity function in a neural network makes the network function like a linear mapping from inputs to outputs

   A) True.

   B) False.

Answer the questions in the spaces provided. You may use the back of each page as scratch space, if you need it.

Name: ____

UBIT name (please print clearly): __________________________

Please indicate which question you are NOT answering by crossing it out in the table below. If you do not, we will grade the first 6 questions.

Please sign below, indicating that you understand that any violations of academic integrity (including, but not limited to, using a phone during the exam, or copying from a neighbor) will result in a grade of zero.

Signature: ____

1. A data scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result. The models should be evaluated based on the following criteria:

    1. Must have a recall rate of at least 80% on the positive class
    2. Must have a false positive rate of 10% or less
    3. Subject to these criteria, we select the one that minimizes business costs

    After creating each binary classification model, the data scientist generates the corresponding confusion matrix. Which confusion matrix represents the model that satisfies the requirements?

    A. TN = 91, FP = 9, FN = 22, TP = 78
    B. TN = 99, FP = 1, FN = 21, TP = 79
    C. TN = 96, FP = 4, FN = 10, TP = 90
    D. TN = 98, FP = 2, FN = 18, TP = 82

Truth

Predicted	+	-
+	TP	FP
-	FN	TN

$$\frac{TP}{TP+FN}$$

given on exam

$$\frac{FP}{FP+TN}$$

2. Which of the following is/are true about weak learners used in ensemble model?

    1. They have low variance and they don't usually overfit
    2. They have high bias, so they cannot solve hard learning problems
    3. They have high variance and they don't usually overfit

        A. 1 and 2

        B. 2 and 3

        C. 1 and 3

        D. None of the above

Suppose, you are working on a binary classification problem. And there are 3 models each with 70% accuracy. The next two questions apply to this setting.

3. If you want to ensemble these models using majority voting method. What will be the maximum accuracy you can get?

   [✓ A.] 100%

   B. 73.38%

   C. 44%

   D. 70%

4. If you want to ensemble these models using majority voting method. What will be the minimum accuracy you can get?

   A. Always greater than 70%

   B. Always greater than and equal to 70%

   [✓ C.] It can be less than 70%

5. Which of the following is **not correct** about the KMeans clustering algorithm?

   [✓ A.] KMeans is a distance based algorithm.

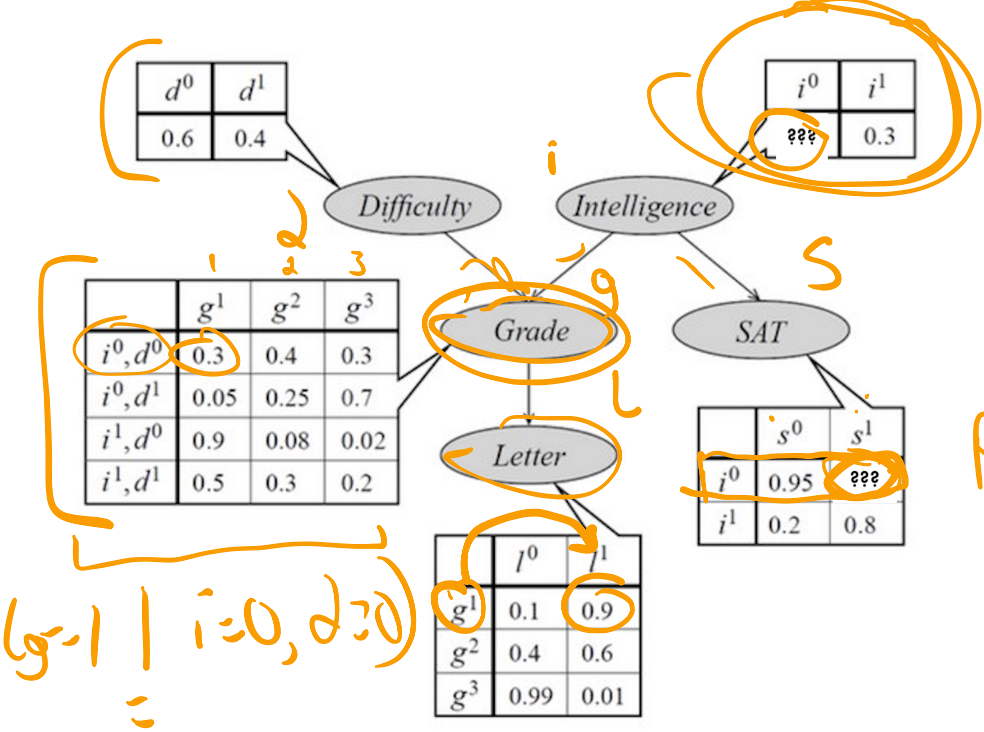
   [✓ B.] KMeans cannot detect overlapping clusters

   C. In higher dimensions distances between data points become more distinguishable and easier for KMeans to detect clusters

   [✓ D.] Relative to Gaussian Mixture Models, Kmeans often performs worse when clusters are irregularly shaped

$p(d, i, g, s, l) = p(l \mid g)\, p(g \mid d, i)\, p(s \mid i)\, p(d)\, p(i)$

Consider the graphical model in Figure 1. Note that the notation d^0 is $p(d = 0)$. On the final, I will make sure to use the notation you have seen in class.



d^0	d^1
0.6	0.4

i^0	i^1
???	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	???
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$p(s=1 \mid i=0)$

$p(g=1 \mid i=0, d=0)$

=

Figure 1: Graphical model for question 12

6. What is $p(i = 0)$ (that is, i^0)?

7. What is $p(s = 1|i = 0)$?

8. What is $p(l = 1|g = 1, d = 0, s = 1, i = 1)$?

$p(g=1 \mid i=1, d=0) \cdot p(i=1) + p(g=1 \mid i=0, d=0) \cdot p(i=0) = \underline{0.48}$

9. What is $\underline{p(g = 1|d = 0)}$?

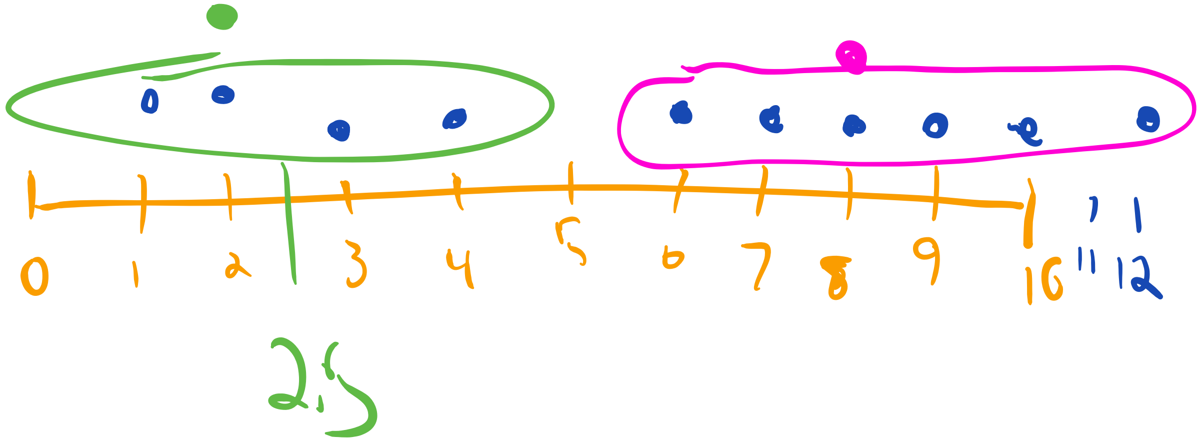
$= (p(g=1, i=1 \mid d=0) +$

$p(g=1, i=0 \mid d=0)$

You are provided with the following 1-dimensional dataset.
Dataset: [3, 8, 6, 2, 1, 4, 12, 9, 7, 10]
Initial centroids:
Cluster 1: K1 = 2
Cluster 2: K2 = 8
Given the initial centroids and following Manhattan distance as a distance metric, cluster the dataset using k-means algorithm on two clusters. After performing the clustering using initial centroids, estimate new centroids for both clusters. Note, the formula for the Manhattan Distance between two points is $MHD(a, b) = \sum_i |a_i - b_i|$.

10. What is the new K1?

11. What is the new K2?



12. Which of the following methods does not avoid overfitting in deep neural networks?

    A. Adding norm penalties to the loss (L1 & L2 regularization)

    B. Using dropout

    C. increasing network capacity

    D. Early Stopping

13. Which of the following statement(s) is/are true about a convolutional layer?

  A. The number of biases is equal to the number of filters.

  [✓ B.] The number of weights is independent of the depth of the input volume. For example, for an Image of size (32,32,3), 3 is the depth.

  C. The number of parameters is independent of the stride

  D. The number of parameters is independent of the padding

14. A student uses a multi-layer perceptron for a classification task and notices that the training error is going down and converges to a local minimum. Then, the student tests on the new data, the test error is abnormally high. What will you recommend the student NOT to do:

    A. The training data size is not large enough. Collect a larger training data and retrain it.

    B. Play with the learning rate and add regularization terms to the objective function.

    C. Use different hyperparameters and train the network several times. Use the average of predictions from all nets to predict test data.

    [✓ D. Use the same training data but add two more hidden layers.]

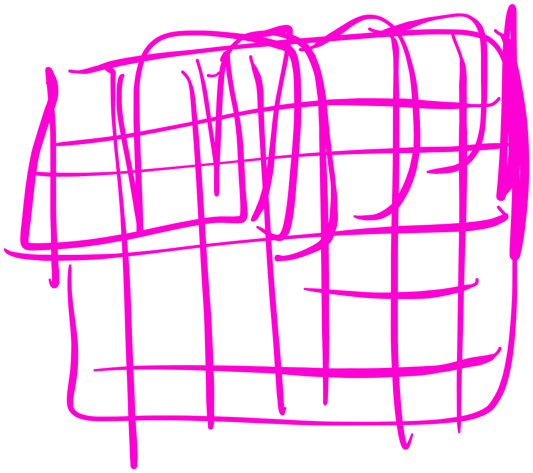
15. When we build a CNN model, the input image is given to us as a matrix of size 32 × 32. We then apply a kernel/filter of size 8 × 8 with a stride of 1 and zero padding. What will be the size of the convoluted matrix?

   A. 7 × 7

   B. 24 × 24

   C. 25 × 25

   D. 8 × 8



Convolution layer: summary

Let's assume input is $W_1 \times H_1 \times C$

Conv layer needs 4 hyperparameters:

- Number of filters **K**
- The filter size **F**
- The stride **S**
- The zero padding **P**

This will produce an output of $W_2 \times H_2 \times K$

where:

- $W_2 = (W_1 - F + 2P)/S + 1$
- $H_2 = (H_1 - F + 2P)/S + 1$

Number of parameters: F^2CK and K biases

Common settings:

K = (powers of 2, e.g. 32, 64, 128, 512)

- F = 3, S = 1, P = 1
- F = 5, S = 1, P = 2
- F = 5, S = 2, P = ? (whatever fits)
- F = 1, S = 1, P = 0

16. Consider the following sample data set:

x_1	x_2	x_3	y
1	1	1	+1
0	1	0	-1
1	0	1	-1
0	0	1	+1

x_1, x_2, x_3 represent features, whereas y represents the label. If we were to train a decision tree on this data set, what feature will we split on at the root?

HINT: You may use the naive approach and select feature that has the lowest training classification error at the root.

A. x_1

B. x_2

C. x_3

D. Non

17. Which of the following statements are true about PCA, SVD, and UMAP (note: UMAP is NOT in the study guide, but we can go beyond that here :))?

    A. PCA is computationally more feasible for large, sparse matrices

    [✓ B.] PCA and SVD are both tools that can be used to perform dimensionality reduction

    C. UMAP and PCA are non-linear approaches to dimensionality reduction

    D. The first dimension of both UMAP and PCA are guaranteed to be the direction of maximal variance in the data

Center +
standardize
non zero
dense
matrix

18. Please judge following descriptions as true or false:

    1. Because decision trees learn to classify discrete-valued outputs instead of real-valued functions it is impossible for them to overfit
    2. In CNN, having max pooling always decrease the parameters
    3. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

    A. False, False, True

    B. True, False, True

    C. True, True, False

    D. False, True, False