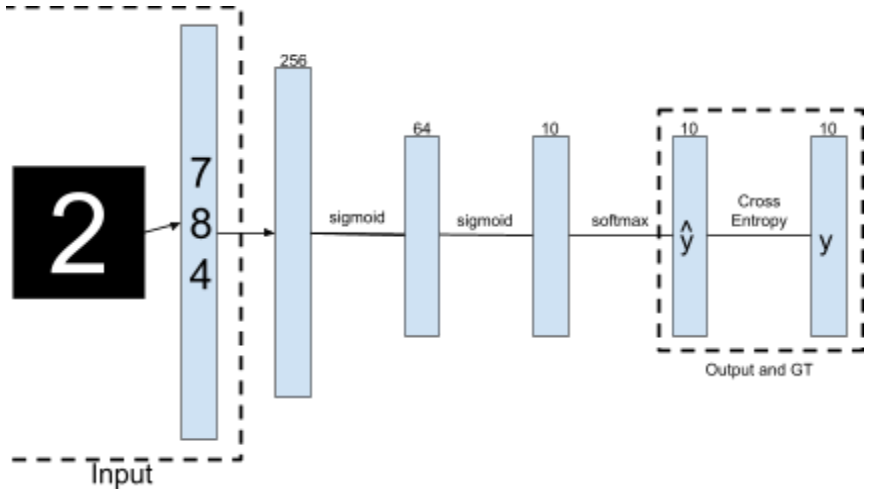
Name:

NetID:

1.) What is grid search and what do we use it for?

2.) Given the network diagram, answer the following questions.



For reference:

The derivative of the sigmoid is: σ(x)(1 - σ(x))
The derivative of Cross-Entropy Loss and Softmax is: ŷ - y

a.) Draw the computation graph for the network. (5 pts.)

b.) Give the derivative chain for calculating the gradient of the bias in the first layer.

$$\frac{\partial \mathcal{L}}{\partial b} =$$

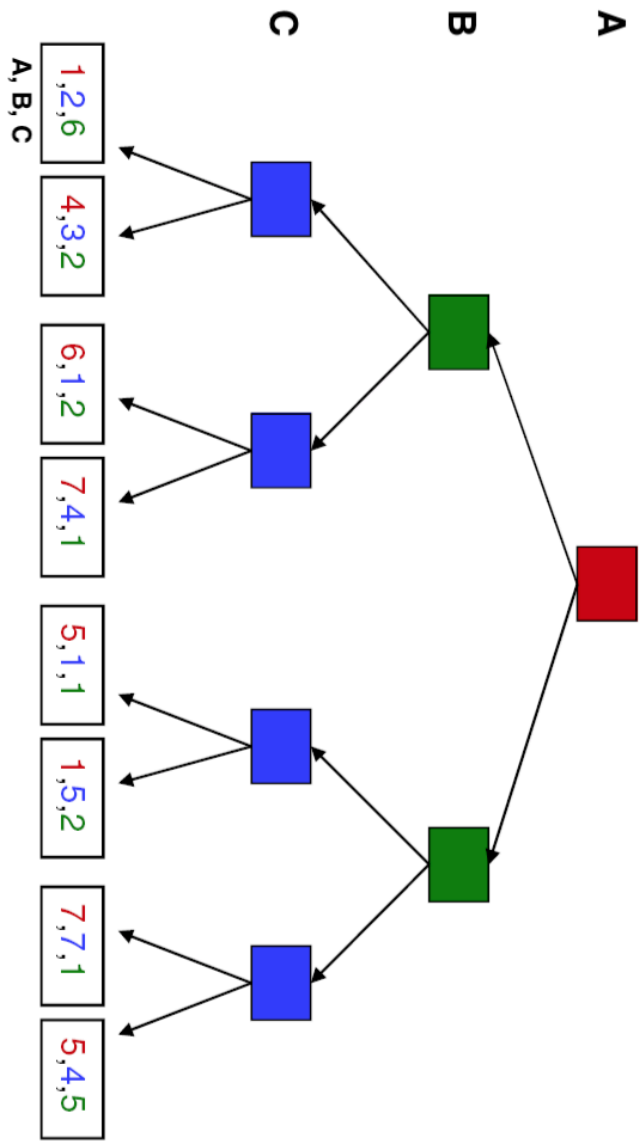
c.) Give the equation you'd use to actually calculate the gradients.

3.) What type of data are RNNs/LSTMs/GRUs used for?

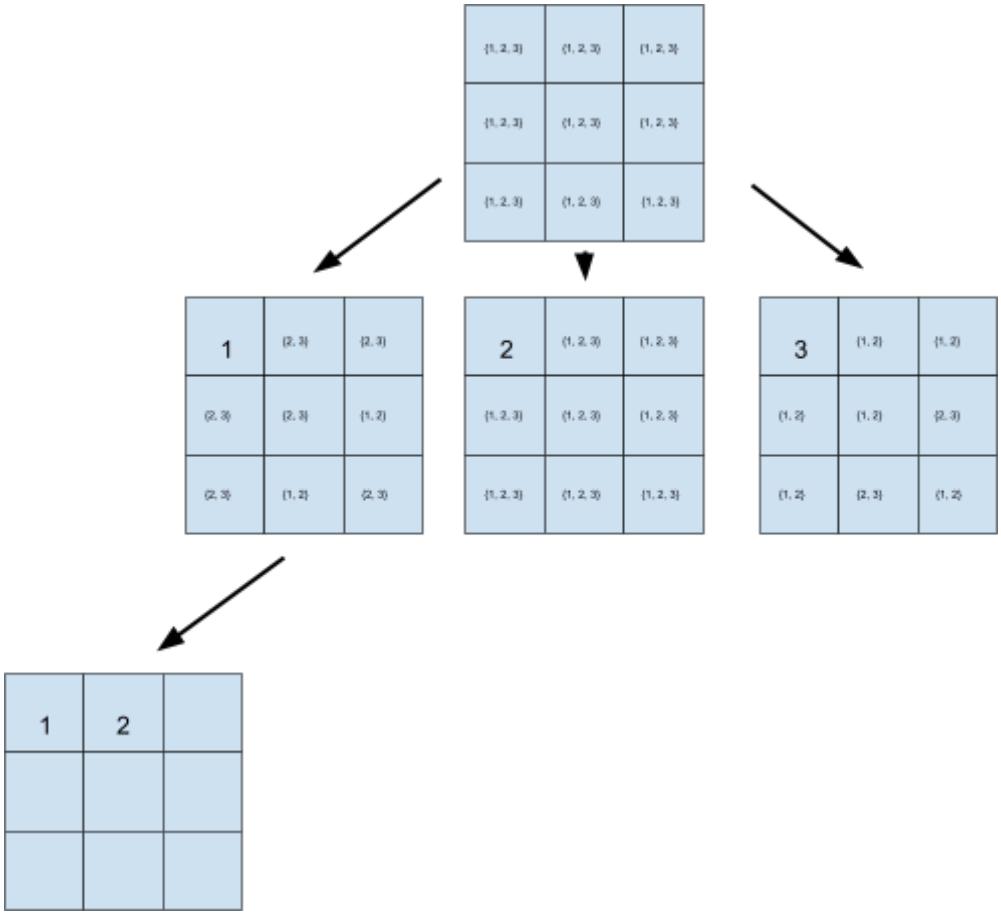
4.) Write the “equations” for a fully connected layer and an RNN. Explain the reasoning behind the difference between the two.

5.) What are vanishing and exploding gradients? Why do they occur?

6.) Given the following tree, perform the minimax algorithm for a three-player game. Assume the tuples are ordered in a top-down oriented fashion relative to the players. Show the expected tuple at each empty node of the tree



7.) Constraint propagation is when, after making a choice of a value to assign to a variable in a CSP, the newly discovered constraints are propagated forward to future states and their domains. Given below is a problem in which each row, each column, and the upper left to lower right diagonal all need to sum to 6. Each square can have a single value in it from the domain of {1, 2, 3}. Shown below are the first four states in the DFS search with constraint propagation applied. Fill in the missing state.



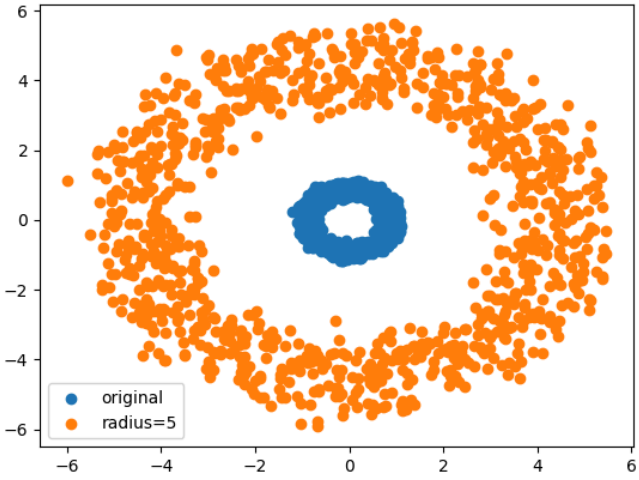
8.) Why would we typically prefer to use DFS rather than BFS for constraint satisfaction problems and backtracking?

9.) Why is it important to separate our dataset into training and testing data?

10.) Why is the Naive Bayes classifier called “Naive”? What is the big assumption we make while using it?

11.) In your own words, what is the difference between agglomerative hierarchical clustering and divisive hierarchical clustering?

12.) Which of the three SVM kernels we've seen so far would be best for separating the data points shown in the plot below?



A.) RBF

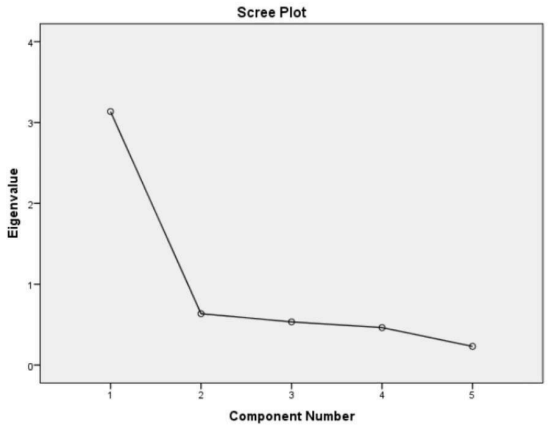
B.) Linear

C.) Polynomial

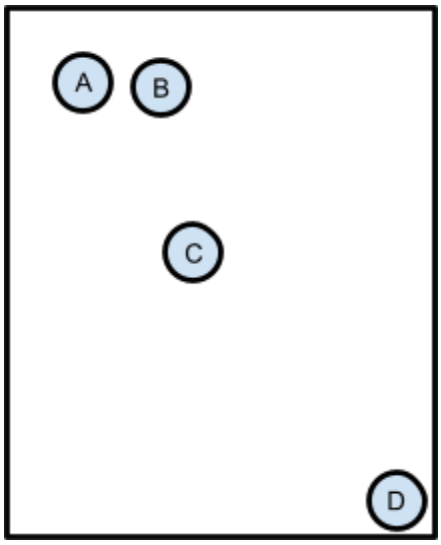
D.) Lagrangian

13.) Is a regression task supervised or unsupervised learning? Explain your reasoning.

14.) When we use PCA, we often get more principal components back than we'd realistically use. Given the scree plot below, indicate how many principal components you'd choose to use. Why don't we use every component?



15.) Create the dendrogram that hierarchical clustering would produce for the following plot.



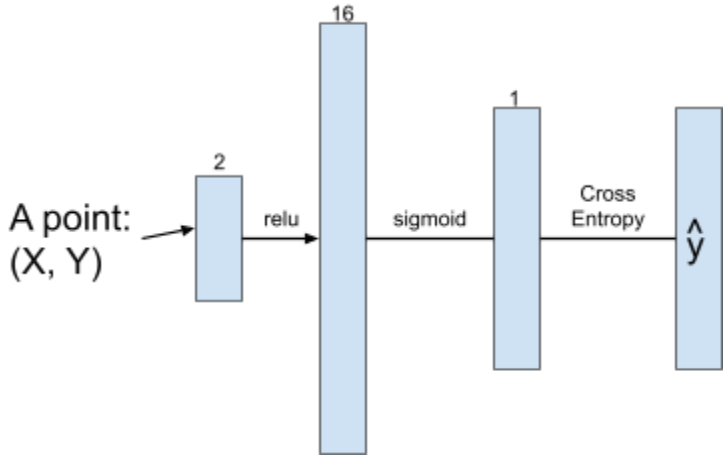
16.) What's an example of a task where a ROC curve is used? When would we prefer AUC?

17.) Explain the difference between Precision, Recall, and the F1 score. Why might we prefer one to any of the others?

18.) What is the difference between gradient descent and backpropagation?

19.) Explain why are sigmoids particularly prone to vanishing gradients, especially in the case of RNNs

20.) Given the network diagram, answer the following questions.



For reference:

The derivative of the sigmoid is: $\sigma(x)(1 - \sigma(x))$
The derivative of Cross-Entropy Loss is: $-((Y / \hat{y}) - ((1 - Y) / (1 - \hat{y})))$

a.) Draw the computation graph for the network.

b.) Give the derivative chain for calculating the gradient of the bias in the hidden layer. (2 pts.)

$$\frac{\partial \mathcal{L}}{\partial b} =$$

c.) Give the equation you'd use to actually calculate the gradients. (5 pts.)

21.) What's an example of sequential data and how do RNNs capture the context?

22.) Why do we call them **recurrent** neural networks? What makes them "recurrent"?

23.) Explain what Q, K, and V are in relation to self-attention. What makes the self-attention mechanism better than RNNs/LSTMs/GRUs?

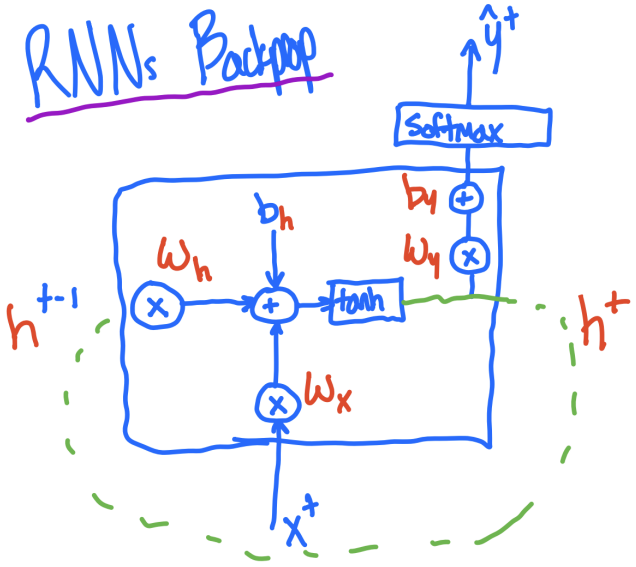
24.) What is the bias/variance tradeoff?

25.) How does k-Nearest Neighbors classify points when used for classification?

26.) What is the purpose of the Gaussian distribution in a Naive Bayes classifier?

27.) What is polynomial feature expansion?

28.) Given the network diagram, draw the computation graph for the network for the sentence “Good Luck” with a word level tokenizer, only producing output for the final token.



29.) What are tokens and embeddings in relation to NLP tasks?

30.) Convolutional Networks use filters or kernels to process images. What do these look like and what purpose do they serve?

31.) What are the two types of pooling layers? Why do we use them?

32.) What is the difference between Djikstra’s and A*?

33.) We say Attention(Q, K, V) is a “dictionary lookup”, explain this metaphor.

34.) What is layer normalization and why do we use it in a transformer?

35.) How are minimax and alpha/beta pruning related? Do we still have to use minimax if we use alpha/beta pruning?

36.) What are the differences between gradient descent, stochastic gradient descent, and mini-batch stochastic gradient descent?

37.) What is a linear transformation and how does it relate to deep learning?

38.) How are softmax and a sigmoid related? What purpose do they serve in a neural network and where do they often appear?

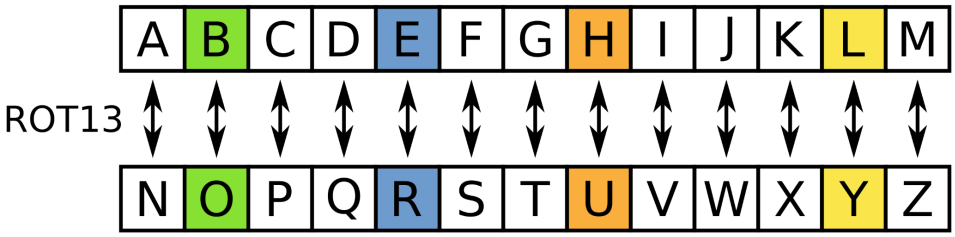
39.) What are L1 and L2 regularization?

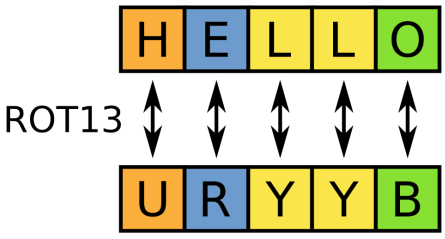
40.) How do we decide where to split a decision tree?

41.) Describe a situation in which each of the following three models may be preferred to the other two: Decision Trees, Logistic Regression, SVMs

42.) One homework05, did it really make sense to use a sequence model like an RNN or transformer for a ROT13 cipher? Why or why not?

Note: ROT13 is a simple letter substitution cipher that replaces a letter with the 13th letter after it in the Latin alphabet.





43.) For homework05 you had to implement Backprop Through Time (bptt), why does the RNN require this?

44.) On homework04 your task was optical character recognition (OCR), what is OCR?

45.) One homework05, our bag-of-words (BoW) comparison had 0% for everything, why is this?

46.) What did we use a Hidden Markov Model (HMM) for on homework02?

47.) Why might we prefer a clustering technique for image segmentation over an SVM as we saw in homework03?