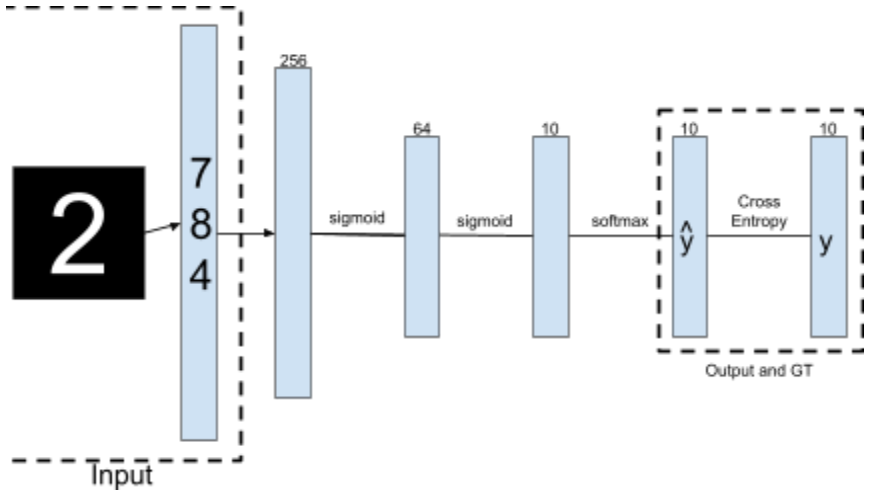
Name:

NetID:

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

We use grid search to test different combinations of hyperparameters to find the best set

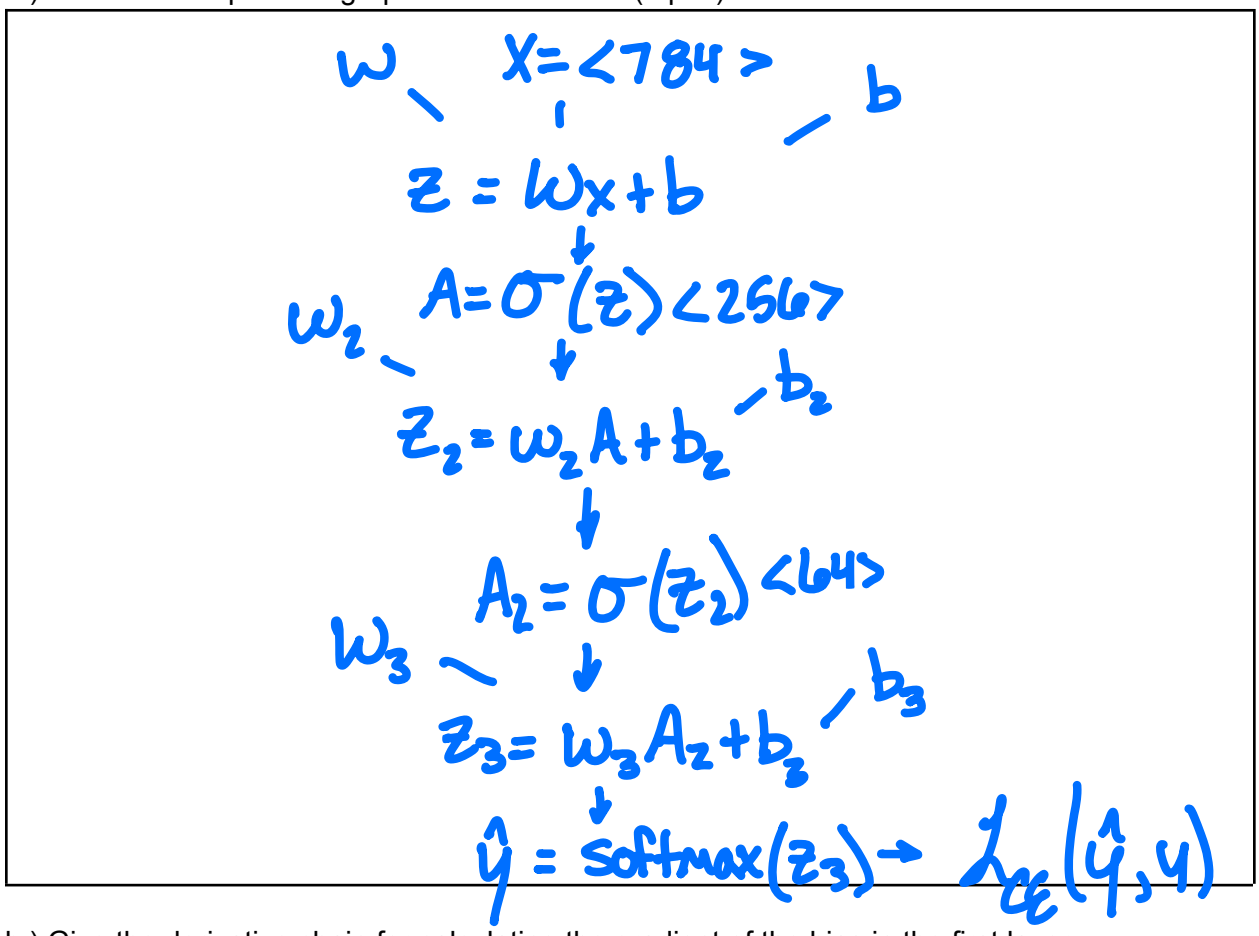
2.) Given the network diagram, answer the following questions.



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.

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_3} \frac{\partial z_3}{\partial A_2} \frac{\partial A_2}{\partial z_2} \frac{\partial z_2}{\partial A} \frac{\partial A}{\partial z} \frac{\partial z}{\partial b}$$

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

$$(\hat{y}-y)(w_3)(\sigma(z_2)(1-\sigma(z_2)))(w_2)(\sigma(z)(1-\sigma(z)))(1)$$

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

"Sequence data" like text or weather every day or stock prices

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

Linear:

$z = wX + b$

RNN:

$h = w_x X + w_h h_{t-1} + b$

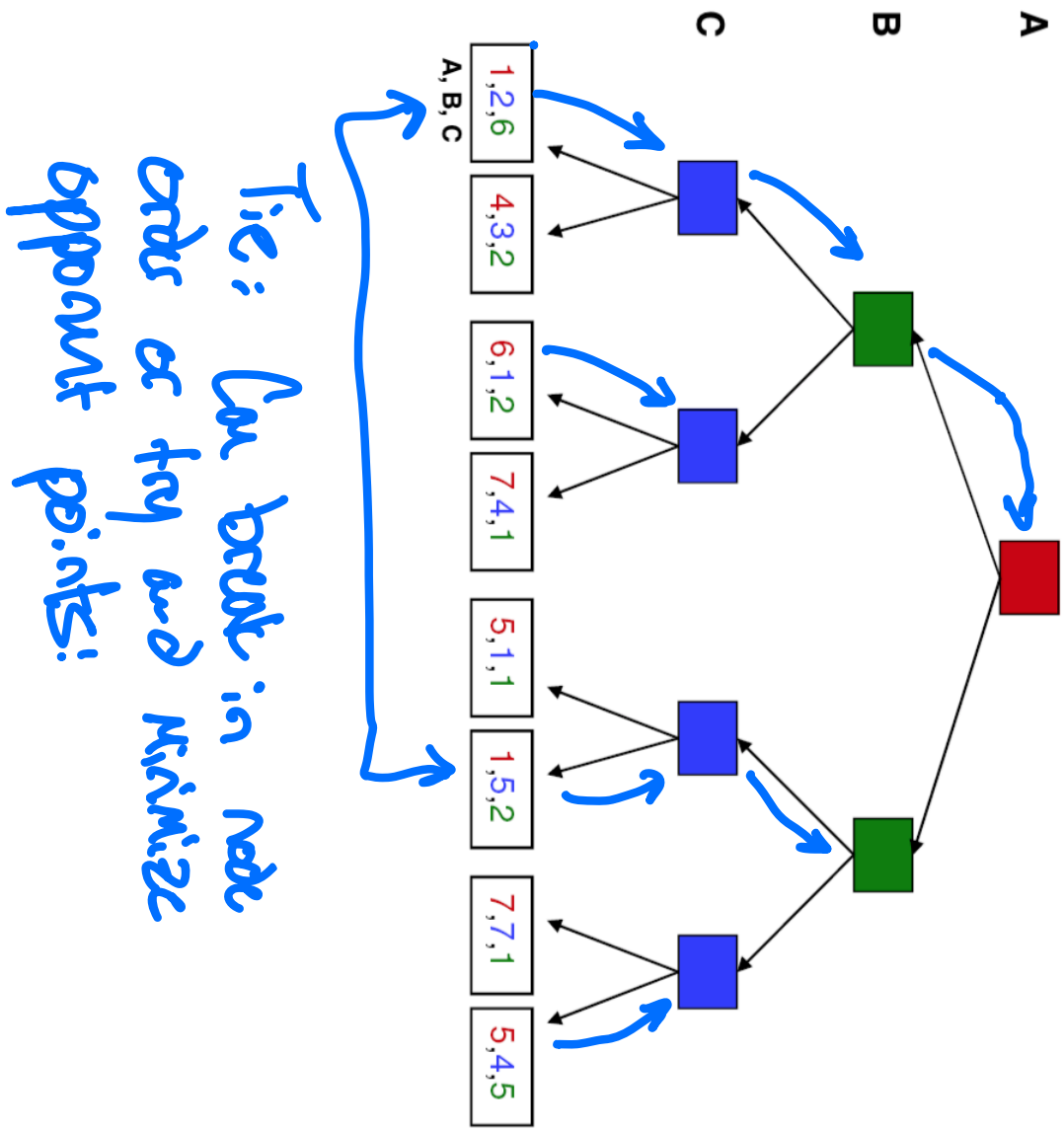
Extra term allows us to capture and maintain "context" over a sequence

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

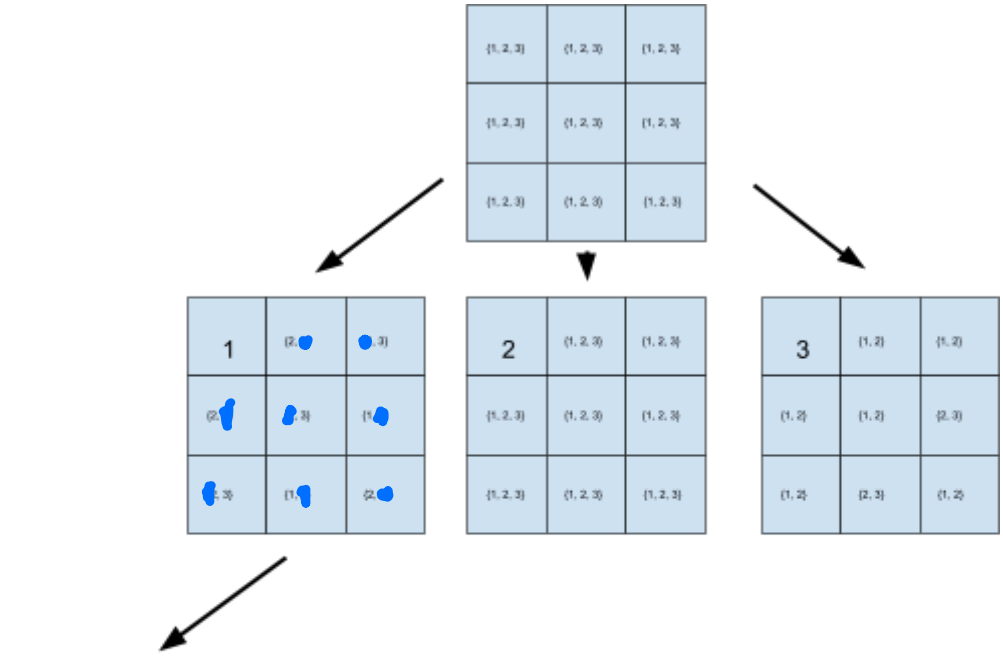
Vanishing → gradients to 0
Exploding → gradients to ∞

Happen when we backprop too many times in one pass multiplying too large or small numbers repeatedly

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.



1	2	3	Σ=6
2	3	1	=6
3	1	2	=6
=6	=6	=6	

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

We get down into the solutions (leafs) faster and can more naturally backtrack (recursion with stack)

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

So our Model doesn't "see" the answers, we want an accurate pic of how our model will perform in the real world with unseen data

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

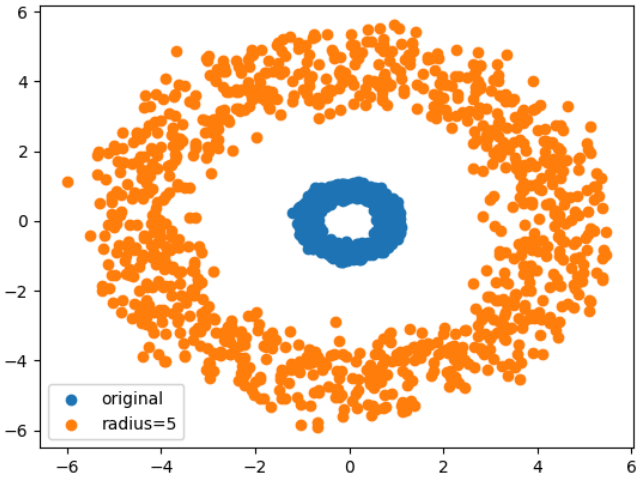
It assumes that features aren't related to each other

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

Agglomerative - bottom up

Divisive - top down

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

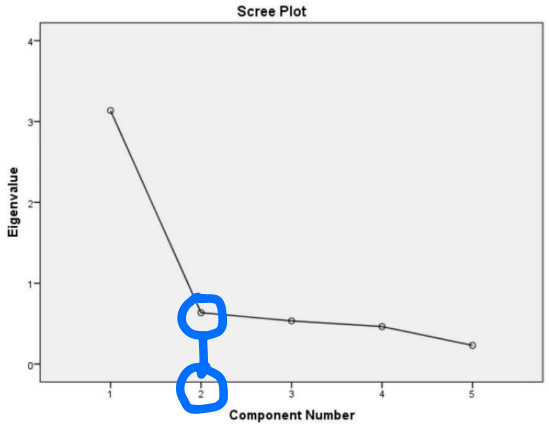
C.) Polynomial

D.) Lagrangian

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

Supervised - we need ground truth numbers to fit our model to

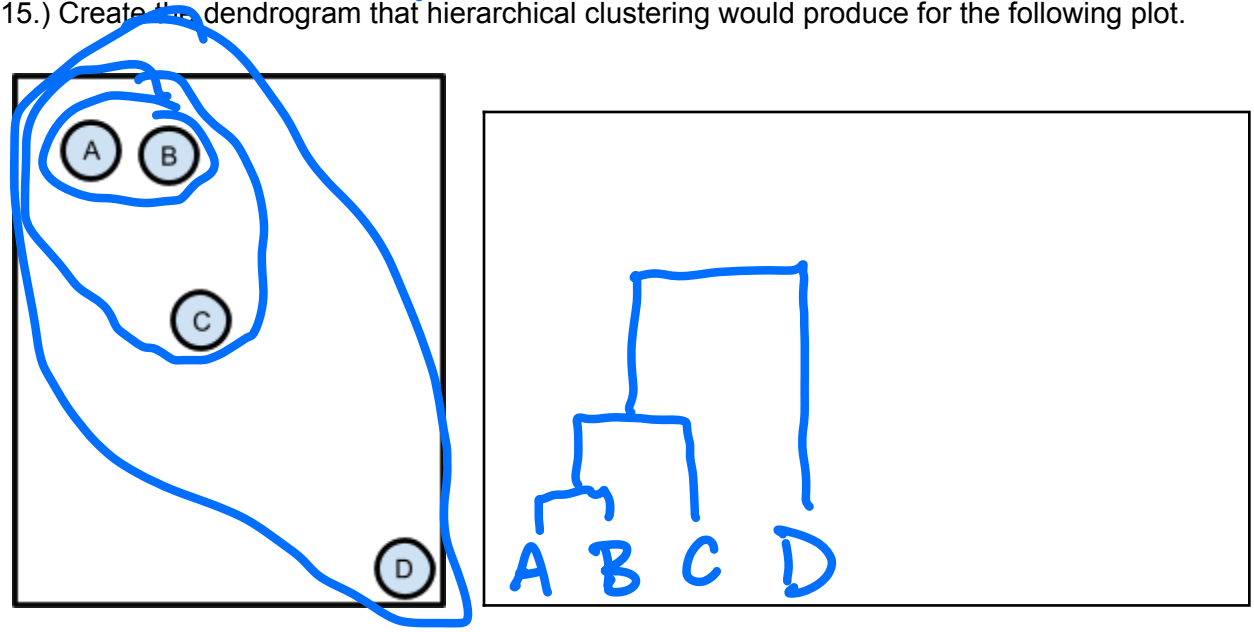
14.) When we use PCA, we often get more principal components back than we'd realistically use. Given the scree plot below, indicate how many principal components you'd choose to use. Why don't we use every component?



Elbow of plot at 2, adding more than this only explains like 0.5 variance so we just don't need it

agglomerative

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?

Roc- comparing TP and FP rates of multiple classifiers at different combos

AUC- we want a single summary measure of our classifiers

↑ I think I didn't end up covering these this semester but still worth knowing (not on exam)

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

Precision - false positives are costly
Recall - false negatives are costly
F1 - balances precision & recall and good for class imbalances

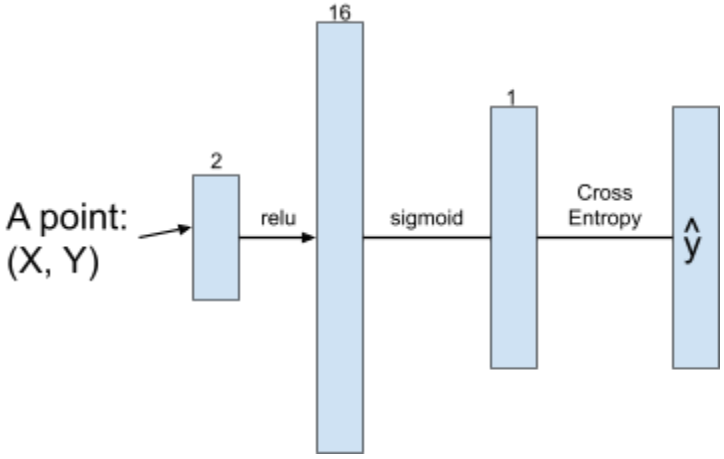
18.) What is the difference between gradient descent and backpropagation?

Gradient descent uses the gradients found via backprop to update the parameters

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

Their graph is ∩ between 0 and 0.25 on y so everything is small

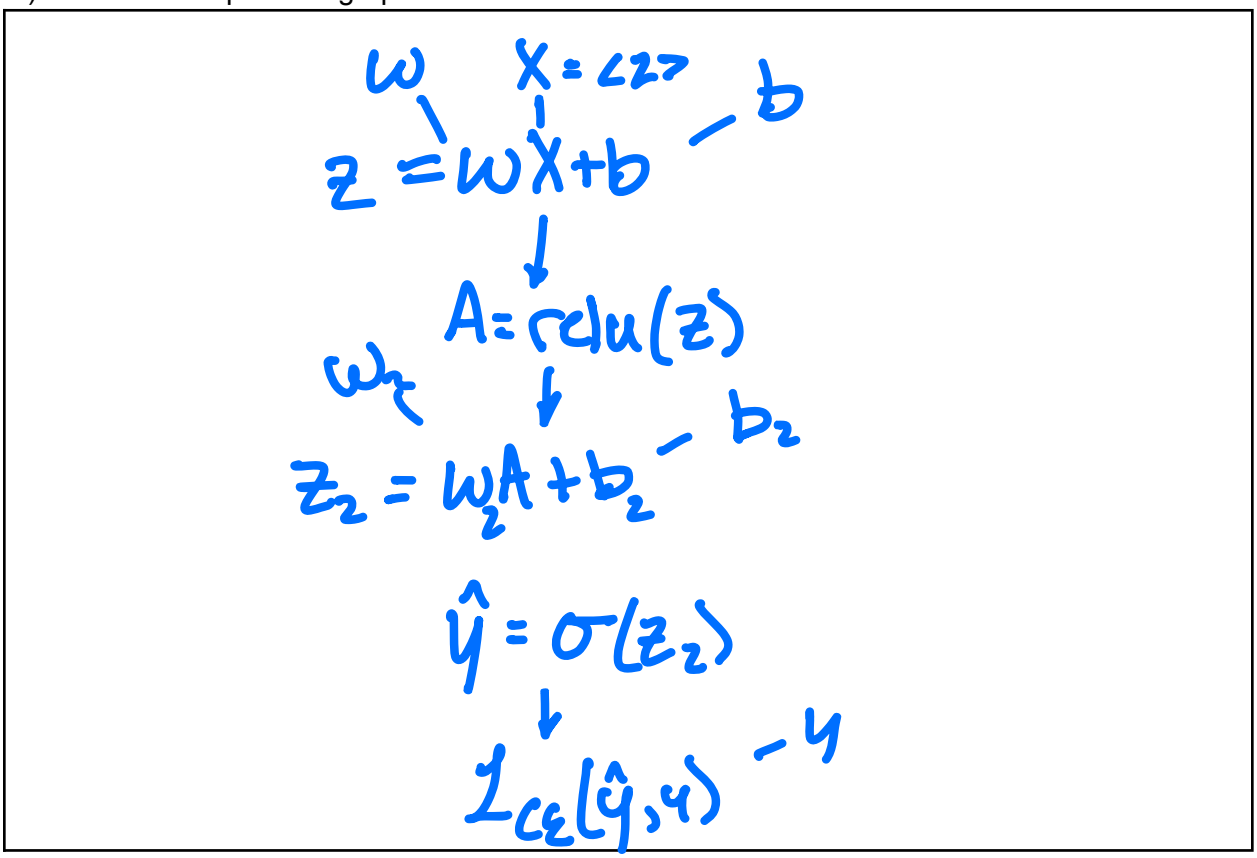
20.) Given the network diagram, answer the following questions.



For reference:

The derivative of the sigmoid is: σ(x)(1 - σ(x))
The derivative of Cross-Entropy Loss is: -((Y / ŷ) - ((1 - Y) / (1 - ŷ)))

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.)

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_2} \frac{\partial z_2}{\partial A} \frac{\partial A}{\partial z} \frac{\partial z}{\partial b}$$

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

$$(\hat{y} - y)(w_2)\begin{pmatrix} x > 0 : 1 \\ x \leq 0 : 0 \end{pmatrix}(1)$$

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

"I like cats"

An RNN would loop over each word (or token) in the input building up a "context" matrix to carry info forward

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

They loop over over token in the input sequence for the forward and backward passes

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

Q - Query - "what each token is looking for"

K - Key - "what each token is"

V - Value - "Context/info of each token"

Self-attention is computed pair-wise so it's way more efficient (fewer loops)

24.) What is the bias/variance tradeoff?

High Bias - model underfits but generalizes better

High Variance - overfit model that isn't robust

We can't have a perfect model so we trade off between these

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

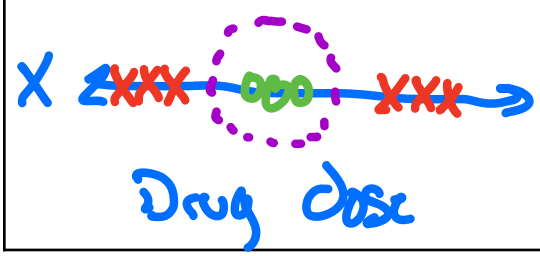
finds the K nearest points to ours and votes using their classes

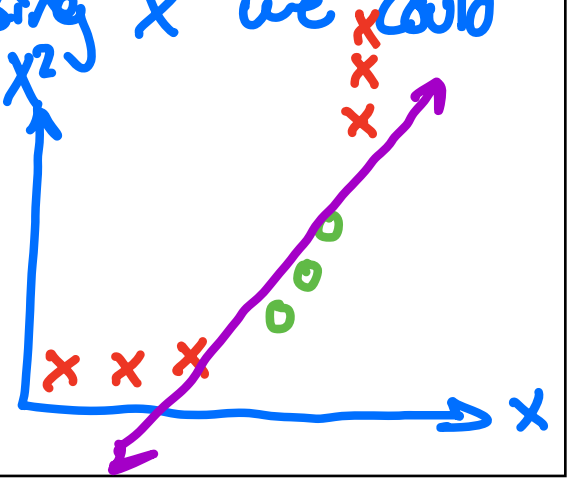
26.) What is the purpose of the Gaussian distribution in a Naive Bayes classifier?

It's how we compute the likelihood of a feature value belonging to a class. We assume feature values follow a normal distribution

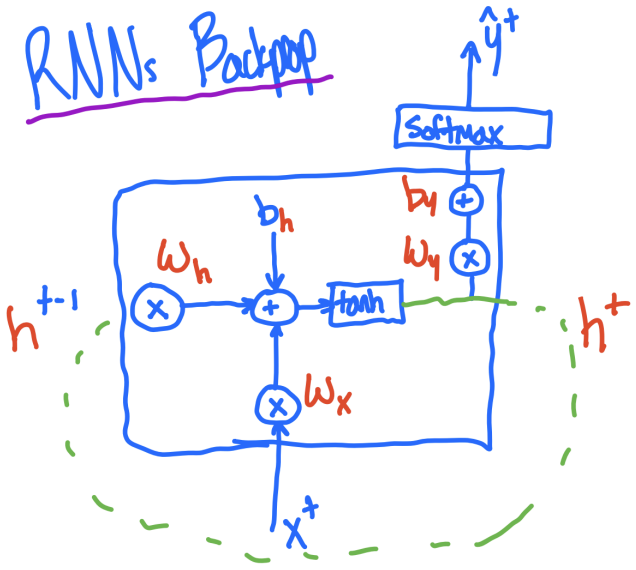
27.) What is polynomial feature expansion?

Instead of just using X we could use X and X^2





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.



Good Luck
x_1 x_2

$z = W_x x_1 + W_h h_0 + b_h$

$h_1 = \tanh(z)$

$z_2 = W_x x_2 + W_h h_1 + b_h$

$h_2 = \tanh(z_2)$

$z_y = W_y h_2 + b_y$

$\hat{y} = \text{Softmax}(z_y)$

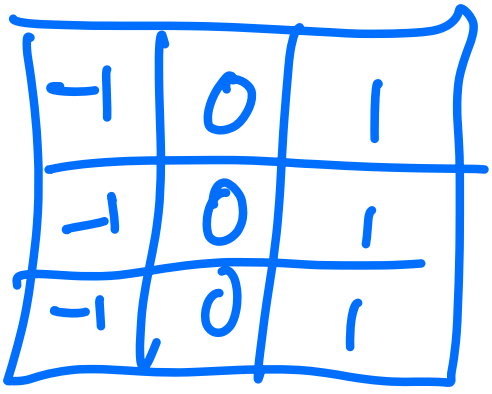
29.) What are tokens and embeddings in relation to NLP tasks?

tokens are mapping input words to numbers and then embeddings are converting these numbers to a fixed dim vector to input to our model

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

Small grids of pixels we use to transform our image to find patterns in it

finds vertical lines →



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

Max: take the highest value in the region

Avg: take the avg value in the region

Used to downsample or blur an image

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

A* is informed and uses the extra heuristic info to ideally find the optimal path faster

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

Query, Key, Value

We do a pairwise comparison of every Q to every K and then using Softmax to "weight" the matches we use the V to modify the context of each token so it's like using Q to look in a {K:V} dict

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

Ensures that features have a consistent distribution, we use it to ensure features don't grow overly large or small as they pass through all the layers

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

α/β pruning runs on top of minimax and may let us ignore some branches but we still use minimax with α/β

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

gd: one sample at a time in a row

Sgd: one random sample at a time

Mbsgd: one batch of random samples at a time

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

It can transform a point from one space to another I.E.: 3D to 2D
We use this in DL by learning the weights used in the transformation to magically transform input into answer space

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

Softmax and sigmoid transform number(s) into "probabilities". We typically see this at the end of a NN to get answer probabilities

39.) What are L1 and L2 regularization?

L_1 → Lasso → leads to sparse weights

L_2 → Ridge → smaller, evenly distributed weights

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

We can use information gain or gini impurity to decide which split seperates the data the "most"

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

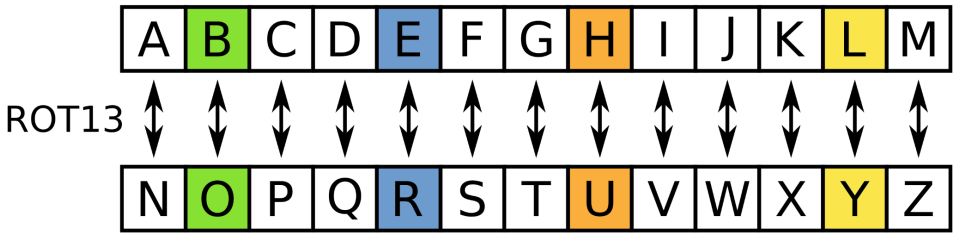
Decision Tree → explainable flow chart Model is needed

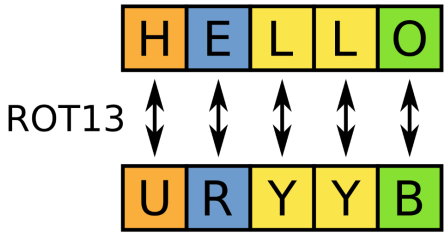
Logistic Regression → simple binary Classification task

SVMs - non-linear, small dataset

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.





Not really because the encoding of the current letter doesn't depend on previous letters in any way nor has any bearing on the encoding of following letters

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

Because we want to build up our context over tokens so the final context relies on every prior one so we need to go backwards

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

Optical Character Recognition -> getting text out of a picture

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

Because th corpuses we were comparing had no shared words between them

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

Using incomplete gps + witness data to calculate th most probable path a suspect took

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

We don't need groundtruth od it's much faster