CSE 446/546 Winter 2025 Final Exam

March 19, 2025

Name _____

UW NetID (not the numbers) _____

Please **wait** to open the exam until you are instructed to begin, and please take out your Husky Card and have it accessible when you turn in your exam.

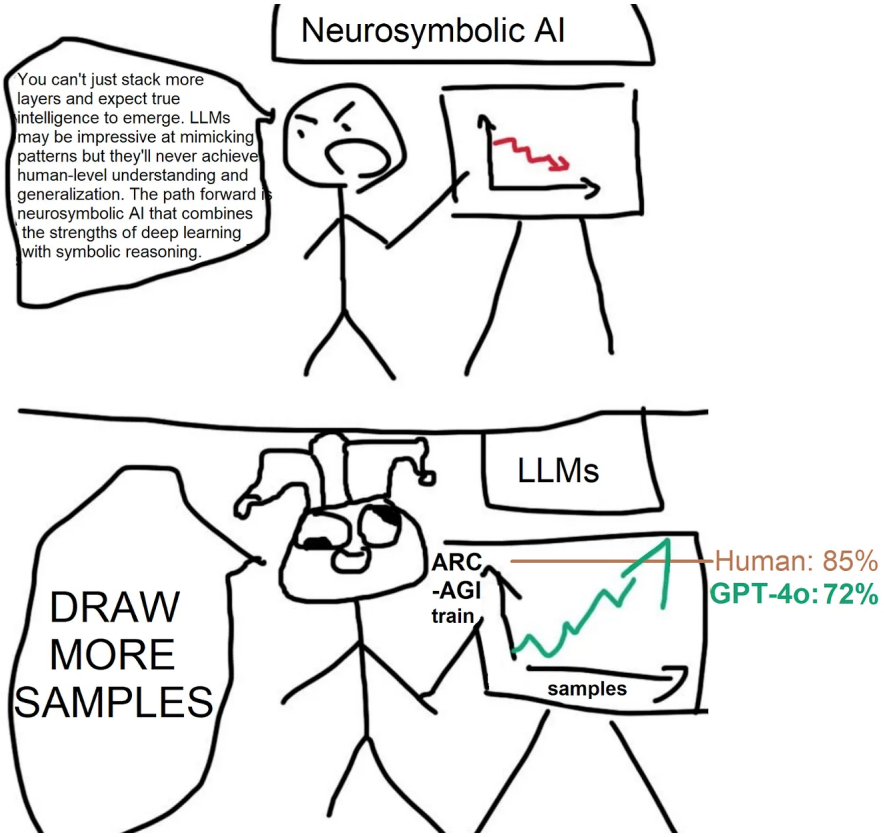
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

- NOTE: Please bubble in your answers. Do not write your answer to the side. Example:

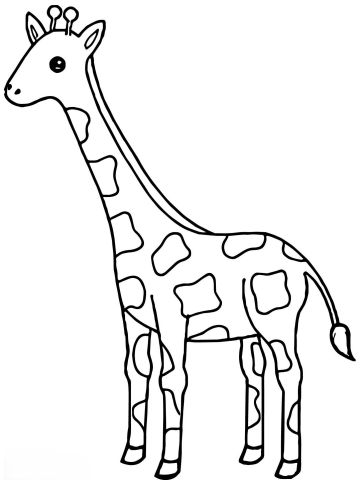
  Not selected answer: (a) Selected answer: (a)

- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- Multiple choice questions marked with [One Answer] should only be marked with one answer. All other multiple choice questions are [Select All That Apply], in which case any number of answers may be selected (**including none, one, or more**).
- For [Select All That Apply] questions, you will receive proportional credit for each option based on whether you get each "option" correct/incorrect. For example if there are 4 options, you will receive 0.25 points for each option that matches the solution.
- For each short answer question, please write your answer in the provided space.
- If you need to change an answer or run out of space, please very clearly indicate what your final answer is and what you would like graded. Responses where we cannot determine the selected option will be marked as incorrect.
- Please remain in your seats for the last 10 minutes of the exam. If you complete the exam before the last 10 minutes, you may turn in your exam to a TA.

These images are included only to cover the back of this page. They have no relation to the exam.
Comic from Natasha:



Giraffe for good luck :)



1. [One Answer] [1 points]

True/False: For a given model, irreducible error can be decreased by improving the model's complexity and increasing the amount of training data.

(a) True

(b) False

2. [One Answer] [1 points]

You're training a classifier model using a neural network built from scratch in PyTorch. You are unable to decide on the depth of the neural network, so you decide to make the network as deep as possible. Despite achieving low training loss, your model performs poorly on the XOR test data. Why? Choose the most accurate explanation.

(a) The neural network is too complex and has too high of a bias squared error.

(b) The neural network is too complex and has too high of a variance error.

(c) The neural network is unable to learn non-linearities since XOR data is not linearly separable.

(d) We need to make the neural network even deeper to capture the complex relationship in the XOR dataset.

3. [2 points]

As dataset sizes increase, would you be more or less inclined to use leave-one-out cross-validation (LOOCV)? Provide reasoning to support your answer.

Answer: _____

_____

_____

4. One Answer | 1 points

You are fine-tuning a model with parameters α, β, and γ, and have decided to perform 7-fold cross-validation to get the best set of hyperparameters. You have 5 candidate values for α, 3 candidate values for β, and 2 candidate values for γ. How many validation errors will you be calculating in total?

(a) Cannot be determined.
(b) 10
(c) 96
(d) 210
(e) 30

5. 3 points

You are analyzing the time until failure for a set of lightbulbs. The data represents the number of months each bulb lasted before failing and is given as follows: x_1, x_2, x_3, x_4. Assuming these times are modeled as being drawn from an exponential distribution. Derive the maximum likelihood estimate (MLE) of the rate parameter λ of this distribution. You must show your work.

Recall probability density function (PDF) for the exponential distribution is given by

$$f(x|\lambda) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$

Hint: You should not have n in your final answer

Answer: $\lambda =$ _____

6. Select All That Apply | 1 points

Which of the following can be convex?

(a) The intersection of non-convex sets
(b) The intersection of convex sets
(c) The union of non-convex sets
(d) The union of convex sets

7. One Answer | 1 points

For convex optimization objectives, taking a gradient step using full-batch GD ensures that your loss shrinks.

(a) True
(b) False

8. One Answer | 1 points

You are building a multi-class classifier using a deep neural network. You notice that your network is training slowly and that the gradients are diminishing quickly. Which activation function for the hidden layers of your network should you switch to, in order to avoid these issues?

(a) $f(x_i) = \frac{1}{1+e^{-x_i}}$
(b) $f(x_i) = \max(0, x_i)$
(c) $f(x_i) = x_i$
(d) $f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$

9. One Answer | 1 points

If two neural networks differ only in the number of hidden layers, the deeper network will always achieve a lower training loss given the same training data.

(a) True
(b) False

10. Select All That Apply | 1 points

Snoopy is training a neural network to classify birds into "Woodstock" and "Not Woodstock". He has a plot of the training and validation accuracy for the neural network model during the training process.

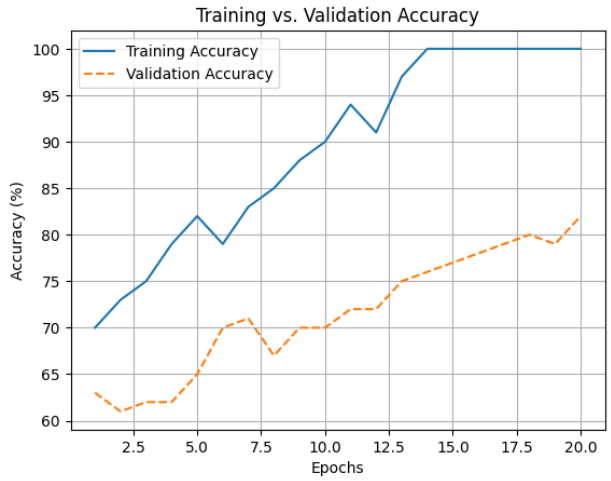


Figure 6: Snoopy's Training Plot

Which of the following actions could Snoopy take to help reduce the difference between training and validation accuracy?

(a) Increase the amount of training data

(b) Apply regularization techniques

(c) Reduce the complexity of the model (e.g., use fewer layers or units)

(d) Train for more epochs without making other changes

(e) Decrease the learning rate

11. One Answer | 1 points

Although both LASSO and PCA can be used for feature selection, they differ in their approach.

True/False: Specifically, LASSO sets some weight coefficients to 0 and selects a subset of the original features, whereas PCA selects features that minimize variance and creates a linear combinations of the original features.

(a) True
(b) False

12. One Answer | 1 points

What is the minimization objective for logistic loss? Here $\hat{y}$ is the prediction, and y is the ground truth label.

(a) $\log(1 + e^{-y\hat{y}})$
(b) $1 + \log(e^{-y\hat{y}})$
(c) $1 + e^{-y\hat{y}}$
(d) $1 + \log(e^{y\hat{y}})$

13. Select All That Apply | 1 points

The L-∞ norm is represented as $|| \cdot ||_\infty$ and is calculated for a vector $\mathbf{x} \in \mathbb{R}^d$ as $||\mathbf{x}||_\infty = \max_i(|x_i|)$. What happens to the parameters in w if we optimize for a standard linear regression objective with L-∞ regularization?

(a) There will be lots of parameters within w that are the same/similar absolute value.
(b) w will be very sparse.
(c) w will not be very sparse.
(d) Not enough information given to determine any of the above.

14. One Answer | 1 points

True/False: In k-means, increasing the value of k never worsens the model's performance on training data.

(a) True

(b) False

15. Select All That Apply | 1 points

Which of the following statements about PCA are true?

(a) The first principal component corresponds to the eigenvector of the covariance matrix with the smallest eigenvalue.

(b) If all the singular values are equal, PCA will not find a meaningful lower-dimensional representation.

(c) The principal components are the eigenvectors of the covariance matrix of the data.

(d) The reconstruction error of the recovered data points with a rank-q PCA strictly decreases as we increase q for all datasets.

16. Select All That Apply | 1 points

Consider the 2×2 matrix:

$$A = \begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix}$$

Let the Singular Value Decomposition (SVD) of A be given by:

$$A = U\Sigma V^T$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix containing the singular values of A. Which of the following statements are correct?

(a) The rank of A is 1.

(b) The nonzero singular value of A is 5.

(c) The columns of V must be $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

(d) The columns of V form an orthonormal basis for $\mathbb{R}^2$.

17. One Answer | 1 points

True/False: In kernel methods, we use a kernel function $k(x, x')$ to implicitly map input data into a feature space with different dimensions without explicitly computing the transformation. If we choose a linear kernel $k(x, x') = x^T x'$, then this is equivalent to mapping data into an infinite-dimensional feature space.

(a) True

(b) False

18. Select All That Apply | 1 points

Which of the following statements about kernels is/are true?

(a) The kernel trick is a technique for computing the coordinates in a high-dimensional space.

(b) If the kernel matrix K is symmetric, it is always a valid kernel.

(c) Eigenvalues of a valid kernel matrix must always be non-negative.

(d) The kernel trick eliminates the need for regularization.

19. Suppose we are doing polynomial kernel regression with training dataset $X \in \mathbb{R}^{n \times d}$.

(a) [1 points]

Let $\mathbf{1} \in \mathbb{R}^n$ denote the vector of ones. Suppose we are using the polynomial kernel with degree up to 1, i.e., degree zero and degree one. Write the corresponding kernel matrix K in terms of X and $\mathbf{1}$.

$K =$ _____

(b) [1 points]

Now suppose we are using the polynomial kernel with degree up to k starting from degree zero. Let M be the corresponding kernel matrix. What is $M_{i,j}$ for row i and column j? Write your answer in terms of $K_{i,j}$.

$M_{i,j} =$ _____

20. [Select All That Apply] [1 points]

Which of the following statements about k-Nearest-Neighbors are true?

(a) The time complexity of the k-NN algorithm for a single query is $O(N \cdot d)$, where N is the number of training samples and d is the number of features.

(b) k-NN is highly efficient for large datasets because it has a low computational cost during the training phase.

(c) k-NN can suffer from the curse of dimensionality, where the effectiveness of the distance metric diminishes as the number of features increases.

(d) Scaling the features is crucial for k-NN performance, as it ensures that all features contribute equally to the distance computation.

(e) k-NN is inherently faster when the number of dimensions (features) is very high, because higher dimensions make the distance between data points more sparse.

21. One Answer | 1 points

When choosing neural network architecture, we generally avoid overparameterization to prevent overfitting.

(a) True

(b) False

22. One Answer | 1 points

When performing stagewise additive modeling, to compute a model at each iteration, we access:

(a) The most recently computed model

(b) The most recently computed ensemble

(c) All previously computed models

(d) All previously computed ensembles

23. Select All That Apply | 1 points

Select the following which is true for the K-means algorithm.

(a) The number of clusters (K) in K-means is a trainable parameter.

(b) The time complexity for running the K-means learning algorithm is agnostic to the number of data points.

(c) The time complexity for matching an unseen data point to k learned centroids is agnostic to the number of data points.

(d) K-means is a parametric model.

(e) K-means algorithm requires labeled data.

(f) K-means performs poorly on data with overlapping clusters.

24. Select All That Apply | 1 points

The following statements describe properties of K-means and Gaussian Mixture Models (GMM). Which of them are correct?

(a) K-means is a "hard clustering" method, while GMM is a "soft clustering" method.

(b) GMM can be used for both clustering and probability density estimation.

(c) Both GMM and K-means assume spherical/circular clusters.

(d) GMM cannot be used when clusters overlap significantly, as it assumes non-overlapping Gaussians.

(e) K-means is sensitive to the selection of initial centroids, which may lead to different clustering results.

25. 1 points

Suppose you are training a GMM with n components. How many parameters need to be learned?

Answer: _____

26. Select All That Apply | 1 points

Which of the following regarding bootstrapping are true?

(a) Bootstrapping is an approach for hyperparameter tuning.

(b) Bootstrapping can be applied to large datasets but is most accurate on small datasets.

(c) For a dataset of size N, there exists 2^N possible unique bootstrap datasets.

(d) Bootstrapping is commonly used to estimate the sampling distribution of a statistic, such as the mean or standard deviation, when the true distribution is unknown.

(e) Since we select N samples when creating the bootstrap dataset, each data point is guaranteed to be selected.

27. 2 points

Suppose you are the hiring manager at "Goggles" (a hypothetical tech giant) and you receive thousands of applicants for a role. Since you took CSE446, you decided to build a model and use past hiring data to automate the resume screening process, which has never been done before in the company. The dataset contains resumes and the labels are whether or not the owner of the resume passed the resume screening stage (previously done by humans). The benefit is two fold. You are able to distill the large pool of applicants quickly and you also eliminate human bias when screening resumes.
Explain why this approach can be problematic.

Answer: _____

28. 2 points

Give an example of a task where we might expect a convolutional neural network to perform better than a deep neural network. Assume both models have roughly the same number of parameters.

Provide reasoning why the CNN might perform better in that setting.

Answer: _____

29. Select All That Apply | 1 points

In the context of linear regression, general basis functions are used to:

(a) Minimize the computational complexity of linear regression models.

(b) Increase the speed of convergence in gradient descent optimization.

(c) Encourage sparsity in the learned weights.

(d) Transform the input data into a higher-dimensional space to capture non-linear relationships.

30. Consider a neural network with 6 layers trained on a dataset of 600 samples with a batch size of 15.

a. 1 points

How many forward passes through the entire network are needed to train this model for 8 epochs?

Answer: _____

b. 1 points

How many forward passes through the entire network are needed to train this model for 5 epochs?

Answer: _____

31. We define a two-layer neural network for a regression task as follows:

Let the input be:

$$x \in \mathbb{R}^d$$

The hidden layer applies a linear transformation followed by a ReLU activation:

$$h = \sigma(W_1 x + b_1), \quad \sigma(z) = \max(0, z), \quad h \in \mathbb{R}^m$$

Where:

- $W_1 \in \mathbb{R}^{m \times d}$ is the weight matrix for the hidden layer.
- $b_1 \in \mathbb{R}^m$ is the bias vector for the hidden layer.
- $\sigma(z)$ is the ReLU activation function, applied element-wise.
- $h \in \mathbb{R}^m$ is the hidden layer output.

The output layer applies a linear transformation without any activation:

$$\hat{y} = W_2 h + b_2, \quad \hat{y} \in \mathbb{R}$$

Where:

- $W_2 \in \mathbb{R}^{1 \times m}$ is the weight matrix for the output layer.
- $b_2 \in \mathbb{R}$ is the bias term for the output layer.
- $\hat{y} \in \mathbb{R}$ is the model prediction.

We use the mean squared error (MSE) as the loss function:

$$L = \frac{1}{2}(\hat{y} - y)^2$$

Where:

- $y \in \mathbb{R}$ is the true target value.
- $\hat{y}$ is the predicted output.

a. [3 points]
Find the gradient of L with respect to W_2.

$\frac{\partial L}{\partial W_2}$: _____

b. [3 points]
Find the gradient of L with respect to b_1.
Hint: Don't forget to take the gradient of the activation function!

$\frac{\partial L}{\partial b_1}$: _____

32. [1 points]

Suppose a dataset has n samples and d features. What is the maximum number of non-empty terminal nodes a decision tree built on this dataset can have? Assume you cannot split on the same feature more than once on any given path.

Answer: _____

33. [3 points]

Prove K-means converges to a local minimum. An english proof (no explicit math) suffices.

Answer:

34. 1 points

Consider $M \in \mathbb{R}^{d \times d}$. Let λ be an eigenvalue of M. Suppose the eigenspace corresponding to λ equals $\mathbb{R}^d$. What is M in terms of λ?

$M =$ _____

END OF EXAM