CSE 446/546 Winter 2025 Midterm Exam

Feburary 7, 2025

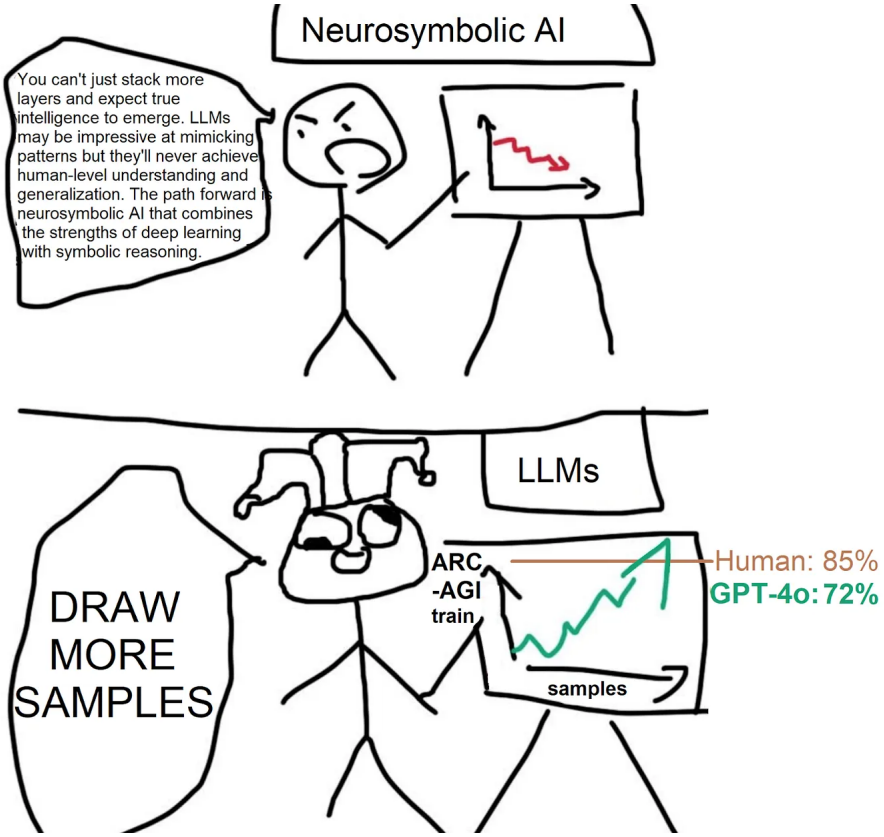
Name _____ **UW NetID** _____

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.

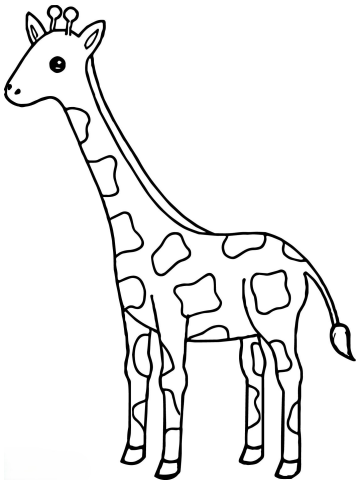
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

- 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. 1 points | One Answer

In a popular gacha game, the probability of pulling an SSR character on a single pull is 0.6% ($P = 0.006$). Assume that each pull is independent and follows a Bernoulli distribution. In such games, players often perform a 10-pull, which means making 10 pulls. Each of these 10 pulls is still independent, meaning the probability of getting an SSR in each pull remains 0.006. If you perform a 10-pull, what is the probability of pulling exactly 2 SSRs? You do not need to know what gacha game is to solve this problem.

(a) 0

(b) $1 - (1 - P)^{10} = 5.84\%$

(c) $(1 - P)^9 \times P \times 10 = 5.68\%$

(d) $(1 - P)^8 \times P^2 \times \frac{10 \times 9}{2} = 0.15\%$

2. 1 points | Select All That Apply

Below are several statements about Gradient Descent (GD) and Stochastic Gradient Descent (SGD). Which of the following are correct?

(a) For GD, each step aims to move along the gradient descent direction at the current point to reduce the value of the objective function.

(b) In SGD, each step computes an estimated gradient based on a single sample, introducing randomness, which may not guarantee that the objective function decreases in every step.

(c) Suppose you have model w_t at the t-th iteration of SGD. The expectation of the direction of the model update for SGD at step t is different from the negative direction of the gradient $-\nabla_w f(w)|_{w=w_t}$.

(d) GD requires the full gradient information of the objective function, while SGD only needs the gradient information on a single sample at each step.

3. 1 points

In a gacha game, the probability of obtaining an SSR character per pull is p, but p is unknown. To estimate p, Bob performed 100 pulls and obtained SSRs $k = 3$ times (i.e., 3 successes). Assume that each pull is independent and follows a Bernoulli distribution.

What is the likelihood of this this scenario occurring?

Likelihood function: $L(p) =$_____

What is the Maximum Likelihood Estimate (MLE) of p as a fractional number?

MLE: $\hat{p} =$_____

4. 1 points | One Answer

Suppose you train a linear regression model (without doing feature expansion), i.e., $f_w(x) = wx + b$, to approximate the cubic function $g(x) = 2x^3 + 7x^2 + 4x + 3$. What's the most likely outcome?

(a) The model will have low bias and low variance

(b) The model will have low bias and high variance

(c) The model will have high bias and low variance

(d) The model will have high bias and high variance

5. 1 points | One Answer

Adding more basis functions to a linear regression model always leads to better prediction accuracy on new, unseen data.

(a) True

(b) False

6. 1 points | One Answer

What datasets from the training/validation/test data split should you utilize during hyperparameter tuning?

(a) Training Data

(b) Training Data, Validation Data

(c) Training Data, Validation Data, Test Data

(d) Training Data, Test Data

7. 1 points | One Answer

Consider $u = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$, $v = \begin{bmatrix} -4 \\ 5 \\ 1 \end{bmatrix}$, $w = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$. Let $x \in \mathbb{R}^3$. Does there exist unique $a, b, c \in \mathbb{R}$ such that $a \cdot u + b \cdot v + c \cdot w = x$?

(a) Yes

(b) No

(c) Not enough information to determine

8. 1 points

Consider data matrix $X \in \mathbb{R}^{n \times d}$, label vector $y \in \mathbb{R}^n$, and regularization parameter $\lambda > 0$. Write the closed form solution for ridge regression.

Answer: _____

9. 1 points | One Answer

Consider a dataset containing three observations for a simple linear regression problem, where y is the dependent variable and x is the independent variable. The dataset is given as follows:

x	y
1	7
2	8
3	9

Find the coefficient β_1 of the linear regression (without bias) $y = \beta_1 x$ using the least squares as loss.

(a) $\frac{46}{14}$

(b) $\frac{14}{46}$

(c) $\frac{50}{14}$

(d) $\frac{14}{50}$

10. 1 points | One Answer

We can find the solution for LASSO by setting the gradient of the loss to 0 and solving for weight parameter w.

(a) True

(b) False

11. [1 points] [One Answer]

You are building a model to detect fraudulent transactions from a dataset of 100K samples. What would be the most effective way to split and utilize your data?

(a) Randomly take an 80-20 data split. Use 80% of the data for training, and 20% for both validation and evaluation.

(b) Use the first 80% of the data for training, the next 10% for validation, and the last 10% for evaluation.

(c) Randomly make a 80-10-10 data split. Use 80% of the data for training, 10% for validation, and 10% for evaluation.

(d) Select a random 80% of the data for training, use the remaining 20% for validation. Evaluate on the training set.

12. [1 points] [One Answer]

You are implementing a model to predict house prices. Your dataset contains 15 features (e.g., location, acres, proximity to city, etc.). However, you believe that many of these features are irrelevant to the house prices. Which method would be most suitable for your model?

(a) Logistic regression with L1 regularization.

(b) Logistic regression with L2 regularization.

(c) Linear regression with L1 regularization.

(d) Linear regression with L2 regularization.

13. [1 points] [Select All That Apply]

While training a model, you notice that it has a small bias but a high variance on the training data. Which of the following are valid strategies to address the high variance?

(a) Increase regularization constant.

(b) Train on a model class that is simpler.

(c) Increase the size of the training dataset.

(d) Use higher-degree features to capture more complex patterns in the data.

14. [1 points] [Select All That Apply]

After a student trains and evaluates a Logistic Regression model, you notice their test accuracy is 99.99%. You know that this was supposed to be a difficult dataset to model, so you investigate. Which of the following are **reasonable explanations** for this excessively high accuracy? Note that if you select multiple answers, not all of them have to be true at the same time.

(a) There was some form of train/test leakage, resulting in the model over-performing on the test set

(b) The data was not linearly separable, making it very easy for the model to classify things correctly

(c) The dataset was incredibly imbalanced, with most of the data points being labeled as positives

(d) The dataset was incredibly imbalanced, with most of the data points being labeled as negatives

15. [1 points] [One Answer]

$W \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^{n \times n}$, $Y \in \mathbb{R}^{p \times m}$, $Z \in \mathbb{R}^{m \times m}$, and $a \in \mathbb{R}^n$. If m, n, p are distinct, which one of the following expressions is valid?

(a) $(X^{-1}aa^\top W^\top)^{-1}(X^\top a)$

(b) $Xa^\top aW^\top(Z^{-1}Y^\top)$

(c) $WXaa^\top XZY^\top$

(d) None of the above

16. [1 points]

For what value of k will k-fold cross-validation create cross-validation splits equivalent to Leave-one-out cross-validation (LOOCV)? Assume you have n data points.

$$k = \_\_\_\_\_$$

17. [1 points] [One Answer]

We can decrease the variance of a model by increasing the model complexity.

(a) True

(b) False

18. [1 points] [Select All That Apply]

Which of the following statements are true about logistic regression? Recall that the sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$ for $x \in \mathbb{R}$

(a) L2 regularization is often used to reduce overfitting in logistic regression by adding a penalty for large coefficient values

(b) The logistic sigmoid function is used to model the probability of the positive class in binary logistic regression

(c) The maximum likelihood estimates for the logistic regression coefficients can be found in closed-form

(d) For any finite input $x \in \mathbb{R}$, $\sigma(x)$ is strictly greater than 0 and strictly less than 1. Thus, a binary logistic regression model with finite input and weights can never output a probability of exactly 0 or 1, and can never achieve a training loss of exactly 0.

19. [1 points] [Select All That Apply]

Below are several statements about the train/test/validation sets and cross-validation. Which of the following are correct?

(a) k-fold cross validation (where $k > 1$) is faster but more biased than leave-one-out (LOO) cross validation.

(b) k-fold cross validation (where $k > 1$) is faster and more accurate than leave-one-out (LOO) cross validation.

(c) The test set can be used to evaluate models during training and for hyperparameter tuning.

(d) The test error gives us an assessment of how our model does on unseen data.

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

Consider the principle of Maximum Likelihood Estimation (MLE), which is a method to estimate the parameters of a statistical model. Which of the following statements is correct?

(a) For MLE, samples must be drawn i.i.d. (independent and identically distributed).

(b) Once we have a log-likelihood function, we maximize it with respect to the parameter θ to find the parameter estimate $\hat{\theta}_{MLE}$.

(c) MLE always provides an unbiased estimator of the true parameter.

(d) MLE identifies the model parameters that maximize the likelihood of the observed data.

21. [1 points] [One Answer]

If we run gradient descent on $f(x)$, gradient descent guarantees that we will converge to the global minimum even if $\nabla^2 f(x) \succeq 0$ does not hold some x, i.e., the Hessian of the objective function is not positive semi-definite for some x.

(a) True

(b) False

22. [1 points] [One Answer]

Let $A_1, A_2, ..., A_n \sim \mathcal{N}(\mu, \sigma^2)$. What is $\mathbb{E}[A_1 + A_2 + A_3]$?

(a) 3μ

(b) 6μ

(c) 9μ

(d) Cannot be determined from the given information.

23. 1 points

Consider a function f(x,y) representing a loss function in a 2-dimensional space, where gradient descent is used to minimize f. Given the function: $f(x, y) = x^2 + 2y^2 + 4xy$ where the initial point is $(x_0, y_0) = (1, 1)$ and the learning rate is 0.1, write down the (x_1, y_1) you get after one step of gradient descent.

Answer: _____

24. 1 points | One Answer

For machine learning models and datasets in general, as the number of training data points grows, the prediction error of the model on unseen data (data not found in the training set) eventually reaches 0.

(a) True

(b) False

25. 1 points | One Answer

Which of the following statements about ridge regression are true?

(a) When there are correlated features, ridge regression typically sets the weights of all but one of the correlated features to 0.

(b) Compared to unregularized linear regression, the additional computational cost of ridge regression increases proportional to the number of data points in the dataset.

(c) Ridge regression reduces variance at the expense of increasing bias.

(d) Using ridge and lasso regularization together (e.g., minimizing a training objective of the form $f(w) = \sum_{i=1}^{n} (y^{(i)} - {x^{(i)}}^\top w)^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$) makes the training loss no longer convex.

26. 1 points | Select All That Apply

Let $n \in \mathbb{N}$ such that $n > 1$. Which of the following functions are convex (with respect to x) over its entire domain?

(a) $f(x) = 5 + \sum_{i=1}^{n} x^{2i}$

(b) $f(x) = 5 + \sum_{i=1}^{n} x^{2i+1}$

(c) $f(x) = 3 \cdot e^{-\frac{x^2}{n}}$

(d) $f(x) = x - \log_{\pi}(x^n)$ on $(0, \infty)$

27. 1 points | Assume $n \neq d$. Suppose $x_1, x_2, ..., x_n$ span $\mathbb{R}^d$. What is the rank of $\sum_{i=1}^{n} x_i x_i^\top$? Write your answer in terms of n and d. Hint: for any matrix A, $\text{rank}(A^\top A) = \text{rank}(A)$.

Answer: _____

28. 1 points

Describe one advantage of full-batch gradient descent over mini-batch gradient descent.

Answer: _____

_____

_____

29. 1 points | Describe one advantage of mini-batch stochastic gradient descent ($1 < B < n$) over stochastic gradient descent with batch size $B = 1$ (e.g., updating the parameters at each iteration based only on one randomly sampled training point).

Answer: _____

_____

_____

END OF EXAM