CSE 446 Spring 2025 Midterm Exam

May 02, 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.

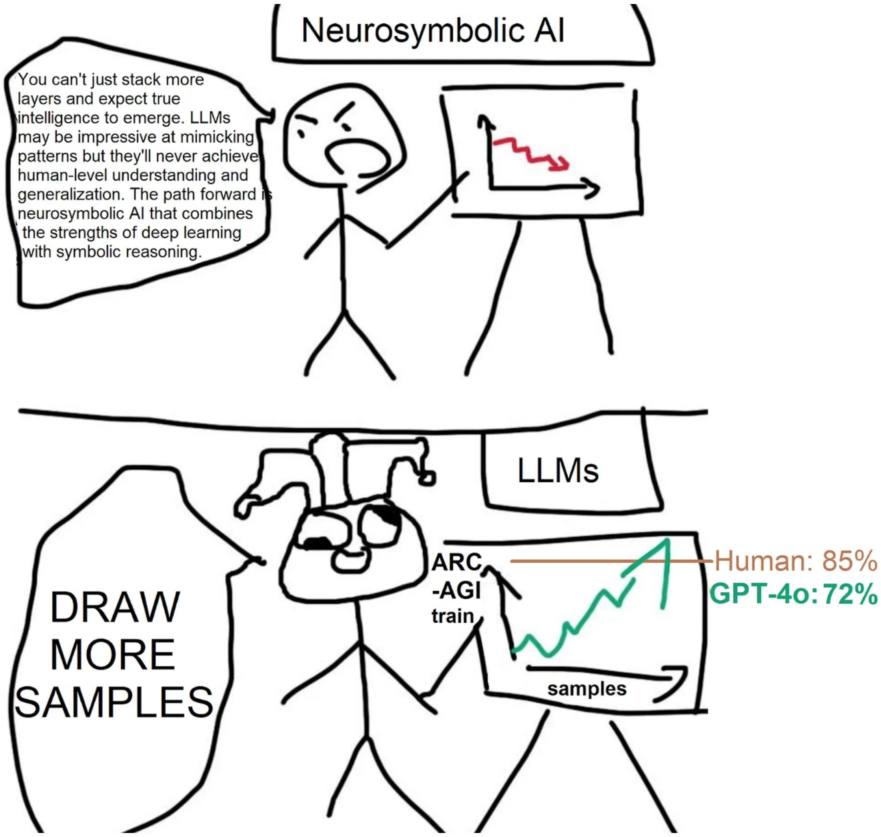
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

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

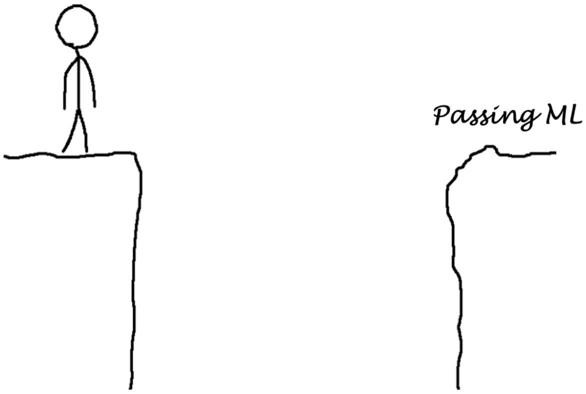
  Not selected answer: ⓐ Selected answer: [✓ a]

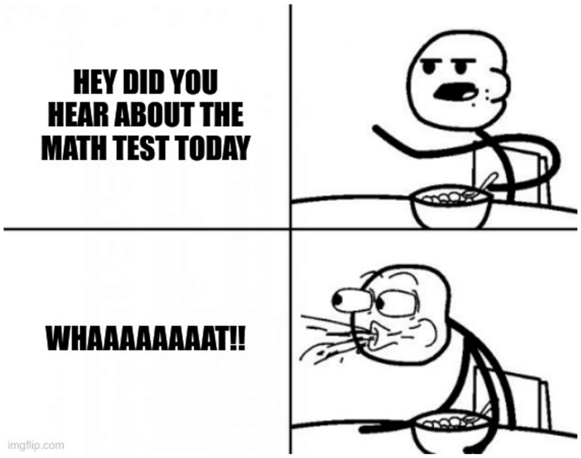
- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- Multiple choice questions marked with One Answer should only be marked with one answer. All other multiple choice questions are Select All That Apply, in which case 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:


Neurosymbolic AI
You can't just stack more layers and expect true intelligence to emerge. LLMs may be impressive at mimicking patterns but they'll never achieve human-level understanding and generalization. The path forward is neurosymbolic AI that combines the strengths of deep learning with symbolic reasoning.
LLMs
DRAW MORE SAMPLES
ARC -AGI train
Human: 85%
GPT-4o: 72%
samples

Something to laugh about :)





1. One Answer | 1 points

Saket is building a classification model from a dataset with two classes of student reactions to the jokes he makes during section: Funny jokes (F) and Dull Jokes (D). The probability of a randomly selected joke being funny is 0.3. The probability the model makes the correct classification given that the joke is dull is 0.7, and the probability the model makes the correct classification given that a joke is funny is 0.2. What is the probability that a randomly selected joke is dull given that it has been classified as dull?

(a) $\frac{42}{73}$

(b) $\frac{49}{73}$

(c) $\frac{21}{73}$

(d) 1

Correct answers: (b)

Explanation: The answer is (b), which is $\frac{49}{73}$.

Let A be the event that the sample is dull.
Let B be the event that the sample is classified as dull.

We have $P(A^c) = 0.3$ and $P(A) = 1 - P(A^c) = 1 - 0.3 = 0.7$.
We have $P(B|A) = 0.7$ and $P(B^c|A^c) = 0.2$.
(Note that $P(B|A^c) = 1 - P(B^c|A^c) = 1 - 0.2 = 0.8$)

We want to calculate $P(A|B)$ here using Bayes Rule and the Law of Total Probability.
$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.7 \cdot 0.7 + 0.8 \cdot 0.3 = 0.73$
$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.7 \cdot 0.7}{0.73} = \frac{0.49}{0.73} = \frac{49}{73}$

Here, our answer is $P(A|B) = \frac{49}{73}$, which is (b).

2. One Answer | 1 points

Suppose we train a model $f(x) = x^\top \widehat{w}$ on dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, by optimizing the following objective:

$$\widehat{w} = \arg\min_w \sum_{i=1}^n (x_i^\top w - y_i)^2 + \log_k ||w||_2^2$$

for $k > 1$. As k increases, what likely happens to the variance of our model?

(a) The variance of our model increases.

(b) The variance of our model decreases.

(c) The variance of our model remains unaffected.

(d) The variance of our model does not change in a predictable way.

Correct answers: (a)

Explanation: As k increases, the penalty term goes to zero, thus increasing the variance of our estimator.

3. One Answer | 1 points

You are building a classification model for your favorite soccer team to determine whether a penalty kick will result in a goal or not, and your data set contains 300 positive examples (resulted in a goal) and 200 negative examples (did not result in a goal). After training, you find that your model has an accuracy of 70% and misclassifies 15% of negative examples as positive. What is the probability that your model will misclassify a positive example as negative?

(a) 15%

(b) 24%

(c) 30%

(d) 40%

(e) Cannot be determined.

Correct answers: (d)

Explanation: We have 500 examples in total, and 500 * 70% = 350 were correctly classified. The rest (150) were incorrectly classified. We also know that 200 * 15% = 30 negative examples were misclassified. This leaves us with 150 - 30 = 120 misclassified examples, and we know these must all be positive examples that were misclassified. This means that $\frac{120}{300} = \frac{2}{5} = 40\%$ of positive examples were misclassified as negative.

4. One Answer | 1 points

True/False: When using the least squares method for linear regression, outliers would have a minimal impact since the least squares method averages out their effects.

(a) True

(b) False

Correct answers: (b)

Explanation: Outliers would cause the regression line to skew, thus leading to a representation that may not be an accurate representation of the data.

5. Select All That Apply | 1 points

In the standard MLE derivation for a parameter in a probability distribution (like the ones we saw in class), why do we apply the logarithm to our equation?

(a) The function is monotonically increasing and therefore does not change the result of our optimization objective.

(b) The function is concave and therefore does not change the result of our optimization objective.

(c) The function is monotonically decreasing and therefore allows us to change our "argmax" to an "argmin".

(d) Applying the logarithm allows us to convert products ($\prod$) into sums ($\sum$), letting us calculate derivatives easier.

Correct answers: (a), (d)

Explanation: The logarithm function is monotonically increasing, which makes (A) true and (C) false. (B) is false. the logarithm function is concave, but that is not why it doesn't change the result of our optimization. (D) is true due to properties of logarithms and derivatives.

6. One Answer | 1 points

After training a model, your model has a low train error and a high test error. Which of the following can be inferred?

(a) The model is underfitting.

(b) The model will generalize well because it has low bias.

(c) Training on more data will likely increase the model's performance on unseen data.

(d) Training on more highly-informative features will decrease the test error.

(e) Reducing model complexity will reduce the irreducible error.

Correct answers: (c)

Explanation: A is incorrect since the model has a low, not high, bias. B is also incorrect since a low bias alone doesn't guarantee that the model will generalize to unseen data. C is correct since more training data generally decreases the variance. D is incorrect since increasing model complexity by increasing the number of features in the dataset will likely increase the variance. E is incorrect since irreducible error comes from underlying noise in the dataset and changing any factor of the model will not change this error.

7. One Answer | 1 points

Given the following Hessian Matrix, which of the following could be the original f(x,y) function?

$$H_f(x,y) = \begin{bmatrix} 2\ln(y) + y^2 e^{xy} & \frac{2x}{y} + e^{xy}(1+xy) \\ \frac{2x}{y} + e^{xy}(1+xy) & -\frac{x^2}{y^2} + x^2 e^{xy} - 6y \end{bmatrix}$$

(a) $f(x,y) = x^2 \ln(y) + e^{xy} - y^3$

(b) $f(x,y) = x^2 \ln(y) + e^{x+y} - y^3$

(c) $f(x,y) = x^3 \ln(y) + e^{xy} - y^3$

(d) $f(x,y) = x^2 \ln(y) + e^{xy} - y^4$

Correct answers: (a)

Explanation: Option A. Take the second derivative for all 4 $\partial^2 x$, $\partial^2 y$, ∂yx, ∂xy and you will get option a matches.

8. Select All That Apply | 1 points

Saket did not pay attention during lecture and did not split the data into a training set and testing set and instead used all the data to train and test a given model. What is the consequence of not splitting the data into a training set and testing set?

- (a) Nothing, Saket knows what he is doing
- (b) The model may overfit and perform poorly on unseen data
- (c) The model may underfit and perform poorly on unseen data
- (d) Saket will overestimate the performance of his model on unseen data

Correct answers: (b), (d)

Explanation: Choice B and D. If we do not split the data into a training set and testing set, then the model will test on seen data. thus resulting in overfitting and a lower error rate than intended. So the performance is not going to be as expected.

9. One Answer | 1 points

Suppose you are designing a model that predicts whether or not a patient will be readmitted into a hospital within a month. The hospital provides a dataset with 25 clinical features per patient (like age, gender, and blood pressure), but not all of them might be relevant to readmission. The data is such that it's possible to draw a straight line (or a higher-dimensional hyperplane) that perfectly divides the patients who were readmitted from those who were not. Which is the most appropriate choice of procedure to train a model in this scenario?

- (a) train an L1 regularized Logistic Regression, then retrain with unregularized Logistic Regression
- (b) train an L2 regularized Logistic Regression, then retrain with unregularized Logistic Regression
- (c) train an L1 regularized Logistic Regression, then retrain with L2 regularized Logistic Regression
- (d) train an L1 regularized Logistic Regression

Correct answers: (c)

Explanation: Train with L1 first, for feature selection. And then retrain with L2 regularization. L2 regularization is required an unregularized model will overfit linearly separable data.

10. 1 points

Describe a problem that might occur if you're training a Logistic Regression model and the data is linearly separable. Around 1-3 sentences.

____

____

____

Explanation: The model will overfit extremely as the magnitudes of the weights increase towards infinity.

11. One Answer 1 points

Suppose we standardize a given dataset. The optimal bias term will be 0 in least-squares linear regression.

ⓐ True

ⓑ False

Correct answers: (b)

Explanation: The offset is the average y value.

12. One Answer 1 points

Let $f, g : \mathbb{R} \to \mathbb{R}$ be convex. Which of the following functions is always convex?

ⓐ $h(x) = f(x) \cdot g(x)$

ⓑ $h(x) = f \circ g(x)$

ⓒ $h(x) = \min(f(x), g(x))$

ⓓ $h(x) = \max(f(x), g(x))$

Correct answers: (d)

Explanation: Pointwise maximum preserves convexity (see section 5)

13. $\boxed{\text{One Answer}}$ $\boxed{\text{1 points}}$

Given a small enough learning rate, gradient descent will converge to the global minima.

ⓐ True

ⓑ False

Correct answers: (b)

Explanation: This is false because non-convex functions can have multiple local minima / saddle points and GD may converge to one of those.

14. $\boxed{\text{1 points}}$

This is the equation for the bias-variance tradeoff. η is the "squared-error-optimal" predictor. D is a dataset $\{(x_i, y_i)\}_{i=1}^n$ sampled from P_{XY}. $\widehat{f}_D \in \mathcal{F}$ is the learned least-squares predictor for some function class $\mathcal{F}$.

Which terms correspond with which concepts?

Write the number of the term next to the concept you think it corresponds with.

$$\begin{aligned}\underbrace{\mathbb{E}_{Y|X}[\mathbb{E}_D[(Y - \widehat{f}_D(x))^2]|X = x]}_{\text{expected squared error}} = &\underbrace{\mathbb{E}_{Y|X}[(Y - \eta(x))^2|X = x]}_{\text{term 1}} \\ &+ \underbrace{(\eta(x) - \mathbb{E}_D[\widehat{f}_D(x)])^2}_{\text{term 2}} \\ &+ \underbrace{\mathbb{E}_D[(E_D[\widehat{f}_D(x)] - \widehat{f}_D(x))^2]}_{\text{term 3}}\end{aligned}$$

Variance: ____________________

Bias: ____________________

Irreducible error: ____________________

Explanation: Variance: term 3.
Bias: term 2
Irreducible error: term 1.
See lecture 4 https://courses.cs.washington.edu/courses/cse446/25sp/schedule/lecture-04/lecture-04-annotated.pdf

15. One Answer | 1 points

Which of the following is an advantage of using ridge regression over unregularized linear regression?

- (a) The ridge objective is concave
- (b) The ridge objective is convex
- (c) The ridge objective always has a unique solution
- (d) The ridge objective has a closed-form solution

Correct answers: (c)

Explanation: (a) Ridge objective is not concave (b) Unregularized linear regression objective is convex as well. (c) Unregularized linear regression does not always have a unique solution, L2 penalty fixes this. (d) Unregularized linear regression has a closed form solution as well

16. One Answer | 1 points

True/False: Lasso Regression uses the square of the L2 norm while Ridge Regression uses the L1 Norm.

- (a) True
- (b) False

Correct answers: (b)

17. One Answer | 1 points

You have independent random variables X, Y such that $X \sim \mathcal{N}(1, 2)$ and $Y \sim \mathcal{N}(3, 4)$. What is $\mathrm{Var}(5X + 6Y + 7)$?

- (a) 35
- (b) 34
- (c) 195
- (d) 194
- (e) Cannot be determined

Correct answers: (d)

Explanation: Because the variables are independent, $\mathrm{Var}(5X + 6Y + 7) = \mathrm{Var}(5X) + \mathrm{Var}(6Y) + \mathrm{Var}(7)$. The variance of a constant is 0, so we remove that term to get $\mathrm{Var}(5X) + \mathrm{Var}(6Y)$. Moving a coefficient outside of the variance function squares it. So we have $25\mathrm{Var}(X) + 36\mathrm{Var}(Y)$. We plug in the given variances for X and Y to get $25 \cdot 2 + 36 \cdot 4 = 194$

18. One Answer | 1 points

The objective function is $L(w) = \|Xw - Y\|_2^2$. What is the gradient of $L(w)$ with respect to w?

ⓐ $2Y^\top(Xw - Y)$

ⓑ $2X^\top(X^\top Xw - Y)$

ⓒ $2X^\top(Xw - Y)$

ⓓ $2Y^\top(X^\top Xw - Y)$

Correct answers: (c)

Explanation:

$$\begin{aligned}
\nabla_w\|Xw - Y\|_2^2 &= \nabla_w((Xw - Y)^T(Xw - Y)) \\
&= \nabla_w((w^TX^T - Y^T)(Xw - Y)) \\
&= \nabla_w(w^TX^TXw - Y^TXw - w^TX^TY + Y^TY) \\
&= 2X^TXw - X^TY - X^TY + 0 \\
&= 2X^T(Xw - Y)
\end{aligned}$$

19. One Answer | 1 points

Which of the following is true, when choosing to use Maximum Likelihood Estimation (MLE)?

ⓐ MLE cannot be used if we do not know the exact distribution of our data.

ⓑ MLE works well for any data distribution, so we do need knowledge of the true distribution.

ⓒ MLE will produce unbiased estimates regardless of the data distribution or the likelihood function that we choose

ⓓ MLE works even if the true distribution of our data isn't known. We can make an educated guess for the distribution of our data for our likelihood function.

Correct answers: (d)

Explanation: We often use MLE in situations where the true distribution of the data is not known. In MLE, we construct a likelihood function based on this chosen distribution and find the parameter values that maximize the probability of observing our data. While ideal to know the true distribution, MLE enables estimation through an educated guess, though the accuracy of the estimates depends on the appropriateness of our chosen distribution.

20. One Answer | 1 points

Consider the function $f(a) = 5a^2 - 3a + 2$. You want to use gradient descent to find the unique minimum, which you know is at $a_* = 0.3$. If at time t you arrive at the point $a_t = 3$, what value for the step size would bring you to a_* at time $t+1$?

- (a) 0.001
- (b) 0.01
- (c) 0.1
- (d) 1

Correct answers: (c)

Explanation: Following the standard gradient descent update formula, we get $0.3 = 3 - \eta * \nabla f(a)$. $\nabla f(a) = 10a - 3$, so $\nabla f(3) = 27$. Plugging this in, we get $0.3 = 3 - \eta * 27$. Solving this equation, we get $\eta = 0.1$.

21. One Answer | 1 points

Donovan is training some machine learning model, and is telling you about it. He needed to *standardize* the data, so he computed the mean and standard deviation of each feature in the entire dataset X and applied the transformation correctly. He then created non-overlapping subsets of X called X_{train}, $X_{validation}$, and X_{test}. To train, validate, and test their model respectively. In this setup, was there train/test leakage?

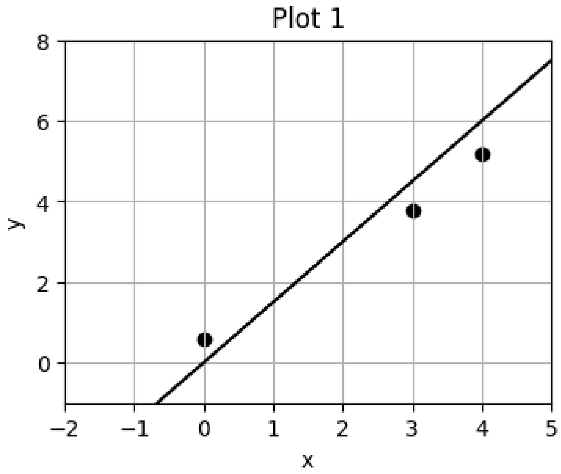
- (a) Yes
- (b) No

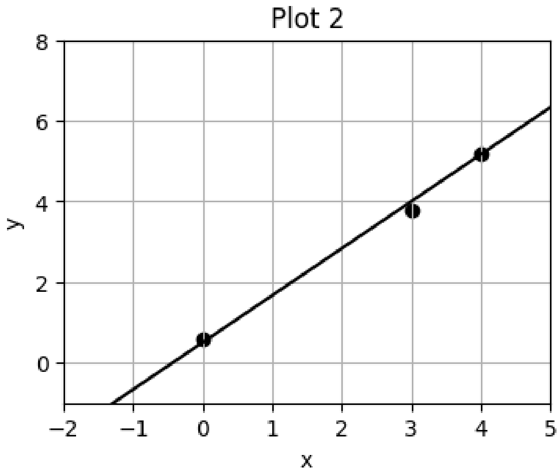
Correct answers: (a)

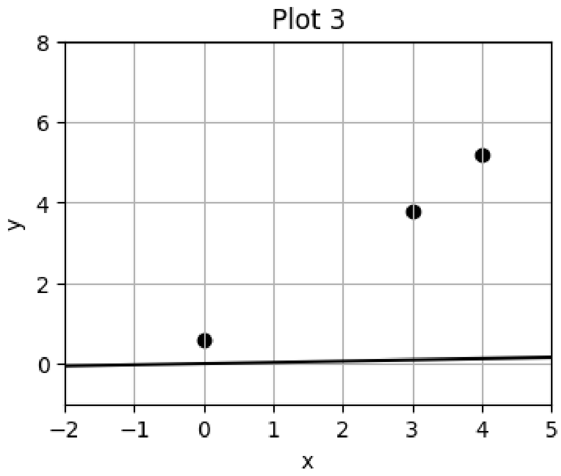
Explanation: They standardized the whole dataset using information from the test set (as it is a subset of X), and this is a form of train/test leakage.

22. $\boxed{\text{1 points}}$

The following plots show 3 data points and 3 models. The data is the same for all 3 models. Match the learned model to the equation used for linear regression.







$\widehat{w} = (X^TX + \lambda I)^{-1}X^Ty$. Plot number: ________________ (for $\lambda > 0$)

$\widehat{w} = (X^TX)^{-1}X^Ty$. Plot number: ________________

$\widehat{w} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^Ty$, where $\widetilde{X} = \begin{bmatrix} X & \vec{\mathbf{1}} \end{bmatrix}$. Plot number: ________________

Explanation: $\widehat{w} = (X^TX + \lambda I)^{-1}X^Ty$: plot 3, because it is overregularized.

$\widehat{w} = (X^T X)^{-1} X^T y$: plot 1, because it doesn't have an offset.
$\widehat{w} = (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T y$: plot 2, because it has an offset.

23. One Answer | 1 points

True/False: The training error is a better estimate of the true error than the cross-validation error.

(a) True

(b) False

Correct answers: (b)

Explanation: Cross validation is better and closer to true error since it deals with somewhat unseen data.

24. Select All That Apply | 1 points

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, smooth function whose derivative $f'(x)$ is also continuous. Suppose f has a unique global minimum $x^\star \in (-\infty, \infty)$, and you are using gradient descent to find $x^\star$. You fix some $x^{(0)} \in R$ and step size $\eta > 0$, and run $x^{(t)} = x^{(t-1)} - \eta f'(x^{(t-1)})$ repeatedly. Which of the following statements are true?

(a) Gradient descent is sure to converge, to some value, for any step size $\eta > 0$.

(b) If f has a local minimum x' different from the global one, i.e., $x' \neq x^\star$, and $x^{(t)} = x'$ for some t, gradient descent will not converge to $x^\star$.

(c) Assuming gradient descent converges, it converges to $x^\star$ if and only if f is convex.

(d) If, additionally, f is the objective function of logistic regression, and gradient descent converges, then it converges to $x^\star$.

Correct answers: (b), (d)

Explanation: A is false because for a large enough step size, gradient descent may not converge. B is correct because $f'(x') = 0$, so gradient descent will never move from a local minimum. C is false because you could "accidentally" initialize GD at $x^\star$ even if f is non-convex. D is correct because the objective of logistic regression is convex.

25. $\boxed{\text{1 points}}$

What is the tradeoff between the size of the validation set and the size of the training set? Around 1-3 sentences.

____

____

____

Explanation: Larger validation set means a better estimate of performance on unseen data. But at the cost of lost training data.

26. $\boxed{\text{1 points}}$

Consider $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$. Suppose $\widehat{w} = \arg\min_w ||Xw - y||_2$ has a unique solution. Fill in the blank for the following vector spaces. Write NA if the there is not enough information to determine the answer.

$\text{Col}(X) =$ ______________________

$\text{Row}(X) =$ ______________________

$\text{Null}(X) =$ ______________________

Explanation: Since in this case the linear regression objective has a unique solution, X must be full rank as well as $n \geq d$, so Row$(X) = \mathbb{R}^d$ and Null$(X) = \{0\}$. Since $n \geq d$, we are unable to guarantee anything about Col(X) other than the fact that it is a subspace of $\mathbb{R}^n$.

27. $\boxed{\text{1 points}}$

For a function $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x, y, z) = xy + x^2 \ln(z) + e^{yz}$. Calculate the gradient of f.

Gradient = ________________________________

Explanation:

$$\nabla f(x) = \begin{pmatrix} y + 2x \ln(z) \\ x + z\, e^{yz} \\ \dfrac{x^2}{z} + y\, e^{yz} \end{pmatrix}$$

28. $\boxed{\text{1 points}}$

Describe a scenario where one would choose to use Ridge regression over Lasso regression. Around 1-4 sentences.

____

____

____

Explanation: Ridge regression is better when all features are important and you don't want to remove any of them. For example, if you are predicting something using many related variables (like gene data), Ridge helps by keeping all the features and just shrinking their values, instead of setting some to zero like Lasso does. It's also useful when there are more features than data points or when features are highly correlated.

29. $\boxed{\text{2 points}}$

Answer the following questions about the Softmax function.

a. Explain how the Softmax function transforms an input vector (logits) and why it is suitable for multi-class classification.

____

____

____

b. Suppose a model outputs the following values/logits for a 3-class classification problem.

$$z = [2, 1, 5]$$

Compute the softmax probabilities. *You can leave the values in terms of exponentiated numbers.*

$$\text{Softmax}(z) = \Bigg[\qquad\qquad , \qquad\qquad , \qquad\qquad \Bigg]$$

Explanation: a. The Softmax function transforms an input vector into a probability distribution over K classes by exponentiating each term and dividing it by the sum of all the exponentiated values in the vector. softmax$(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$. This is helpful for multiclass classification tasks because it shows the model's uncertainty over multiple classes in a normalized way.

b. $\sum_{j=1}^{K} e^{z_j} = e^2 + e^1 + e^5$

Softmax$(z) = \left[\frac{e^2}{e^2+e^1+e^5}, \frac{e}{e^2+e^1+e^5}, \frac{e^5}{e^2+e^1+e^5}\right]$

END OF EXAM