

CSE 446/546 Autumn 2025 Midterm Exam

October 28, 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: For multiple choice questions, clearly indicate your answer by filling in the letter(s) associated with your choice. Bubble in your answers. Do not write your answer to the side.
Example:

Not selected answer: ☐ a Selected answer: ☒ a

- Multiple choice questions marked with should only be marked with one answer. All other multiple choice questions are , in which case any number of answers may be selected (**including none, one, or more**).
- For questions, you will receive proportional credit for each option based on whether you get each “option” correct/incorrect. For example, if the question is worth 1 point and 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 write your NetID at the top of every page.
- 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.

ID: _____

Page 2

This image is included only to cover the back of this page. It has no relation to the exam.



1. 1 point

When a model underfits, what happens to its bias and variance? When the model overfits, what happens to its bias and variance?

Solution: The bias-variance tradeoff describes the likely trend in models that when bias is high, variance is low and when bias is low, variance is high. When the bias is high and variance is low, the model is underfitting and making systematic errors. When the bias is low and variance is high, the model is overfitting and overly sensitive to fluctuations within the data.

2. The probability mass function (PMF) for the geometric distribution counting the number of trials until the first success is given by $P(X = x) = p(1 - p)^{x-1}$ for $x \in \{1, 2, 3, \dots\}$.

a. 2 points

Suppose you get the following sample data from a geometric distribution, which counts the number of trials until the first success: $x_1, x_2, \dots, x_n$

Write down the expression for $\hat{p}$, the MLE for the success probability p of this geometric distribution, as a function of $x_1, x_2, \dots, x_n$.

$$\hat{p} = \underline{\hspace{2cm}}$$

b. 1 point

Would $\hat{p}$ be more likely to be accurate (i.e., closer to the true underlying p^* of the data-generating distribution) with less sample data or more sample data? Briefly explain.

Solution: The answer is $p = \frac{n}{\sum_{i=1}^n x_i}$.

First, we want to calculate the likelihood function $L(x_1, \dots, x_n|p)$ below.

$$L(x_1, \dots, x_n|p) = P(x_1|p) \cdot P(x_2|p) \cdot \dots \cdot P(x_n|p) = p(1-p)^{x_1-1} \cdot p(1-p)^{x_2-1} \cdot \dots \cdot p(1-p)^{x_n-1} \\ = p^n \cdot (1-p)^{(x_1+x_2+\dots+x_n)-n}$$

Now, we calculate the log-likelihood function:

$$\ln L(x_1, \dots, x_n|p) = \ln(p^n \cdot (1-p)^{(x_1+x_2+\dots+x_n)-n}) = n \cdot \ln(p) + (x_1+x_2+\dots+x_n-n) \cdot \ln(1-p)$$

To find the argmax of p (and thus the MLE) of this log-likelihood expression, we need to take its derivative with respect to p and set it equal to 0.

$$\frac{d}{dp} \ln L(x_1, \dots, x_n|p) = \frac{d}{dp} (n \cdot \ln(p) + (x_1 + x_2 + \dots + x_n - n) \cdot \ln(1-p)) = \frac{n}{p} - (x_1 + x_2 + \dots + x_n - n) \cdot \frac{1}{1-p} = 0$$

$$\implies n(1-p) - p(x_1 + x_2 + \dots + x_n - n) = 0 \implies n - np - p(x_1 + x_2 + \dots + x_n) + pn = 0$$

$$\implies n - p(x_1 + x_2 + \dots + x_n) = 0 \implies p = \frac{n}{x_1 + x_2 + \dots + x_n}$$

Thus, the MLE here of p is given by $p = \frac{n}{\sum_{i=1}^n x_i}$.

For part b, more sample data leads to less variance of $\hat{p}$

3. 1 point

When conducting ridge regression, the regularization constant λ serves as a tunable hyperparameter. Which split of your dataset should you use to select λ : the training set, the validation set, or the test set? Briefly explain your answer by describing what might go wrong if you instead chose the other two out of those three options.

Solution: The validation set. If you choose the training set, you will end up with $\lambda = 0$; if you use the test set, you will overfit to the test set.

4. 1 point

What does it mean for a predictor to “generalize?”

Solution: We say a predictor generalizes if it performs as well on unseen data as on training data.

5. One Answer 1 point

You change your model from a complex one (e.g., higher degree polynomial expansion) to a much simpler one (e.g., linear model) on the same dataset. What is the most likely effect?

- (a) **Bias $\uparrow$, Variance $\downarrow$, Irreducible error unchanged**
- (b) Bias $\downarrow$, Variance $\uparrow$, Irreducible error $\downarrow$
- (c) Bias $\uparrow$, Variance $\uparrow$, Irreducible error unchanged
- (d) Bias $\downarrow$, Variance $\downarrow$, Irreducible error $\uparrow$

Solution: The answer is A. Simplifying the model usually increases bias and reduces variance; the irreducible error (data noise) doesn't change with model choice.

6. One Answer 1 point

Consider a regression setting where you have standardized all features (mean 0, variance 1). In this setting, many features are highly correlated. Which statement is most accurate?

- (a) Ridge (L2) typically yields many exact zeros, while Lasso (L1) rarely sets coefficients to zero.
- (b) Lasso has a general closed-form solution; Ridge does not.
- (c) **Ridge tends to spread weight across correlated features, while Lasso often selects one feature from a correlated group and sets the others to zero.**
- (d) Lasso regression only works when the features do not have zero mean, so Lasso regression will not induce sparsity on standardized data.

Solution: Correct answer is C. Ridge shrinks coefficients continuously and rarely to exactly zero, distributing weight among correlated features. Lasso applies an L1 penalty that induces sparsity, often picking one feature from a correlated set and zeroing the rest. Ridge has a closed-form solution; Lasso generally does not (except in special cases). Lasso regression works regardless of the mean of the features, and standardization is still common so that the penalty treats features comparably.

7. 1 point

Give a value for the constant $c \in \mathbb{R}$ that makes f a valid PDF:

$$f(x) = \begin{cases} cx, & \text{if } 0 < x < 10 \\ 0, & \text{otherwise} \end{cases}$$

$c =$ _____

Solution: If $c = \frac{1}{50}$, $f(x)$ is a valid PDF.

One requirement for a valid PDF is that $\int f(x) = 1$:

$$\int_0^{10} cx = 1 \tag{1}$$

$$\left[\frac{1}{2}cx^2\right]_0^{10} = 1 \tag{2}$$

$$\frac{1}{2}10^2c - \frac{1}{2}0^2c = 1 \tag{3}$$

$$50c = 1 \tag{4}$$

$$c = \frac{1}{50} \tag{5}$$

If $c = \frac{1}{50}$, $\int f(x) = 1$ this also satisfies the non-negativity requirement ($f(x) \geq 0$).

8. Select All That Apply 1 point

Recall that the expression for the closed-form solution for ridge regression is $\hat{w} = (X^\top X + \lambda I)^{-1} X^\top y$, where $X \in \mathbb{R}^{n \times d}$ is the matrix of n data points each with d features; $y \in \mathbb{R}^n$ is the vector of n labels, one for each data point; and $\lambda \geq 0$ is the regularization constant. However, your friend who is taking a linear algebra class reminds you that this expression involves a matrix inverse, and not all matrices are invertible. Which of the following statements regarding this matrix inverse are true?

☐ a) X must be square for $(X^\top X + \lambda I)$ to be invertible.

- ⓑ $(X^\top X + \lambda I)$ is invertible if and only if $X^\top X$ is invertible.
- ⓒ $(X^\top X + \lambda I)$ is always invertible if $\lambda > 0$, since it ensures that $(X^\top X + \lambda I)u \neq 0$ for all vectors $u \in \mathbb{R}^d$ except for the vector of all zeros.
- ⓓ $X^\top X$ is always invertible because it is a symmetric positive semidefinite matrix. Therefore, $(X^\top X + \lambda I)$ is also invertible.

Solution: A is incorrect because the squareness of X has no bearing on the invertibility of $X^\top X + \lambda I$, since $X^\top X$ and λI will both always be square.

B is incorrect because adding λI allows for invertibility of $X^\top X + \lambda I$ regardless of if $X^\top X$ alone has a nontrivial nullspace.

C is correct because when $\lambda > 0$:

$$(X^\top X + \lambda I)u = (X^\top X)u + (\lambda I)u \quad (6)$$

$$(X^\top X)u + (\lambda I)u \geq (\lambda I)u \quad (7)$$

$$(\lambda I)u > 0 \quad (8)$$

$$\therefore (X^\top X + \lambda I)u > 0 \quad (9)$$

So

$$(X^\top X + \lambda I)u \neq 0, \forall u \in \mathbb{R}^d \setminus \{0\}$$

$X^\top X + \lambda I$ has an empty nullspace, and so is invertible.

D is incorrect because not all symmetric positive semidefinite matrices are invertible; only symmetric positive definite matrices are.

9. 1 point

Assume we have n training data points sampled from some population distribution, which we then use to train a model. Is leave-one-out cross validation (LOOCV) error (1) typically a slight underestimate, (2) an unbiased estimator, or (3) typically a slight overestimate of the true (population) error of the final model trained on all n data points? Briefly explain your answer.

Solution: It is typically a slight overestimate of the true error. This is because LOOCV uses a slightly smaller training size $n - 1$ in each fold, so the model performs slightly worse than the final model.

10. Consider the following function.

$$f(x) = \max \{-x, e^x - 1\}.$$

(a) One Answer 1 point

Is f convex on $\mathbb{R}$?

☒ **Yes**

☐ No

(b) 1 point

Give the interval of all subgradients of f at $x = 0$.

Recall that a vector $g \in \mathbb{R}^d$ is a subgradient at x if it satisfies

$$f(y) \geq f(x) + g^T(y - x) \text{ for all } y \in \mathbb{R}^d.$$

Hint: It might be helpful to visualize this function and its subgradients instead of trying to work through this formula.

Answer: [,]

Solution: Part (a). Yes, the max of convex functions is convex.

Part (b). $[-1, 1]$

11. Consider the following linear regression setting. Let $X \in \mathbb{R}^{n \times d}$, $\lambda \geq 0$, and suppose $y \in \mathbb{R}^n$ is drawn from a Gaussian, $y \sim \mathcal{N}(Xw, \sigma^2 I)$ for some $w \in \mathbb{R}^d$, $\sigma^2 > 0$. Let

$$\hat{w} = \arg \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2.$$

(a) 1 point

If $n < d$ and $\lambda = 0$, how many solutions does the above optimization problem have?

Answer: _____

(b) 1 point

If $n < d$ and $\lambda > 0$, how many solutions does the above optimization problem have?

Answer: _____

(c) 1 point

Fill in the blanks with **increases/decreases/does not change**. As σ^2 increases:

The bias of $\hat{w}$ _____

The variance of $\hat{w}$ _____

The irreducible error _____

Solution: (a) Infinite; problem is underspecified. (b) 1, ridge regression always has a unique solution.

(c) Bias remains unaffected. Variance and irreducible error increases.

12. One Answer 1 point

Consider gradient descent on a loss function with many local minima and one global minimum. What is a potential benefit of using (mini-batch) stochastic gradient descent small mini-batch size (e.g., 1), compared to full-batch gradient descent?

- (a) There is no potential benefit.
- (b) It guarantees that SGD will find the global minimum.
- (c) **It could lead to noisy gradients computed at each step, helping overcome a local minimum.**
- (d) It will provide more accurate gradient estimates across the dataset, and thus lead to quicker convergence.

Solution: Using SGD with a batch size of 1 will lead to very noisy updates, which may help overcome a local minima. B is false since using SGD never guarantees finding the global minimum. D is false since using SGD doesn't guarantee more accurate estimates, and also because using SGD will usually lead to slower convergence.

13. 1 point

You are training a logistic regression model for your job as a data scientist. You observe that both the training error and test error are very high, even though you have many features to represent your data and used a high regularization parameter (λ). Given this information, what is the most likely issue, and what could you change to improve model performance?

Solution: Issue is underfitting due to model being too simple; this is the most likely issue based on the high bias and high regularization. Possible solutions include increasing model complexity, decreasing regularization, and better feature engineering.

14. 1 point

We can think of using gradient descent for training a Logistic Regression model with Cross-Entropy loss as an analogy of a ball rolling down a hill. In this analogy, what do the ball's location (e.g., its x and y coordinates on a map) and its height (e.g., its z coordinate) represent?

Solution: The location represents the parameters of the model and the height represents the loss of the model with those parameters. So, the lowest point on this “loss-landscape” is the configuration of model parameters where loss is minimized.

15. 1 point

Given training data points $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ for each $i = 1, 2, \dots, n$, standard least-squares linear regression solves the following optimization problem:

$$\hat{w}_{\text{sq}} = \arg \min_{w \in \mathbb{R}^d} \sum_{i=1}^n (w^\top x_i - y_i)^2.$$

Consider instead solving the following *least-absolute-deviation* regression problem:

$$\hat{w}_{\text{ad}} = \arg \min_{w \in \mathbb{R}^d} \sum_{i=1}^n |w^\top x_i - y_i|.$$

Assume that the dataset contains a few noisy outliers, i.e., points where because of noise, our predicted $\hat{y}_i$ is very different from the observed y_i . In this setting, what is one potential advantage of least absolute deviation regression compared to least squares regression? Briefly explain your answer.

Solution: A potential advantage of least absolute deviation regression is that it is more robust to outliers than least squares regression. By squaring the errors, least squares regression gives disproportionately large weight to outliers. Least absolute deviation's linear penalty treats all errors in proportion to their magnitude, making the resulting model $\hat{w}_{\text{ad}}$ much less sensitive (i.e., more robust) to the presence of a few noisy outliers.

16. 1 point

A well-funded health-tech startup in Seattle creates an AI model to help dermatologists detect dangerous skin cancer from images of skin lesions. Their goal is to sell this AI as a diagnostic tool to hospital systems across the United States.

- **The Product:** An AI diagnostic tool that analyzes images of skin lesions and classifies them as either benign or potentially malignant.
- **Training Data:** The model is trained exclusively on a massive, high-quality dataset of anonymized images collected over 10 years from several large hospital networks in the **Seattle metropolitan area**.
- **Performance:** The model achieves 98% accuracy and a very high score when validated on a held-out test set of images from the **same** Seattle-area hospitals.

Provide one reason (related to this class) for why the company might not be successful deploying their service.

Solution: Example answer: images of skin types collected in seattle may not transfer to other areas in the US whether other skin types may be more common.

17. One Answer 1 point

You train an unregularized logistic regression model with two features, x_1 and x_2 . You observe that small perturbations¹ in the training data cause large fluctuations in the learned weights w_1 and w_2 , but the model's overall predictions on the training data remain almost unchanged.

¹Small perturbations are slight changes that do not systematically shift the underlying distribution (ie. changes from re-sampling from the same data-generating distribution or adding small, zero-mean noise to features/labels).

What best explains this behavior?

- (a) The model is underfitting due to too few parameters.
- (b) The optimization did not converge to the global minimum.
- (c) **The features contain redundant information, making the weight estimates unstable.**
- (d) The features have nonlinear interactions that logistic regression cannot capture.

Solution: When features are nearly linearly dependent, multiple weight combinations can yield almost identical predictions. This redundancy (multicollinearity) makes individual coefficients unstable—small data perturbations can flip their magnitudes or signs—while overall predictions remain consistent.

18. One Answer 1 point

(Continuation from question above) Suppose you instead fit the logistic regression model on the same data but with an L1 regularization penalty. What effect would you expect?

- (a) Lasso keeps both w_1 and w_2 nonzero, smoothly sharing weight across them.
- (b) Lasso makes x_1 and x_2 independent by altering $X^\top X$, so both w_1 and w_2 become stable together.
- (c) Lasso shrinks both w_1 and w_2 toward zero while keeping both coefficients nonzero.
- (d) **As the regularization coefficient grows, Lasso will tend to set one of w_1 or w_2 to zero and keep the other nonzero.**

Solution: The ℓ_1 penalty induces sparsity. With two correlated features, lasso typically keeps one and zeros the other.

19. You train a logistic regression classifier with features $x = (x_1, x_2)$, using

$$P(y = 1|x) = \sigma(w_0 + w_1x_1 + w_2x_2), \quad \text{where } \sigma(z) = \frac{1}{1 + e^{-z}}.$$

a) 1 point

Describe the geometric shape of the decision boundary in this feature space.

b) 1 point

Suppose you expand the model to include quadratic terms:

$$P(y = 1|x) = \sigma(w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2).$$

Describe how this changes the shape of the decision boundary, and what new kinds of patterns the model can capture.

c) 1 point

If your training accuracy improves but validation accuracy worsens after this expansion, what modification to your training objective could help, and why?

Solution: (a) In the original space, the decision boundary $w_0 + w_1x_1 + w_2x_2 = 0$ is a straight line. (b) Adding quadratic features maps the data into a higher-dimensional space, where the boundary becomes nonlinear (e.g., curved or elliptical). This increases model flexibility (lower bias) but also variance, making it more sensitive to noise. (c) Regularization (e.g., ℓ_2 penalty) discourages overly large coefficients, controlling variance and improving generalization.

20. 1 point For ridge regression with bias, we can write the objective function as

$$J(w, b) = \|Xw + b\mathbf{1} - y\|_2^2 + \lambda\|w\|_2^2,$$

where the parameters are $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, the data is $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$, and $\lambda \geq 0$ is the regularization constant. $\mathbf{1}$ refers to the vector of all 1s in $\mathbb{R}^n$.

Compute $\nabla_w J(w, b)$, the gradient with respect to w .

Answer: _____

Solution: Using matrix calculus,

$$\nabla_w J = 2X^\top(Xw + b\mathbf{1} - y) + 2\lambda w.$$

The first term is from differentiating the squared error, and the second term comes from the ℓ_2 regularization penalty.

21. **[Extra credit. Not attempting this question will not affect your grade.]** Consider linear regression on 2-dimensional data, where the features of the i -th data point are denoted $(x_1^{(i)}, x_2^{(i)}) \in \mathbb{R}^2$ and the target of the i -th data point is $y^{(i)} \in \mathbb{R}$. There are n training data points, $\{(x_1^{(i)}, x_2^{(i)}), y^{(i)}\}_{i=1}^n$. Assume that the following always holds in the training data, for all $i = 1, 2, \dots, n$:

$$\begin{aligned} x_2^{(i)} &= 2x_1^{(i)} \\ y^{(i)} &= 5x_1^{(i)}. \end{aligned}$$

1. 1 point Suppose we train a ridge regression model, i.e., we solve

$$\hat{w} = \arg \min_{(w_1, w_2) \in \mathbb{R}^2} \sum_{i=1}^n (w_1 x_1^{(i)} + w_2 x_2^{(i)} - y^{(i)})^2 + \lambda(w_1^2 + w_2^2)$$

with $\lambda > 0$. What is $\hat{w}$ in the limit as $\lambda \rightarrow 0$ from above? Hint: It is not necessary to formally take the limit; just assume λ is vanishingly small but positive.

$$\hat{w}_1 = \underline{\hspace{2cm}}$$

$$\hat{w}_2 = \underline{\hspace{2cm}}$$

2. 1 point Now suppose we train a Lasso regression model, i.e., we solve

$$\hat{w} = \arg \min_{(w_1, w_2) \in \mathbb{R}^2} \sum_{i=1}^n (w_1 x_1^{(i)} + w_2 x_2^{(i)} - y^{(i)})^2 + \lambda(|w_1| + |w_2|)$$

with $\lambda > 0$. What is $\hat{w}$ in the limit as $\lambda \rightarrow 0$ from above? Hint: It is not necessary to formally take the limit; just assume λ is vanishingly small but positive.

$\hat{w}_1 =$ _____

$\hat{w}_2 =$ _____

Solution: Because the data is collinear, note that any w that satisfies $w_1 + 2w_2 = 5$ will minimize the MSE term of the objective, because the MSE will be 0. There are an infinite number of such solutions. Since λ is positive but vanishingly small, we will always choose a solution that achieves zero MSE, because the MSE term will dominate in the objective compared to the regularization term. Thus, these problems are asking which of these infinite solutions will we pick with different regularizers.

For ridge regression, we solve $\min w_1^2 + w_2^2$ subject to $w_1 + 2w_2 = 5$. Setting the derivative of the objective to zero and plugging in $w_1 + 2w_2 = 5$ gives us $\hat{w}_1 = 1, \hat{w}_2 = 2$.

For lasso regression, we similarly obtain $\hat{w}_1 = 0, \hat{w}_2 = 2.5$. Note that lasso promotes sparsity by only selecting x_2 because intuitively, it gets more bang for its buck by increasing w_2 over w_1 .

ID: _____

Page 16

END OF EXAM

ID: _____

Page 17

EXTRA PAPER