

CSE 446/546 Winter 2026 Midterm Exam

February 9, 2026

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).

- For multiple choice questions, fill in the circle(s) to indicate your answers. Do not write your answer to the side. Example:



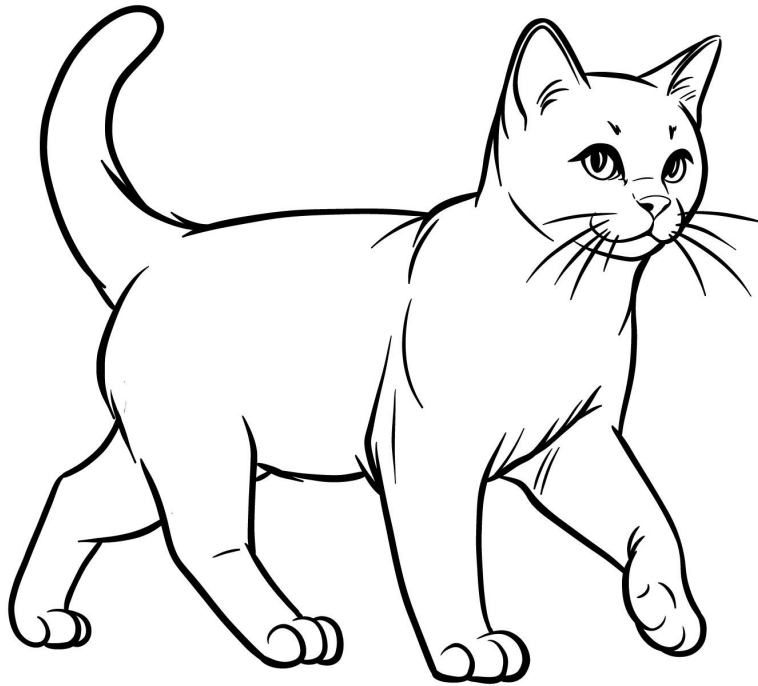
Selected answer.



Unselected answer.

- 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, we will give partial credit.
- 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, clearly indicate what you would like graded as your final answer. If we cannot determine your final answer, we cannot give credit.
- If you complete the exam before the last 5 minutes, you may turn in your exam to course staff. Otherwise, please remain in your seats for the last 5 minutes of the exam.
- If you need to use the restroom, please first leave your exam with an instructor at the front of the room. You can then collect and resume your exam when you return.

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



1. 2 points

Recall that $\|x\|_2 = \sqrt{x^\top x}$ is a norm over $\mathbb{R}^n$. Use the properties of norms from Homework 2 to answer the following questions.

Hint: a function g is a norm (over $\mathbb{R}^n$) if it satisfies all of the following:

- i. $g(x) \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $x = 0$*
- ii. $g(ax) = |a|g(x)$ for all $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$*
- iii. $g(x + y) \leq g(x) + g(y)$ for all $x, y \in \mathbb{R}^n$*

(a) Is $f(x) = \|x\|_2^2$ a norm? Explain why or why not.

(b) Is $g(x) = \|Ax\|_2$ a norm for any $A \in \mathbb{R}^{n \times n}$? Explain why or why not.

Solution: $\|x\|_2^2$ is not a norm as it does not satisfy (ii.) since for $a \in \mathbb{R}$ such that $|a| \neq 1$,

$$\begin{aligned}\|ax\|_2^2 &= a^2\|x\|_2^2 \\ &\neq |a|\|x\|_2^2.\end{aligned}$$

$\|x\|_2^2$ also does not satisfy (iii.) since for $x, y \in \mathbb{R}^n$ such that $2x^\top y > 0$,

$$\begin{aligned}\|x + y\|_2^2 &= \|x\|_2^2 + 2x^\top y + \|y\|_2^2 \\ &> \|x\|_2^2 + \|y\|_2^2.\end{aligned}$$

$\|Ax\|_2$ is not always a norm since it does not always satisfy (i.). A can have a non-trivial null space such that $Ax = 0$ for $x \neq 0$ (and therefore $\|Ax\|_2 = 0$).

2. 1 points Select All That Apply

Let $X_1, \dots, X_n$ be i.i.d. samples from a continuous distribution with density $p(x \mid \theta)$, where θ is an unknown parameter. Define the likelihood function $\ell(\theta) = \prod_{i=1}^n p(X_i \mid \theta)$ and the log-likelihood function $\log \ell(\theta) = \sum_{i=1}^n \log p(X_i \mid \theta)$.

Which statements about maximum likelihood estimation (MLE) are true?

- (a) **Any θ that maximizes $\ell(\theta)$ also maximizes $\log \ell(\theta)$.**
- (b) The value of the maximized $\ell(\theta)$ is equivalent to the value of the maximized $\log \ell(\theta)$.
- (c) $\hat{\theta}_{MLE}$ is always unbiased given any sample size n .
- (d) **For a Bernoulli model $X_i \sim \text{Bernoulli}(p)$ (X_i is 1 with probability p and 0 with probability $1 - p$), the MLE is $\hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^n X_i$.**
- (e) If the log-likelihood function is concave, then any θ that satisfies $\nabla_{\theta} \log \ell(\theta) = 0$ is guaranteed to be a global minimum.

Solution: Correct choices: A, D.

(A) True: $\log(\cdot)$ is strictly increasing, so maximizing the product likelihood is equivalent to maximizing the sum log-likelihood. (B) False: In general $\ell(\theta) \neq \log \ell(\theta)$ (C) False: MLEs can be biased in finite samples (e.g., variance in a Normal model with unknown mean). (D) True: maximizing $p^{\sum X_i} (1 - p)^{n - \sum X_i}$ gives $\hat{p} = \bar{X}$. (E) False: concavity implies critical points are global maxima (or part of a flat maximum set), not minima.

3. 1 points

Let $X_1, \dots, X_n$ be i.i.d. samples from an Exponential distribution with unknown rate $\lambda > 0$:

$$p(x | \lambda) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

(a) Write out the data log-likelihood.

(b) Derive the maximum likelihood estimate $\hat{\lambda}_{MLE}$

Solution: (A) In the exam, we told the students that any correct mathematical expression for the MLE is fine (it could be product or sum form).

The likelihood is

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^n X_i\right),$$

so

$$\ell(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n X_i.$$

(B) Differentiate:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n X_i.$$

Setting $\ell'(\lambda) = 0$ gives

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n X_i},$$

so $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n X_i$ is incorrect.

4. 1 points

You are modeling sensor readings using a linear model: $Y = Xw + \epsilon$. Your engineering team is considering the following noise models:

- **Model A:** $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- **Model B:** $\epsilon \sim \text{Laplace}(0, b)$

In lecture, we showed that MLE under **Model A** amounts to solving the optimization problem:

$$\hat{w} = \arg \min_w \|Y - Xw\|_2^2.$$

In section, we showed that MLE under **Model B** amounts to solving the optimization problem:

$$\hat{w} = \arg \min_w \|Y - Xw\|_1.$$

(a) One Answer

A sensor occasionally fails and reports extreme measurements. Which estimator is more robust to such outlier measurements?

☐ a) Model A

☒ b) **Model B**

(b) Why?

Solution: B. The Laplace (L1) estimator is more stable because its loss grows linearly with the residual magnitude, while the Gaussian (L2) loss grows quadratically. As a result, extreme outliers have disproportionately large influence under L2 but limited influence under L1.

5. 1 points

True or False: In linear regression, one of the reasons to add regularization is to ensure the loss function is convex. Explain your answer in one or two sentences.

Solution: Statement is False. The linear regression loss is already convex. It may not be *strictly* convex

6. 1 points One Answer

In linear regression, increasing the number of features will:

- ☐ (a) Always increase training error.
- ☐ (b) Always decrease training error.
- ☒ (c) **Never increase training error.**
- ☐ (d) Never decrease training error.
- ☐ (e) None of the above.

Solution: (C) Consider $X \in \mathbb{R}^d$, $y \in \mathbb{R}^d$. Let $\hat{w} = \arg \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2$. Suppose we add p more features to X , i.e. we obtain new data matrix

$$\tilde{X} = \begin{bmatrix} X & F \end{bmatrix} \in \mathbb{R}^{n \times (d+p)},$$

where $F \in \mathbb{R}^{n \times p}$ represents the newly added features. Set

$$\tilde{w} = \begin{bmatrix} \hat{w} \\ \mathbf{0}_p \end{bmatrix} \in \mathbb{R}^{d+p},$$

where $\mathbf{0}_p \in \mathbb{R}^p$ is the vector of all zeroes. Observe that

$$\begin{aligned} \min_{w \in \mathbb{R}^{d+p}} \|\tilde{X}w - y\|_2^2 &\leq \|\tilde{X}\tilde{w} - y\|_2^2 \\ &= \|X\hat{w} - y\|_2^2 \\ &= \min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2. \end{aligned}$$

7. 2 points

Consider a linear regression problem with an intercept term. You are given the following small dataset with labels y and data matrix X :

$$y = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 2 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

(note that the data matrix includes a prepended column of ones).

- (a) Compute the least-squares solution $\hat{w} = \begin{pmatrix} \hat{b} \\ \hat{m} \end{pmatrix}$ where $\hat{b}$ is the intercept and $\hat{m}$ is the slope.
- (b) Compute training mean squared error (MSE) over the training labels y and predicted labels $\hat{y} = X\hat{w}$.
- (c) Explain why the training MSE is not zero. (Hint: Look at how a few specific rows in X map to y).

Solution:

Part A:

First compute

$$X^\top X = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \Rightarrow (X^\top X)^{-1} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Next compute

$$X^\top y = \begin{pmatrix} 4 \\ 2 \end{pmatrix}.$$

Therefore,

$$\hat{w} = (X^\top X)^{-1} X^\top y = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

so $b = 1$ and $m = 1$.

Part B:

$$\hat{y} = Xw = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 2 \end{pmatrix}$$

$$r = y - \hat{y} = \begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}.$$

The sum of squared residuals is $0^2 + (-1)^2 + 1^2 + 0^2 = 2$, so the training MSE is $2/4 = 1/2$.

Part C:

Several potential correct answers to why MSE is not zero.

- Simple intuitive answer - if you observe the 2nd and 3rd row in X , the same vector maps to different y . It is impossible to fit such data points exactly and hence, MSE $\neq 0$.
- There are two data samples with the same x and different y . Thus the irreducible error is non-zero, and the MSE (which includes the irreducible error) must also be non-zero.

8. 1 points

You are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}$. You decide to model this data using the function:

$$f(x_i) = w_0 + w_1 \sin(x_i) + w_2 x_i^2 + w_3 e^{x_i}$$

- (a) Is this a linear regression model? Briefly explain why or why not in 1-2 sentences.
- (b) Let the matrix H and weight vector w be such that the model can be written in the matrix form $\hat{y} = Hw$. Explicitly write out the contents of the i -th row of H and the vector w .

$$H_i =$$

$$w =$$

Solution:

- a. Yes it is a linear regression model. The model is linear in its parameters. The non-linearity of the features do not affect the linearity of the regression.
- b.

$$H_i = [1 \quad \sin(x_i) \quad x_i^2 \quad e^{x_i}]$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

9. 1 points Select All That Apply

You are given a dataset of input-output pairs $(x_1, y_1), \dots, (x_n, y_n)$ where each $x_i, y_i \in \mathbb{R}$. You wish to model the relationship between x and y using Linear Regression (Ordinary Least Squares). You are allowed to transform the inputs x using any fixed basis function

$$h(x) = [h_1(x), \dots, h_k(x)]^\top \in \mathbb{R}^k$$

before learning the weights. Which of the following regression models **can** be learned using the standard closed-form solution $\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$, where X is defined as

$$X = \begin{bmatrix} h(x_1)^\top \\ \vdots \\ h(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n \times k},$$

for some choice of h ?

- ☐ **a** $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$
- ☐ **b** $y = w_0 + e^{w_1 x}$
- ☐ **c** $y = w_0 + w_1 \sin(x) + w_2 \cos(x)$
- ☐ **d** $y = w_0 + w_1 e^x + w_2 \log(x)$

Solution: All TRUE except for Option b. The weight w_1 is in the exponent—the model is not linear with respect to the parameters. The rest of the models are linear in parameters, making them solvable using the closed-form solution.

10. 2 points

Recall the ridge regression estimator $\hat{w}_{\text{ridge}} = (X^\top X + \lambda I_d)^{-1} X^\top y$, where $X \in \mathbb{R}^{n \times d}$ is the data matrix with n data points, each consisting of d features. Assume $X^\top X$ has eigenvalues $\mu_1 \geq \mu_2 \geq \dots \geq \mu_d \geq 0$.

(a) Select All That Apply Which of the following are True in the setting above?

- ☐ (a) $(X^\top X)$ is guaranteed to be invertible.
- ☐ (b) $(X^\top X) \succeq 0$
- ☐ (c) When $\lambda > 0$, $X^\top X + \lambda I_d$ is guaranteed to be invertible.
- ☐ (d) When $\lambda > 0$, $X^\top X + \lambda I_d \succeq 0$

(b) One Answer What are the eigenvalues of the regularized matrix $X^\top X + \lambda I_d$?

- ☐ (a) $\lambda \mu_i$ for $i = 1, \dots, d$
- ☐ (b) $\mu_i + \lambda$ for $i = 1, \dots, d$
- ☐ (c) $\lambda \mu_i + 1$ for $i = 1, \dots, d$
- ☐ (d) μ_i / λ for $i = 1, \dots, d$

Solution: Part (a): $X^\top X$ is not guaranteed to be invertible, but it has non-negative eigenvalues $\implies$ matrix is positive semi-definite. Hence, option (b) is correct. $X^\top X + \lambda I_d$ is always invertible and positive definite. Since all positive definite matrices are also positive semi-definite, (c) and (d) are correct.

Part (b): Let $v_i \neq 0$ be an eigenvector of $X^\top X$ with eigenvalue μ_i , so $(X^\top X)v_i = \mu_i v_i$. Then $(X^\top X + \lambda I_d)v_i = (X^\top X)v_i + \lambda v_i = (\mu_i + \lambda)v_i$. Thus the eigenvalues of $X^\top X + \lambda I_d$ are $\mu_i + \lambda$.

11. 1 points

You are fitting a linear model with many similar (correlated and similar magnitude) features and relatively few samples. Hyperparameters are chosen via cross-validation. You train both ridge and lasso regression with a large regularization parameter λ .

(a) How will ridge and lasso distribute weight across these features differently?

- (b) Which solution $\hat{w}_{\text{ridge}}$ or $\hat{w}_{\text{LASSO}}$ is more stable (i.e. lower variance)? In other words, if we were to sample a fresh dataset of the same size, which solution would change less? Explain.

Solution: (a) Ridge shrinks all coefficients and distributes weight across correlated features, while lasso tends to select a single feature and set the other correlates to zero.

(b) The ridge regression solution will be more stable. With highly similar features, LASSO will tend to select either one or the other similar feature, so weights will vary significantly across new datasets.

12. 1 points Select All That Apply

Consider the ridge regression problem. Let $X \in \mathbb{R}^{n \times d}$ be the data matrix and $y \in \mathbb{R}^n$ be the label vector. The ridge objective is

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2,$$

where $\lambda \geq 0$ is the regularization parameter.

Which of the following statements are true?

- ☐ a The objective is convex when $\lambda > 0$.
- ☐ b The objective is convex when $\lambda = 0$.
- ☐ c The solution is unique when $\lambda > 0$ and $d > n$.
- ☐ d The solution is unique when $\lambda > 0$ and $n > d$.
- ☐ e The solution is more likely to be sparse when $\lambda > 0$.

Solution: A, B, C, D. The least squares objective $\|y - Xw\|_2^2$ is a quadratic function in w with Hessian $2X^\top X$. Since $X^\top X$ is always positive semidefinite, the least squares objective is convex regardless of the dimensions of X . However, if $X^\top X$ is singular, the objective is not strictly convex and may have multiple minimizers. For $\lambda > 0$, the total Hessian becomes $2(X^\top X + \lambda I)$, which is positive definite. Consequently, the objective is strongly convex with a unique minimizer, even when $d > n$. Sparsity is typically induced by ℓ_1 .

13. 1 points You are modeling delivery time in minutes for an online grocery service using squared loss. Features include distance, time of day, driver experience, and number of items. You observe the following behavior as model complexity increases:
- Training error decreases monotonically.
 - Test error decreases at first, then increases.
 - Even the best-performing model has high test error in regions with heavy traffic.
- (a) Explain why the model has high test error in regions with heavy traffic (third bullet), even with a highly flexible model and an extremely large dataset. Your answer should reference the conditional distribution of Y given X .
- (b) Suppose you could collect infinite training data, but must keep the same feature set. Which components of the prediction error would go to zero, and which would remain non-zero? Explain using the bias–variance decomposition.

Solution: (a) The test error has a lower bound given by the irreducible error

$$\mathbb{E}[(Y - \mathbb{E}[Y | X])^2 | X],$$

which arises from randomness in the conditional distribution of Y given X (e.g., unpredictable traffic incidents). This error cannot be reduced by collecting more data or using a more complex model as long as the feature set is fixed.

(b) With infinite data, the variance goes to zero because the learned predictor converges to its expected value. Bias remains if the model class cannot represent the true conditional mean, and irreducible error remains non-zero.

14. 1.5 points

You're building a regression model for data generated by

$$y = f(x) + \varepsilon, \quad f(x) = 2x, \quad \varepsilon \sim \mathcal{N}(0, 1).$$

Consider the point $x = 1$, so that $f(1) = 2$. After training the model four times on four different datasets with the same data generation process, you obtain the fitted predictors $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, and $\hat{f}_4$. At $x = 1$, $\hat{f}_1(1) = 2$, $\hat{f}_2(1) = 2.2$, $\hat{f}_3(1) = 2.4$, and $\hat{f}_4(1) = 2.2$.

We want to understand the bias-variance trade-off of this model. Using the decomposition of expected test mean squared error (MSE), estimate the following model properties:

(a) Bias²

(b) Variance

(c) Irreducible error

Solution: Note: it's okay if students do not simplify their results.

$$\mathbb{E}[\hat{f}(1)] = \frac{2+2.2+2.4+2.2}{4} = 2.2$$

$$\text{(a) Bias}^2: \left(\mathbb{E}[\hat{f}(1)] - f(1) \right)^2 = (2.2 - 2)^2 = (0.2)^2 = 0.04 \quad \text{(b) Variance: } \mathbb{E} \left[\left(\hat{f}(1) - \mathbb{E}[\hat{f}(1)] \right)^2 \right] = \frac{1}{4} ((2 - 2.2)^2 + (2.2 - 2.2)^2 + (2.4 - 2.2)^2 + (2.2 - 2.2)^2) = \frac{1}{4} (0.04 + 0 + 0.04 + 0) = 0.02$$

(c) Irreducible Error: Variance of ε , which is 1

15. 2.5 points

You're overseeing a machine learning project to classify rare heart murmurs from stethoscope recordings, and there are two different models to examine. Model A is trained on a few manually selected features, including heart rate and volume. Meanwhile, Model B is trained on a large feature set extracted from the raw audio.

- (a) Which of the models would you expect to have higher bias? What about higher variance? Irreducible error? For each source of error, circle the correct option and give a brief justification.

Higher Bias: Model A Model B Neither

Higher Variance: Model A Model B Neither

Higher Irreducible Error: Model A Model B Neither

- (b) For the model with higher bias, what are some strategies that could decrease bias? For the model with higher variance, what strategies would decrease variance?

To Decrease Bias:

To Decrease Variance:

Solution:

- (a) Model A: Higher bias, since the simple features likely underfit the ground truth. Lower variance, because the lack of flexibility ensures predictions will not change much between datasets. Irreducible error is caused by external noise, so is equal between both models.
Model B: Lower bias, since training on richer features will better capture the classification signal. Higher variance, because more features leads to more sensitivity to dataset variations. Irreducible error is the same as Model A.
- (b) Model A can train on higher-dimensional features or use a more complex model. Model B (if overfitting becomes a problem) can use regularization techniques, more training data, or cross-validation. Irreducible error cannot be eliminated.

Solution: The optimal predictor is $\mathbb{E}[Y|X] = f(x)$. It has true MSE of 1.

16. 2 points

You are training a model to predict whether a user will click an ad. Each data point corresponds to a single ad impression, and each user appears in many impressions. You have $n = 1,000,000$ total impressions but only 50,000 distinct users.

Your goal is to generalize to new users that do not appear in the training data.

- (a) Explain why randomly splitting individual impressions into train/validation/test sets can lead to overly optimistic performance.

- (b) Design a correct evaluation procedure. Your answer should specify: (i) how to split the data, (ii) how to perform cross-validation for hyperparameter tuning, and (iii) how to obtain the final test performance.

Solution: (a) If we split individual impressions at random, the same user will likely appear in both training and validation/test sets. The model can then exploit user-specific patterns (e.g., embeddings or historical click behavior) learned from training impressions to perform well on validation/test impressions from the same users. This creates data leakage and leads to overly optimistic performance estimates, especially when the goal is to generalize to new users.

(b) Use a group-based split where the unit of splitting is the user:

- Split the set of users into disjoint training, validation, and test user groups (e.g., 80/10/10). All impressions from a given user must belong to exactly one split.
- For hyperparameter tuning, perform k -fold cross-validation over users within the training set: partition training users into k folds, train on $k - 1$ folds of users and validate on the held-out fold of users. (All impressions of a user remain together.)

- After selecting hyperparameters, retrain the model on all training users (optionally training + validation users), and evaluate once on the held-out test users to report final performance.

17. 1 points

You've just joined a biotech startup that's building machine learning models to predict disease risk based on genomic data. On your first day, you're given two datasets. Dataset A contains 3.5 million anonymized patient records with hundreds of features each, collected over 10 years. Dataset B contains 500 samples from a rare disease study.

For each dataset, would you recommend using a single train-val split or k -fold cross-validation to evaluate model performance? Briefly justify your answer with two reasons per case.

(a) Dataset A:

(b) Dataset B:

Solution:

- (a) Single train-val split. (1) Cross-validation is computationally expensive due to the large dataset and number of features. (2) The variance in model evaluation is low due to the large dataset, so cross-validation is not as necessary since a train-test split still leaves millions of samples in the training set.
- (b) Cross-validation. (1) The dataset is very small, so cross-validation allows for better use of the limited data by having each data point used for both training and validation at some point. (2) A single train-val split might lead to misleadingly high or low performance results based on how the data was partitioned. Cross-validation averages results over multiple iterations, ensuring a more reliable evaluation.

18. 1 points Select All That Apply

You're given two datasets: $\mathcal{D}$ and $\mathcal{D}_{\text{test}}$, where $\mathcal{D}$ is 100 times larger than $\mathcal{D}_{\text{test}}$. You perform k -fold cross-validation with $\mathcal{D}$ and observe that the mean cross-validation accuracy is 92%, but the final test accuracy on $\mathcal{D}_{\text{test}}$ is 64%. Which of the following could be a reasonable explanation

for the discrepancy between these two accuracies?

- (a) The model is overfitting to the data distribution $\mathcal{D}$.
- (b) **The test set is too small.**
- (c) **There is a difference in distribution between $\mathcal{D}$ and $\mathcal{D}_{\text{test}}$.**
- (d) Cross-validation always overestimates performance.

Solution:

- (a) Incorrect, the validation accuracy is 92%, which means we're not overfitting to $\mathcal{D}$.
- (b) Correct, too small of a test set could mean the 62% could have wide confidence intervals.
- (c) Correct, the model will perform poorly if the test data is completely different from its training data.
- (d) Incorrect, cross-validation does not always overestimate performance.

19. 1 points Select All That Apply

You are training a machine learning model on a dataset of 1,000 samples and want to evaluate its generalization performance. Consider the statements below regarding cross-validation.

Which of the following statements are **TRUE**?

- ☐ a Using k-fold cross-validation with a large k (e.g., $k = n$) reduces bias in the estimate of test error but increases computational cost.
- ☐ b Cross-validation allows you to detect overfitting and select hyperparameters, but the estimate of error is still subject to variance due to the random splitting of the data.
- ☐ c Cross-validation can be applied to any supervised learning algorithm.
- ☐ d Cross-validation error is an unbiased estimate of future test error.

Solution: A (Correct) LOOCV ($k = n$) has low bias because each training set is nearly the full dataset, but it requires training the model n times, increasing computational cost.

B (Correct) Cross-validation provides a more accurate estimate of generalization error than a single train/test split, but the estimate still varies depending on random splits.

C (Correct) Cross-validation can be applied to any supervised learning algorithm.

D (Incorrect) Distribution shift in future test scenarios may change the error.

20. 1 points Select All That Apply

Consider logistic regression for binary classification with labels $y \in \{0, 1\}$. The model is defined by

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

Which of the following statements are true?

- ☐ a The predicted probability $P(y = 1 | x)$ is always between 0 and 1.
- ☐ b Increasing $x^\top w$ increases $P(y = 1 | x)$.
- ☐ c Classifying an example as $y = 1$ when $P(y = 1 | x) \geq 0.5$ is equivalent to checking whether $x^\top w \geq 0$.
- ☐ d The predicted probability $P(y = 1 | x)$ is a linear function of x .

Solution: (a) TRUE. The sigmoid function maps any real value to the interval $(0, 1)$

- (b) TRUE. The sigmoid function is monotonically increasing, so larger values of $x^\top w$ lead to higher predicted probabilities.
- (c) TRUE. Because $\sigma(0) = 0.5$, thresholding the probability at 0.5 is equivalent to thresholding the linear score $x^\top w$ at zero.
- (d) FALSE. Although $x^\top w$ is linear in x , applying the sigmoid makes the predicted probability a nonlinear function of x .

21. 1 points

Here is a list of machine learning tasks that we have learned so far:

- (a) Binary Classification
- (b) Multi-class classification
- (c) Regression

For each of the numbered scenarios below (1-4), select the single best machine task formulation from the list of options (A–C). Note: Options A–C may be used once, more than once, or not at all.

- (a) A bank wants to analyze a customer’s financial history to decide whether to “Approve” or “Reject” their loan application.

Task: _____

- (b) An agricultural tech company wants to predict the total yield of corn (in kilograms) for a specific farm based on rainfall and soil pH levels.

Task: _____

- (c) A tool takes an audio waveform of spoken speech and determines which specific word from a dictionary was spoken.

Task: _____

- (d) A security system analyzes a photo of a face and determines whether it matches a claimed employee’s ID photo.

Task: _____

Solution:

- (a) A. This is predicting between two classes.
- (b) C. This is predicting a continuous value.
- (c) B. This is predicting a word (1 class) from a dictionary of known words (k classes).
- (d) A. This is predicting between two classes: matches or doesn’t match.

22. 1 points Select All That Apply

Which of the following statements are **TRUE**?

- ☒ **a** $f(x) = |x|$ is convex

ⓑ A convex function can have multiple global minima.

Solution: (a) Using the triangular inequality definition of convexity, we can write $|\theta x + (1 - \theta)y| \leq \theta|x| + (1 - \theta)|y|$ for all x and y . (b) Convexity is not uniqueness. $f(x) = 0 \forall x$ is still convex. Intuitively, flat regions at the bottom are allowed in convex functions.

23. 1 points Select All That Apply

Which of the following statements are **TRUE**?

- ☒ (a) In stochastic gradient descent, the stochastic gradient is an unbiased estimator of the (empirical) full gradient.
- ☐ (b) Gradient descent always converges for any step size $\eta > 0$ as long as the function f is differentiable.

Solution: (a) Even though the gradient itself is noisy, under the uniform sampling assumption, we can show that $E[g(w)] = \nabla f(w)$ where $g(w)$ is the stochastic gradient and $\nabla f(w)$ is the true gradient. (b) If step size is too big, GD might oscillate or grow even for fully differentiable functions. Example: $f(x) = x^2$, $x_{t+1} = x_t - \eta * 2x_t$. If $\eta = 1$, $x_{t+1} = -x_t \implies$ oscillation.

24. 1 points

Show that $f(x) = e^{-x}$ is a convex function.

Solution: $\frac{d^2}{dx^2} e^{-x} = e^{-x} > 0$, so e^{-x} is convex.

25. 1 points

Is $f(x) = e^{-x^2}$ a convex function? Explain why or why not.

Solution: $f''(x) = e^{-x^2}(4x^2 - 2)$. Since $e^{-x^2} > 0$ for all x , the sign of $f''(x)$ is determined by $4x^2 - 2$, and $4x^2 - 2 \geq 0 \iff |x| \geq \frac{1}{\sqrt{2}}$. Thus $f''(x) < 0$ for $|x| < \frac{1}{\sqrt{2}}$ and $f''(x) > 0$ for $|x| > \frac{1}{\sqrt{2}}$. Hence e^{-x^2} is not convex on $\mathbb{R}$; it is concave near 0 and convex for large $|x|$ (bell shape curve).

26. 1 points

Consider minimizing a differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ using gradient descent. Starting from an initial point w_0 , gradient descent updates are given by $w_{t+1} = w_t - \eta \nabla f(w_t)$, where

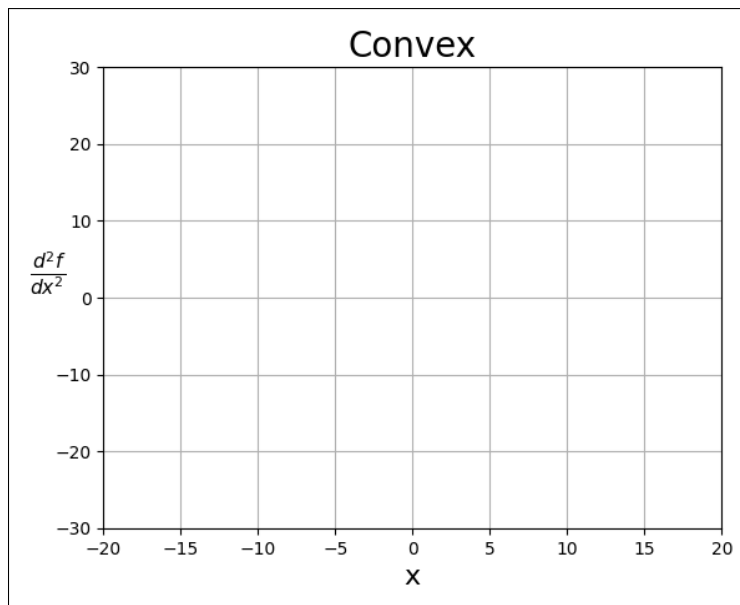
$\eta > 0$ is the step size (learning rate).

- (a) Under what conditions is gradient descent guaranteed to converge to a global minima?
- (b) Give an example setting in which increasing the learning rate leads to gradient descent converging in fewer iterations?
- (c) Give an example setting in which decreasing the learning rate leads to gradient descent converging in fewer iterations?

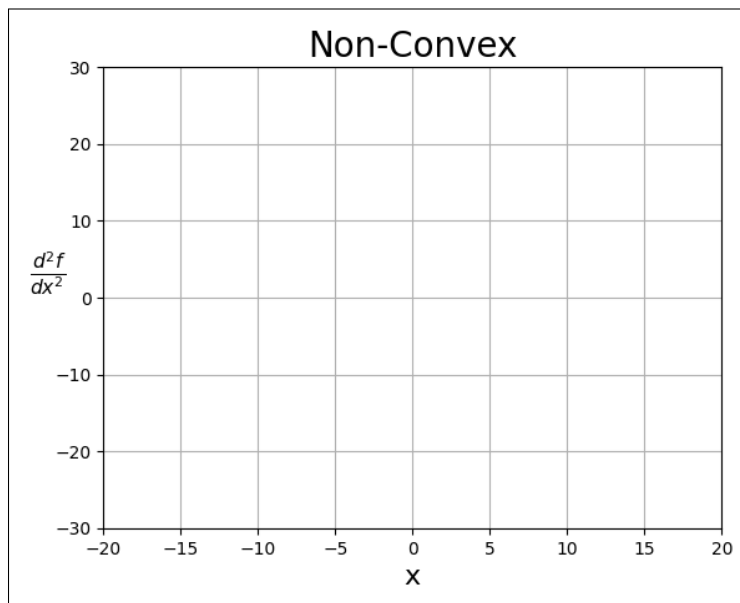
Solution: (a) The learning rate must be sufficiently small (or follow a decaying schedule) AND the function f must be *strongly* convex. E.g., minimizing an affine function with nonzero slope will never converge. (b) When the function is very smooth. (c) When the function is not very smooth. Here, the update can overshoot the minimum, causing the iterates to oscillate or diverge instead of converging.

27. 1 points

(a) Draw a function $\frac{d^2f}{dx^2}$ such that $f(x)$ is convex on $[-20, 20]$.



(b) Draw a function $\frac{d^2f}{dx^2}$ such that $f(x)$ is non-convex.



Solution: Plot for a : $\forall x \in [-20, 20], \frac{d^2 f}{dx^2}(x) \geq 0$. Plot for b : $\exists x \in [-20, 20], \frac{d^2 f}{dx^2}(x) < 0$

28. 2 points

In this question, we are going to compare gradient descent (GD) and stochastic gradient descent (SGD) and look at the benefits and drawbacks of each. For each objective below, select the best algorithm, and provide a brief explanation for your choice. Select “depends on context” if the better choice depends on the problem and/or context.

(a) **Computational efficiency of a single update step:**

GD SGD depends on context

Explanation:

(b) **Gradient accuracy (i.e. minimize deviation from the true gradient):**

GD SGD depends on context

Explanation:

(c) **Ability to escape local minima:**

GD SGD depends on context

Explanation:

(d) **Faster convergence (measured in time, *not* in number of iterations):**

GD SGD depends on context

Explanation:

Solution: SGD is better for computational efficiency, since it only needs to compute the gradient for one data point at a time.

GD generally provides more accurate gradient updates, since it uses the whole dataset for computing the gradient rather than a single data point (like SGD).

SGD is better suited for escaping a local minima, since computing the gradient based on a single data point leads to noisy updates, thus potentially helping overcome a local minima.

Faster convergence isn't a clear advantage for either, since it totally depends on the problem. Sometimes GD will lead to faster convergence, while in other cases SGD will lead to faster convergence. This may depend on your hardware (e.g., can you efficiently parallelize gradient calculations across data points?).

