CSE 446 Winter 2023 Final Exam

March 15, 2023

Please WAIT to open the exam until you are instructed to begin. You can write your name on this page.

Please write your name and ID on your notes page (if you have one). We will collect this with your exam.

Please take out your student ID and leave it on the corner of your desk, as we will come around and check them while you work on the exam.

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

- Write your name and ID number in the provided spaces on every page of the exam.
- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter associated with your choice.
- 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 and note sheet by handing them to a TA.

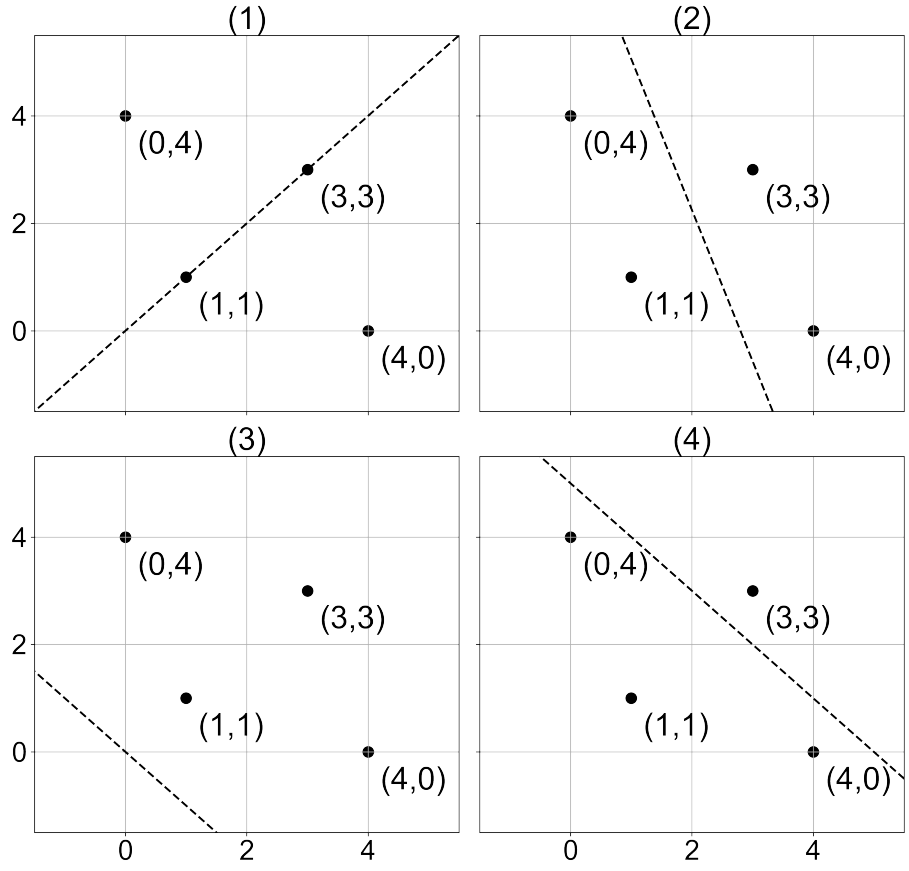
1. Suppose you have a data matrix $X \in \mathbb{R}^{n \times 10,000}$ and you want the 3 principal components of X. What is an efficient algorithm to compute these?

   Answer: _____

2. Suppose you have a data matrix $X \in \mathbb{R}^{10,000 \times 10,000}$ where $x_{ij} \sim_{iid} \mathcal{N}(0, \sigma^2)$ for each $i, j \in [10,000]$ and you want to understand how many principal components are needed to have reconstruction error $\leq 5/10,000$. What would be an efficient way to answer this question?

   Answer: _____

3. Consider the following scatter plots of a data matrix X with four data points in $\mathbb{R}^2$. Choose the plot whose line represents the direction of the first principal component of $X - \mu$, where $X \in \mathbb{R}^{n \times d}$ the vector $\mu \in \mathbb{R}^d$ is the featurewise mean of X.



(a) Plot 1

(b) Plot 2

(c) Plot 3

(d) Plot 4

4. In PCA, the following words go together (draw lines to match the words on the left with the words on the right)

Variance	Minimization
SVD	$A = USV^T$
Reconstruction error	Maximization
PageRank	Power Method

5. True/False: Given a set of points in a d-dimensional space, using PCA to reduce the dataset to $d' < d$ dimensions will **always** lead to loss of information.

   (a) True
   (b) False

6. Given a dataset X in a d-dimensional space, using PCA to project X onto $d_1 < d_2 < d$ dimensions leads to the d_1 dimensional projection to have higher _____ compared to the d_2-dimensional projection.

   Answer: _____

7. True/False: Given a dataset X in a d-dimensional space, using PCA to project X onto $d_1 < d_2 < d$ dimensions leads to the d_1 dimensional projection to being a subspace of the d_2-dimensional projection.

   (a) True
   (b) False

8. Consider a kernel matrix P that is given by $P_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ for a kernel map ϕ, inner product $\langle \cdot, \cdot \rangle$, and data samples $x_i, x_j \in \mathbb{R}^d$. Write the closed-form solution for the $\hat{\alpha}$ that minimizes the loss function $L(\alpha) = ||y - P\alpha||_2^2 + \lambda \alpha^T P \alpha$.

   Answer: $\hat{\alpha} =$ _____

9. Which of the following statements about kernels is/are **true**? Select **all** that apply.

   (a) A kernel feature map $\phi(x) : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ *always* maps to higher dimensional space (i.e., $k > d$).
   (b) Kernel matrices depend on the size of the dataset.
   (c) Kernel matrices are square.
   (d) Kernel matrices are used for data dimensionality reduction.

10. Fix a kernel K and corresponding feature map ϕ. True/False: One can train and evaluate a kernelized SVM (with this kernel) in polynomial time only if $\phi(x)$ runs in polynomial time for every x.

    Extra credit: explain your answer. _____

    _____

    (a) True
    (b) False

11. True/False: The number of clusters k is a hyperparameter for Lloyd's Algorithm for k-means clustering.
    (a) True
    (b) False

12. k-means refers to optimizing which of the following objectives? Here $\mu_{C(j)}$ is the mean of the cluster that x_j belongs to. m is the number of points.

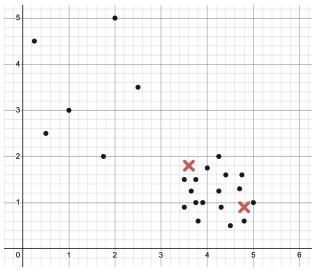
(a) $F(\mu, C) = \sum_{j=1}^{m} \|\mu_{C(j)} - x_j\|_2^2$

(b) $F(\mu, C) = \min_{j=1}^{m} \|\mu_{C(j)} - x_j\|_2^2$

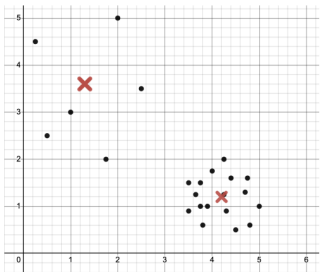
(c) $F(\mu, C) = \sum_{j=1}^{m} \|\mu_{C(j)} - x_j\|_2$

(d) $F(\mu, C) = \max_{j=1}^{m} \|\mu_{C(j)} - x_j\|_2^2$

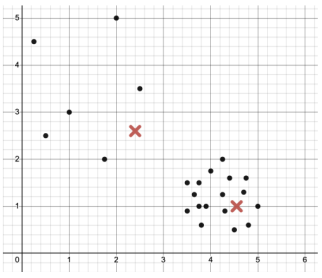
13. You are using Lloyd's algorithm (the algorithm described in class) to perform k-means clustering on a small dataset. The following figure depicts the data and cluster centers for an iteration of the algorithm. Dataset samples are denoted by markers and cluster centers are denoted by markers $\times$.



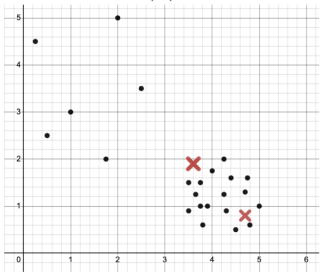
Which of the following depicts the *best* estimate of the cluster center positions after the next single iteration of Lloyd's algorithm? Hint: a single iteration refers to *both* update steps.



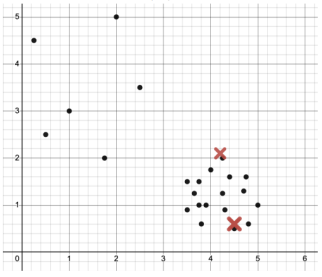
(A)



(B)



(C)



(D)

(a) Plot A

(b) Plot B

(c) Plot C

(d) Plot D

14. Consider a dataset X where row X_i corresponds to a complete medical record of an individual $i \in [n]$. Suppose the first column of X contains each patient's name, and no other column contains their name. True/False: Removing the first column from X gives a dataset $X_{\cdot,2:d}$ where no individual (row) is unique.

(a) True

(b) False

15. Suppose that a model finds that towns with more children tend to have higher rates of poverty compared to towns with fewer children. Upon seeing this, a local mayor suggests that children be banished from the town in order to reduce poverty. What is the flaw of this reasoning?

(a) The reasoning is correct.

(b) We cannot make policy decisions based on a machine learning model.

(c) Correlation does not imply equal causation.

16. What method can be described as a resampling method used to estimate population parameters by repeatedly sampling from a dataset?

(a) Power method

(b) Bootstrapping

(c) k-means

(d) SVD

17. True/False: The bootstrap method can be applied to both regression and classification questions.

(a) True

(b) False

18. Which of the following can be done to reduce a model's bias?

(a) Add more input features.
(b) Standardize/normalize the data.
(c) Add regularization.
(d) Collect more data.

19. For ridge regression, how will the bias and variance in our estimate $\hat{w}$ change as the number of training examples N increases? Assume the regularization parameter λ is fixed.

(a) $\downarrow$ bias, $\uparrow$ variance
(b) same bias, $\downarrow$ variance
(c) same bias, $\uparrow$ variance
(d) $\downarrow$ bias, $\downarrow$ variance
(e) same bias, same variance

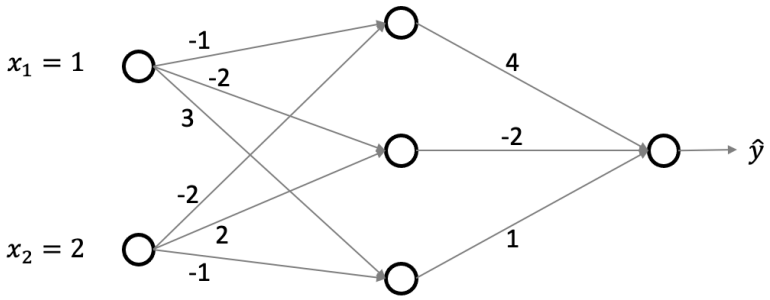
20. Let $\eta(X)$ be an unknown function relating random variables X and Y, D be a dataset consisting of sample pairs (x_i, y_i) drawn *iid* from the probability distribution P_{XY}, and $\hat{f}_D$ an estimator of η. Draw lines to match the expressions on the left with the words on the right.

$\mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$	Prediction error
$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2 \mid X = x]]$	Learning error
$\mathbb{E}[\mathbb{E}_D[(Y - \hat{f}_D(x))^2] \mid X = x]]$	Irreducible error

21. SVM models that use slack variables have _____ bias compared to SVM models that do not use slack variables (circle answer below).

(a) equal
(b) lower
(c) higher

22. Consider the following neural network with weights shown in the image below. Every hidden neuron uses the ReLU activation function, and there is no activation function on the output neuron. Assume there are no bias terms. What is the output of this network with the input $x = (1, 2)$? Give a numerical answer.



Answer: _____

23. Consider a neural network with 8 layers trained on a dataset of 800 samples with a batch size of 10. How many forward passes through the entire network are needed to train this model for 5 epochs?

Answer: _____

24. In neural networks, the activation functions sigmoid, ReLU, and tanh all

(a) always output values between 0 and 1.

(b) are applied only to the output units.

(c) are essential for learning non-linear decision boundaries.

(d) are needed to speed up the gradient computation during backpropagation (compared to not using activation functions at all).

25. While training a neural network for a classification task, you realize that there isn't a significant change to the weights of the first few layers between iterations. What could NOT be a reason for this?

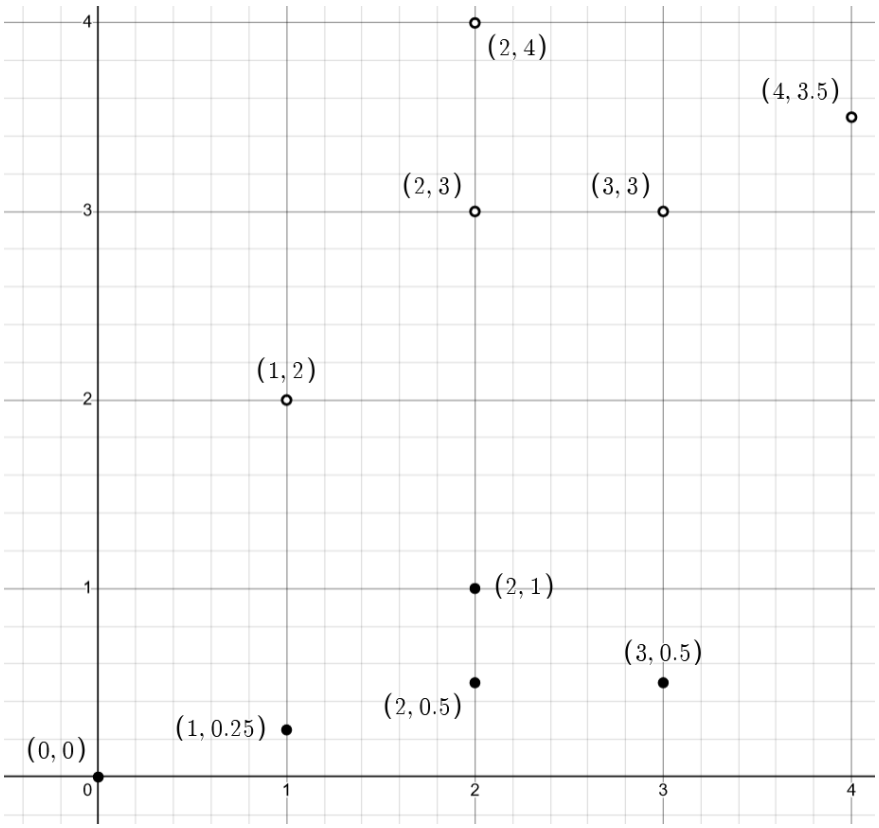
(a) The model is stuck in a local minimum.

(b) The network is very wide.

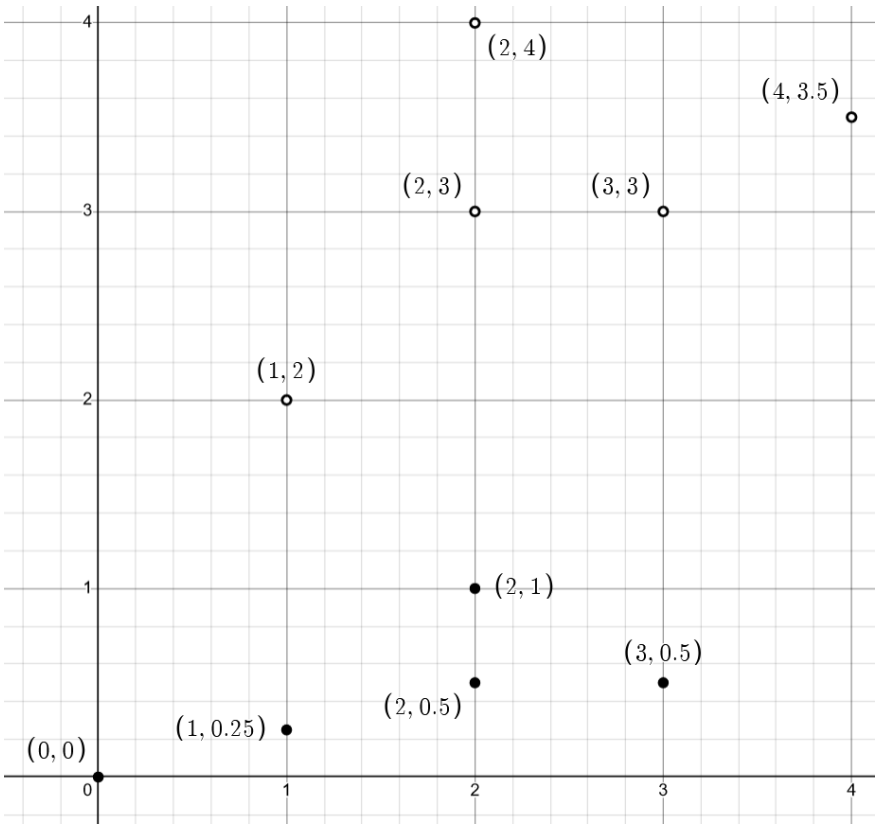
(c) The weights of the network are all zero.

(d) The learning rate is very small.

26. Shade in the region where decision boundaries that lie inside it have equal training error.



27. Draw the maximum margin separating boundary between the hollow and filled points.



28. What are support vectors in an SVM without slack?

(a) The data points that don't fall into a specific classification.

(b) The most important features in the dataset.

(c) The data points on the margin of the SVM.

(d) All points within the dataset are considered support vectors.

29. You have a batch of size N 256 x 256 RGB images as your input. The input tensor your neural network has the shape (N, 3, 256, 256). You pass your input through a convolutional layer like below:

```
Conv2d(in_channels=3, out_channels=28, kernel_size=9, stride=1, padding=1)
```

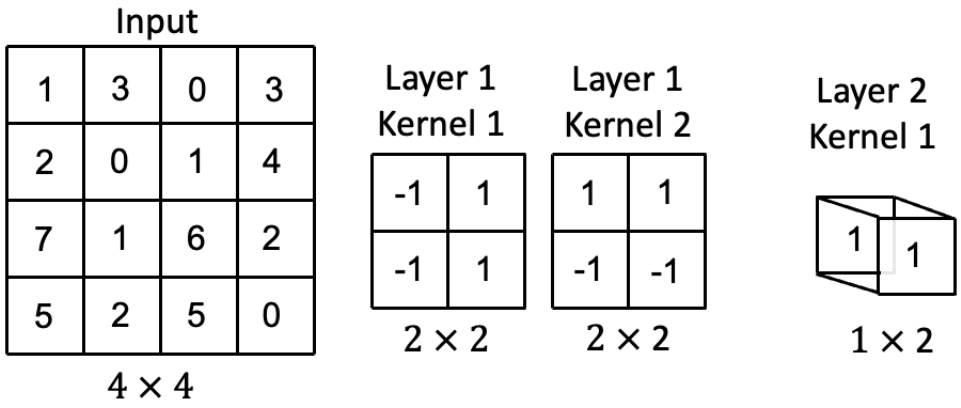
What is the shape of your output tensor?

Answer: (_____, _____, _____, _____)

30. Suppose that you have a convolutional neural network with the following components:

```
1. One 2D-convolutional layer with two 2x2 kernels, stride 2, and no zero-padding
2. A max pooling layer of size 2x2 with stride 2.
3. One 2D-convolutional layer with one 1x1 kernel, stride 1, and no zero-padding
```

Suppose you propagate the input below (left) through the CNN with the following kernel weights. Assume there are no bias terms.



What is the output of this network given the current weights and input?

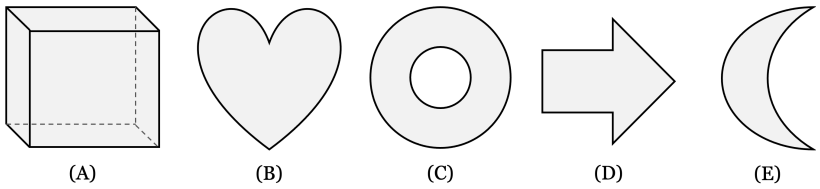
(a) 0

(b) 4.5

(c) 8

(d) 9

31. Which of the following shapes are convex? Select **all** that apply.



(a) Shape A.
(b) Shape B.
(c) Shape C.
(d) Shape D.
(e) Shape E.

32. Given differentiable functions $f(x) : \mathbb{R} \longrightarrow \mathbb{R}$ and $g(x) : \mathbb{R} \longrightarrow \mathbb{R}$, which of the following statements is false?

(a) if $-f(x)$ is concave, then $f(x)$ is convex.
(b) if $f(x)$ and $g(x)$ are convex, then $h(x) := \max(f(x), g(x))$ is also convex.
(c) if $f(x)$ and $g(x)$ are convex, then $h(x) := \min(f(x), g(x))$ is also convex.
(d) $f(x)$ can be both convex and concave on the same domain.

33. Which of the following loss functions are convex? Select **all** that apply.

(a) 1-0 loss.
(b) Squared loss (MSE).
(c) Sigmoid loss.
(d) Logistic loss.
(e) Hinge loss.

34. True/False: Both forward and backward passes are a part of the backpropagation algorithm.

    (a) True

    (b) False

35. Both LASSO and PCA can be used for feature selection. Which of the following statements are true? Select **all** that apply.

    (a) LASSO selects a subset (not necessarily a strict subset) of the original features

    (b) If you use the kernel trick, principal component analysis and LASSO are equivalent learning "techniques"

    (c) PCA produces features that are linear combinations of the original features

    (d) PCA is a supervised learning algorithm

36. Which of the following statements about choosing L1 regularization (LASSO) over L2 regularization (Ridge) are true? Select **all** that apply.

    (a) LASSO (L1) learns model weights faster than Ridge regression (L2).

    (b) L1 regularization can help us identify which features are important for a certain task.

    (c) L1 regularization usually achieves lower generalization error.

    (d) If the feature space is large, evaluating models trained with L1 regularization is more computationally efficient.

37. Which of the following techniques can be helpful in reducing the original dimensions of input data? Select **all** that apply.

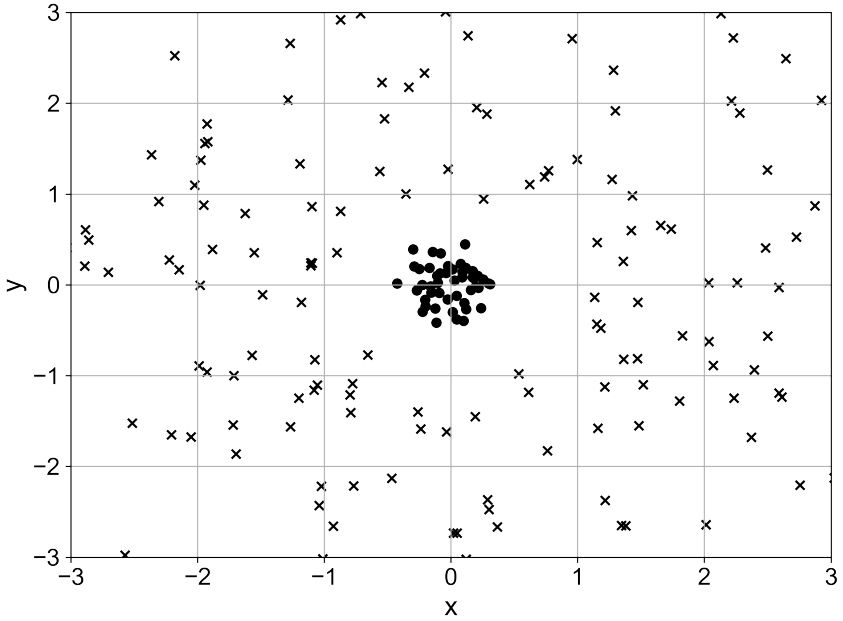
    (a) L1 Regularization (LASSO)

    (b) L2 Regularization (Ridge)

    (c) Principal Component Analysis (PCA)

    (d) k-means Clustering

38. Which of the following features could allow a logistic regression model to perfectly classify all data points in the following figure? Select **all** that apply.



(a) $|x_i|, |y_i|$

(b) $x_i + y_i, x_i - y_i$

(c) x_i^2, y_i^2

(d) x_i^3, y_i^3

39. What is the expression for logistic loss? Here $\hat{y}$ is a prediction, and y is the corresponding ground truth label.

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

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

(c) $1 + e^{-y\hat{y}}$

(d) $\log(1 + e^{y\hat{y}})$

40. The following expression for $\hat{\Theta}_2$ will appear twice in this exam. Consider a distribution X with unknown mean μ and variance σ^2. We define the population variance to be as follows

$$\hat{\Theta}_2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \hat{\Theta}_1)^2\right) \text{ for } \hat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Is $\hat{\Theta}_2$ unbiased?

(a) Yes

(b) No

41. The following expression for $\hat{\Theta}_2$ will appear twice in this exam. Consider a distribution X with unknown mean μ and variance σ^2. We define the population variance to be as follows

$$\hat{\Theta}_2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \hat{\Theta}_1)^2\right) \text{ for } \hat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

What is the expected value of Θ_2?

Answer: _____

42. What is the biggest advantage of k-fold cross-validation over Leave-one-out (LOO) cross-validation?

(a) It provides a more accurate estimation of model performance

(b) Prevents overfitting

(c) Easier to compute

(d) Minimizes impact from sample size

43. Consider a data matrix $X \in \mathbb{R}^{n \times d}$. What is the smallest upper bound on $\text{rank}(X)$ which holds for every X?

    Answer: $\text{rank}(X) \leq$ _____

44. Let A be an $n \times n$ matrix. Which of the following statements is true?

    (a) If A is invertible, then A^T is invertible

    (b) If A is PSD, then A is invertible

    (c) If A is symmetric, then A is invertible

    (d) None of these answers.

45. Let $A \in \mathbb{R}^{m \times m}$ and x in $\mathbb{R}^m$. What is $\nabla_x x^T A x$?

    Answer: $\nabla_x x^T A x =$ _____

1. **Extra Credit** Consider one of the "semi-fresh" datasets $\hat{X}$ generated using the bootstrap method for a dataset X, where n is large and $X_i \sim_{iid} \mathcal{D}$. Let f_X be the model trained on X. $err(f_X, \hat{X})$ is a/an _____ of $err_{\mathcal{D}}(f_X)$.

    (a) unbiased estimate

    (b) slightly biased upwards

    (c) slightly biased downwards

    (d) very biased estimate (either upwards or downwards), to the point where this value by itself is not useful.

2. **Extra credit:** Suppose that we have $x_1, x_2, \ldots, x_{\mathbf{2n}}$ are independent and identically distributed realizations from the Laplacian distribution, the density of which is described by

$$f(x \mid \theta) = \frac{1}{2} e^{-|x-\theta|}$$

Find the M.L.E of θ. Note that for this problem you may find the sign function useful, the definition of which is as follows

$$\mathsf{sign}(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

Answer: _____

3. **Extra credit:** Consider a nearest neighbor classifier that chooses the label for a test point to be the label of its nearest neighboring training example. What is its leave-one-out cross-validated error for the data in the following figure? ("+" and "−" indicate labels of the points).

Answer: _____