CSE 446/546 Autumn 2023 Final Exam

December 13, 2023

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

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).

- Write your name and UW NetID (<netID>@uw.edu) 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(s) 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.

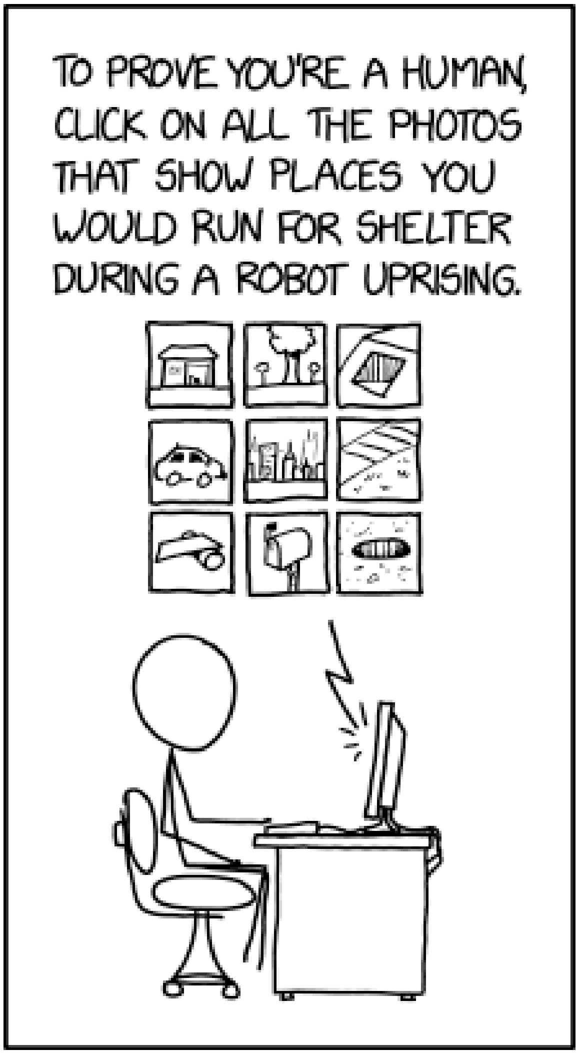

TO PROVE YOU'RE A HUMAN, CLICK ON ALL THE PHOTOS THAT SHOW PLACES YOU WOULD RUN FOR SHELTER DURING A ROBOT UPRISING.

Figure 1: This image is included only to cover the back of this page. It has no relation to the exam.

Beginning of Exam

This exam consists of 33 multiple choice questions, followed by 11 short answer questions. Each question is worth 1 point and each multiple choice question has one correct option, unless otherwise stated.

1. A fair six-sided die is rolled twice. What is the conditional probability that the first roll showed a 2, given that the sum of the two rolls is 6?

   (a) 1/6

   (b) 1/5

   (c) 3/11

   (d) 2/5

2. If matrix A has distinct eigenvalues, what can be said about its eigenvectors?

   (a) The eigenvectors form a linearly independent set

   (b) The eigenvalues are orthogonal to each other

   (c) A must be positive semi-definite

   (d) None of the above

3. In the context of multi-class logistic regression, which statement most accurately describes the decision boundaries?

   (a) They are linear and distinctly separate distinct classes.

   (b) They are non-linear and may overlap.

   (c) They remain unchanged, regardless of any transformations of the data.

   (d) They may be linear or non-linear, depending on the distribution of the data.

4. Which of the following is true about linear and logistic regression?

   (a) Both models output a probability distribution.

   (b) Both models are good choices for regression classes of problems.

   (c) Both models are good choices for classification.

5. Suppose you train a binary classifier in which the final two layers of your model are a ReLU activation followed by a sigmoid activation. How will this affect the domain of your final predictions?

   (a) This will cause all predictions to be positive.

   (b) This will have no effect on the distribution of predictions.

   (c) This will cause cause all predictions to be negative.

   (d) None of the above.

6. You are tasked with building a regression model to predict whether an email is spam [label=1] or not spam [label=0] based on various features. You are debating using linear or logistic regression. What type of regression is most suitable and why?

   (a) Linear regression, because it is optimized for learning the influence of multiple features.

   (b) Linear regression, because logistic regression cannot predict the comparative magnitude of the likelihood that an email is spam.

   (c) Logistic regression, because it models the probability of an instance belonging to a particular class

   (d) Logistic regression, because it allows for complex non-linear interactions between features and thus will be more accurate.

7. Which of the following matrices represents some kernel function $K : X \times X \to \mathbb{R}$ evaluated on two points?

   (a)

   $$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

   (b)

   $$\begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$$

   (c)

   $$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

   (d)

   $$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

8. Consider kernel ridge regression

   $$\hat{w} = \text{argmin}_w \frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top \phi(x_i))^2 + \lambda ||w||_2^2$$

   where $\phi : \mathbb{R}^d \to \mathbb{R}^q$ denotes the feature mapping and $d \neq q$, and $K_{i,j} := \langle \phi(x_i), \phi(x_j) \rangle$ denotes the entry (i, j) in the kernel matrix K. Which of the following statements are true? **Select all** that apply.

   (a) The optimal $\hat{w}$ is always a linear combination of x_i's for $i = 1, 2, ..., n$.

   (b) The optimal $\hat{\alpha}$ is $\hat{\alpha} = (KK^\top + \lambda I)^{-1} Y$.

   (c) The kernel method will still work even if the feature mapping is not one-to-one.

   (d) If K is positive semi-definite, then we can find a solution even when $\lambda = 0$.

9. The bootstrap method **cannot** be used to estimate the distribution of which of the following statistics?

(a) Mean

(b) Median

(c) Variance

(d) The bootstrap method can be applied to all of the above statistics.

10. True/False: Bootstrapping is a resampling technique that involves generating multiple datasets of size d by randomly sampling observations without replacement from the original dataset of size n (where $d << n$). True/False: Bootstrapping can be computationally prohibitive for large datasets.

(a) True, False

(b) True, True

(c) False, True

(d) False, False

11. Which of the following statements best describes the differences between Random Forests and Boosting in the context of decision tree-based ensemble methods?

(a) Random Forests and Boosting both reduce variance by averaging multiple deep decision trees, with no significant differences in their approach.

(b) In Random Forests, trees are built independently using bagging, while Boosting builds trees sequentially, with each tree learning from the errors of the previous ones.

(c) Boosting reduces bias by building shallow trees, whereas Random Forests use deep trees to address variance and do not focus on reducing bias.

(d) Both Random Forests and Boosting are identical in their handling of bias and variance, differing only in computational efficiency.

12. Which of the following statements is true about a single Decision Tree and Random Forest?

    (a) Random Forest has lower training error because it aggregates multiple trees

    (b) A good Random Forest is composed of decision trees that are highly correlated

    (c) Random Forest is useful because it's easy to explain how a decision is made

    (d) A single Decision Tree can result in comparably low training error in classification task compared to Random Forest

13. How is the performance of a distance-based machine learning model **typically** impacted when the data dimensionality is very high?

    (a) The performance significantly improves because there are more distinguishing features.

    (b) The performance decreases because the data points tend to appear equidistant in high-dimensional space.

    (c) The computational complexity of the distance calculations is reduced.

    (d) The performance remains unaffected as high-dimensionality uniformly impacts the positional relationships among the data points.

14. Which of the following is true about selecting k=1 for a k-nearest neighbors model of high dimensional data?

    (a) k=1 will make the model more sensitive to noise in the data.

    (b) k=1 will more accurately represent the real world distribution because it is more specific.

    (c) k=1 is a good option because it will lead to the highest number of different groupings, to match the high dimensionality of the data.

    (d) k=1 means that there will only be one grouping.

15. Which of the following is true about k-means clustering?

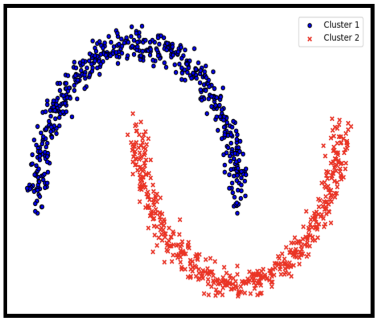
    (a) k-means diverges and is non-convex.

    (b) k-means diverges and is convex.

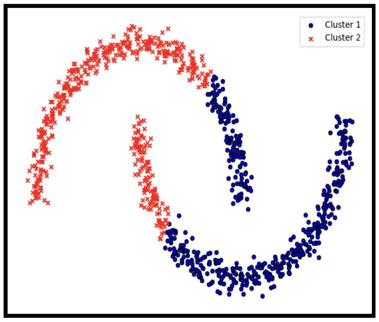
    (c) k-means converges and is non-convex.

    (d) k-means converges and is convex.

16. In which of the following plots are the points clustered by k-means clustering?



(a)



(b)

    (a) Plot (a)

    (b) Plot (b)

17. What is the main purpose of the softmax activation function in the output layer of a neural network?

    (a) To introduce non-linearity.

    (b) To normalize the output to represent probabilities.

    (c) To speed up convergence during training.

    (d) To prevent overfitting.

18. Suppose you have a fully-connected neural network with an input layer (3 neurons), one hidden layer (4 neurons), and an output layer (2 neurons). Each neuron (excluding those in the input layer) additionally includes a bias unit with its own weight. What is the total number of parameters in this network?

    (a) 9

    (b) 11

    (c) 24

    (d) 26

19. When training a neural network, which of the following will guarantee we do not overfit to the training data?

    (a) Normalize the data before training.

    (b) Increase the number of layers in our network until the final training loss stops decreasing.

    (c) Neither of the above

20. Given a simple two-layer neural network:

- Weights from input to hidden layer: $W^{(1)} = \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix}$, Bias for hidden layer: $[b_1^{(1)}, b_2^{(1)}]$, Activation function: $\sigma(z) = \frac{1}{1+e^{-z}}$
- Weights from hidden to output layer: $W^{(2)} = [w_1^{(2)}, w_2^{(2)}]$, Bias for output layer: $b^{(2)}$, Activation function: $\sigma(z) = \frac{1}{1+e^{-z}}$
- Target output: y; predicted output: $\hat{y}$
- Loss function: $\frac{1}{2}(y - \hat{y})^2$

After performing a forward pass with input $[x_1, x_2]$ and computing the loss, you execute a backward pass to calculate the gradients of the loss with respect to the weights and biases. What are the correct gradients for the weight, $w_{11}^{(1)}$, after one round of backpropagation?

Hint: Use chain rule to compute the gradients for $W^{(2)}$ and $W^{(1)}$. $\sigma'(z)$ is $\sigma(z) \cdot (1 - \sigma(z))$.

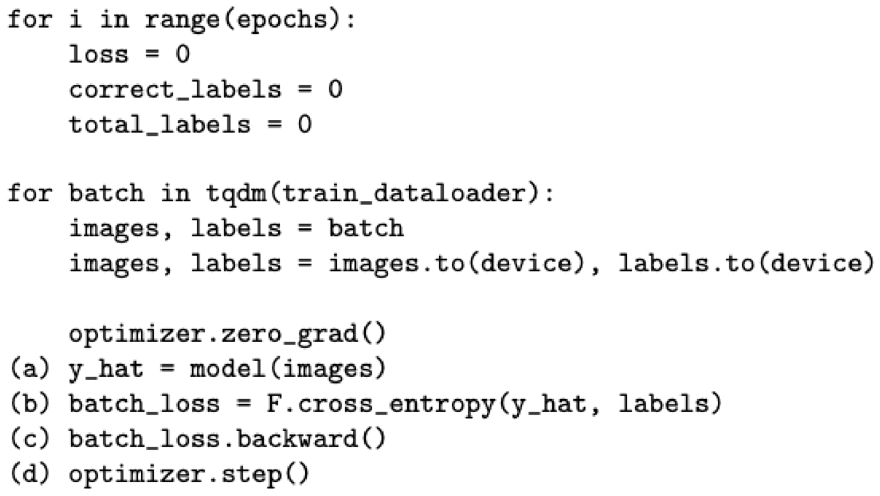
(a) $\frac{\partial \text{Loss}}{\partial w_{11}^{(1)}} = (y - \hat{y})^2 \cdot w_1^{(2)} \cdot \sigma'(z_1^{(1)}) \cdot x_1$

(b) $\frac{\partial \text{Loss}}{\partial w_{11}^{(1)}} = (y - \hat{y}) \cdot \hat{y} \cdot w_1^{(2)} \cdot \sigma'(z_1^{(1)}) \cdot x_1$

(c) $\frac{\partial \text{Loss}}{\partial w_{11}^{(1)}} = (y - \hat{y}) \cdot \hat{y} \cdot (1 - \hat{y}) \cdot w_1^{(2)} \cdot \sigma'(z_1^{(1)}) \cdot x_1$

(d) $\frac{\partial \text{Loss}}{\partial w_{11}^{(1)}} = (y - \hat{y}) \cdot x_1$

21. Which of the following statement is true about the following code snippet?



```
for i in range(epochs):
    loss = 0
    correct_labels = 0
    total_labels = 0

for batch in tqdm(train_dataloader):
    images, labels = batch
    images, labels = images.to(device), labels.to(device)

    optimizer.zero_grad()
(a) y_hat = model(images)
(b) batch_loss = F.cross_entropy(y_hat, labels)
(c) batch_loss.backward()
(d) optimizer.step()
```

(a) Step (a) completes the forward pass in backward propagation.

(b) Step (b) calculates the batch loss using a loss function that consists of its own trainable parameters, and weighs each sample differently based on those parameters.

(c) Step (c) never changes the weight parameters of any previous layer.

(d) Step (d) by itself performs the stochastic gradient descent by calculating the gradients and updating parameterized weights (you may assume we are using `torch.optim.SGD` for `optimizer`).

22. Which of the follow is true about using backpropagation to train a neural network using a package such as PyTorch or TensorFlow?

(a) You need to create a method that computes the gradient of each node of your neural network to call in the backpropagation step.

(b) Automatic differentiation executed by these packages takes advantage of the fact that the gradients of most functions can be pre-computed.

(c) The back-propagation executed by these packages is the process of computing the derivative of the nodes of a neural network starting with the first node at the beginning of the network and then proceeding to the next node(s).

(d) These packages fail on models with ReLU layers because the ReLU function is not differentiable everywhere, and thus the packages cannot execute backpropagation.

23. How is Singular Value Decomposition (SVD) typically utilized in image compression?

    (a) Selecting important pixels
    (b) Discarding low-rank components
    (c) Enhancing color information
    (d) Increasing image resolution

24. In the context of image processing, which of the following will directly impact the total number of trainable weights in a convolutional layer of a convolutional neural network (CNN)?

    (a) The resolution of the input image
    (b) The kernel size of the layer
    (c) The stride of the layer
    (d) The amount of padding used

25. What is the key advantage of using Gaussian Mixture Models (GMMs) over k-means clustering for data clustering tasks?

    (a) GMMs are computationally more efficient than k-means and are better suited for large datasets due to their simpler calculations.
    (b) GMMs, unlike k-means, can automatically determine the optimal number of clusters in a dataset without requiring this as an input parameter.
    (c) GMMs can model complex cluster shapes and densities, accommodating elliptical shapes, as they do not assume clusters to be spherical like k-means.
    (d) GMMs inherently handle missing data and noise better than k-means due to their probabilistic approach, which accounts for uncertainty in the data.

26. Which of the following statements are true? **Select all** that apply.

    (a) The sum of two convex functions is always convex.
    (b) The sum of two concave functions is always concave.
    (c) The sum of a convex and concave function is always concave.

27. Which of the following is **not** true about an arbitrary convex function $f : \mathbb{R} \to \mathbb{R}$ without any other assumptions? **Select all** that apply.

    (a) For all $x \in \mathbb{R}$, $f''(x) \geq 0$

    (b) The set
    $$\{(x, y) \in \mathbb{R}^2 \mid y \geq f(x)\}$$
    is convex

    (c) If c is a subgradient of f at x, then for all $y \in \mathbb{R}$:
    $$f(y) \geq f(x) + c(y - x)$$

    (d) f cannot be concave

28. Suppose $f(x) = ax^2 + bx + c$, where $a, b, c \in \mathbb{R}$. Which of the following statements are true about the convexity of f?

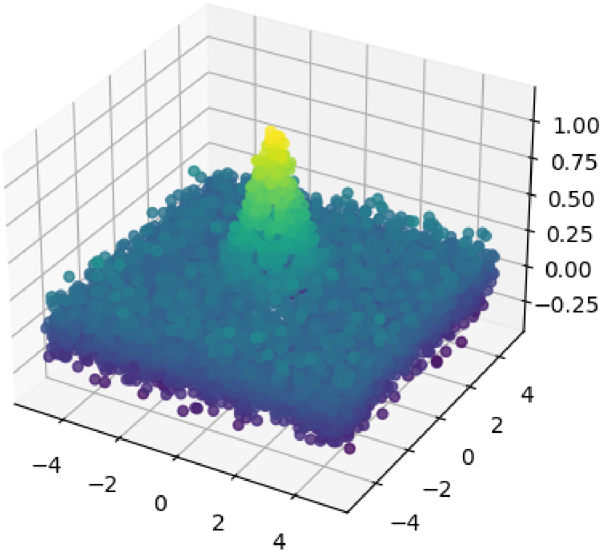
    (a) f is always convex since it is a polynomial.

    (b) f is convex only when $a > 0$, $b > 0$, and $c > 0$.

    (c) If $a > 0$ then f is convex.

    (d) If $a = 0$ then f is never convex.

29. Given this 3-D scatter plot, which of the following basis functions would you use for linear regression?



(a) $\phi(x, y) = \begin{bmatrix} 1 \\ x \\ y \\ xy \\ x^2 \\ y^2 \end{bmatrix}$

(b) $\phi(x, y) = \begin{bmatrix} e^{-x^2} \\ e^{-y^2} \\ e^{-(x^2+y^2)} \end{bmatrix}$

(c) $\phi(x, y) = \begin{bmatrix} \cos(x) \\ \cos(y) \end{bmatrix}$

(d) $\phi(x, y) = \begin{bmatrix} \sin(x) \\ \sin(y) \end{bmatrix}$

30. Suppose that we want to train a predictor $\hat{f}(x) = \hat{w}^\top x$ and we assume that $y = w^\top x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Which of the following statements about bias-variance tradeoff is true?

(a) (bias^2 + variance) is equal to the expected error between our trained predictor $\hat{f}(x)$ and the true data points (y's).

(b) Regularization is usually used to increase the variance of our trained predictor $\hat{f}(x)$.

(c) Irreducible error comes from the variance of the data points y's.

31. Consider a dataset $x_1, x_2, \ldots, x_n$ drawn from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with the density function given by $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Which of the following is true about the maximum likelihood estimation of μ and σ^2?

(a) The MLE for both μ and σ^2 cannot be determined without additional information.

(b) The MLE of μ is the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, but the MLE of σ^2 cannot be determined without additional information.

(c) The MLE of μ is the sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the MLE of σ^2 is the sample variance, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

(d) The MLE of μ is the sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the MLE of σ^2 is given by $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

32. In k-fold cross-validation, what is the primary advantage of setting k to a higher value (e.g., k=10) compared to a lower value (e.g., k=2)?

(a) It increases the accuracy of the model on unseen data.

(b) It provides a more reliable estimate of model performance.

(c) It reduces computational time.

(d) It eliminates the need for a separate test set.

33. Which of the following statements is true for ridge regression if the regularization parameter is too large?

    (a) The loss function will be the same as the ordinary least squares loss function.

    (b) The loss function will be the same as the Lasso regression loss function.

    (c) Large coefficients will not be penalized.

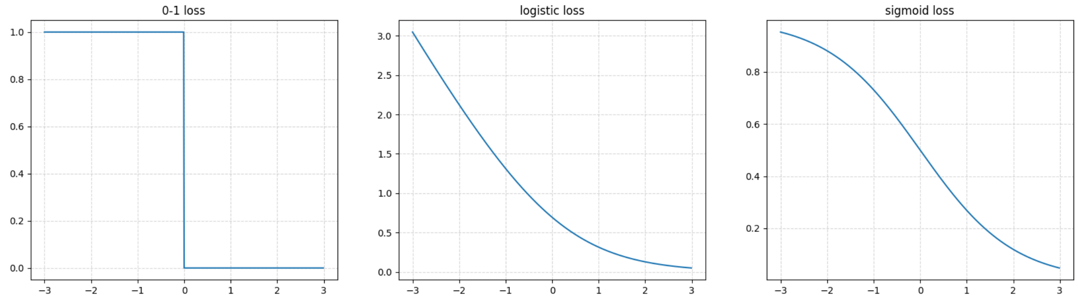
    (d) The model will overfit the data.

    (e) The model will underfit the data.

34. Consider a binary classification task, where $\hat{y}$ denotes the prediction and $y = +1$ or $y = -1$. Briefly describe the strength of minimizing logistic loss as opposed to 0-1 loss and sigmoid loss. The losses are formally defined as

$$\begin{aligned} \text{0-1 loss}(\hat{y}, y) &= \begin{cases} 0 & \text{if } \operatorname{sign}(y) = \operatorname{sign}(\hat{y}) \\ 1 & \text{otherwise} \end{cases} \\ \text{logistic loss}(\hat{y}, y) &= \log(1 + \exp(-y\hat{y})) \\ \text{sigmoid loss}(\hat{y}, y) &= \frac{1}{1 + \exp(y\hat{y})} \end{aligned}$$

The followings are example plots of each loss when $y = +1$.



Strength of logistic loss compared to 0-1 loss: ______________________________

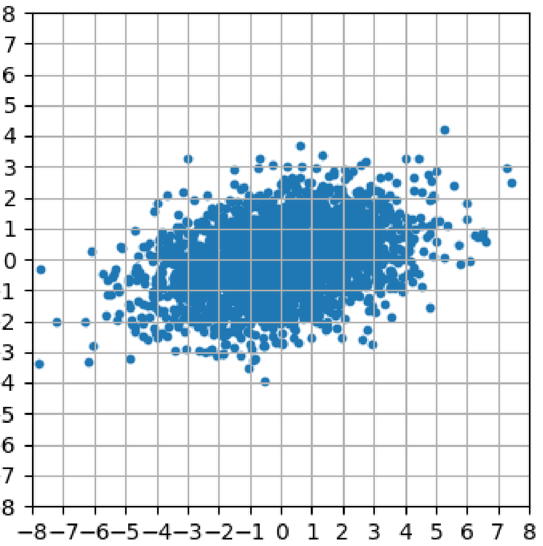
Strength of logistic loss compared to sigmoid loss: ______________________________

35. Name one advantage and one disadvantage of applying k-nearest neighbors for classification.

    Advantage: ____

    Disadvantage: ____

36. Consider the following scatter plot of a multivariate Gaussian distribution. Draw the approximate first and second principal components on the graph and say which is which.



First principal component:________________________

Second principal component:________________________

37. You are training a regression model to predict house prices. You decide to use zip-code as a feature in your model. Describe one possible problem with using zip-code as a feature in your model. [Note: zip-codes are numbers assigned to geographic regions. For example, going west from UW the zip-codes are 98105, 98103, then 98107.]

    Answer: ____

    ____

    ____

38. The real matrix A has the following singular value decomposition, $A = USV^\top$. Describe the structure of S and what S contains. (Note: for your answer you can consider this to be the compact or regular formulation of the SVD.)

    Answer: ____

    ____

    ____

39. Describe two advantages of using non-linear pooling layers, such as max-pooling, in a convolutional neural network (CNN).

    Answer: ____

    ____

    ____

40. What is the purpose of using multiple filters for a single convolutional layer in a neural network?

    Answer: ____

    ____

    ____

41. You are tasked with designing a convolutional neural network (CNN) to classify images as one of three classes; cat, dog, and other. You have a dataset of 10,000 images each to train and test your model, each with an image resolution of 1024 × 1024 pixels. You also have a dataset with the same 10,000 images, but with a downsampled image resolution of 16 × 16 pixels. What is one advantage and one disadvantage of using the dataset with lower resolution images?

    Answer: ____

42. You design a deep learning model. Describe one thing that could happen if you start training your model with too high of a learning rate:

    Answer: ____

43. You are creating a deep learning model. Why is it important to perform hyper-parameter tuning on a different set of data than the data that you used to test your model?

    Answer: ____

44. You design a machine learning model. In your own words, conceptually describe what the bias of your model means. [Note: this is asking about the theoretical bias of the model, NOT the social biases that may be influencing training dataset.]

    Answer: ____

    ____

    ____

This page is intentionally left blank as scratch paper.

This page is intentionally left blank as scratch paper.

Thank you for joining us for CSE 446/546 this quarter and we hope you enjoy your break!