CSE 446/546 Winter 2025 Final Exam

March 19, 2025

Name _____

UW NetID (not the numbers) _____

Please **wait** to open the exam until you are instructed to begin, and please take out your Husky Card and have it accessible when you turn in your exam.

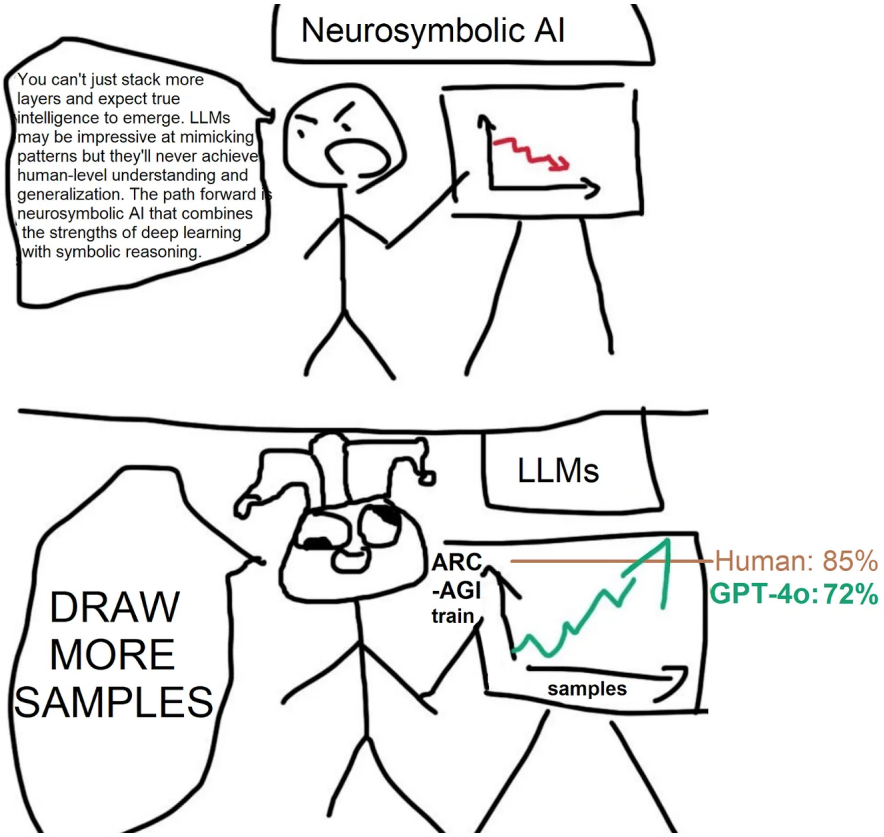
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

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

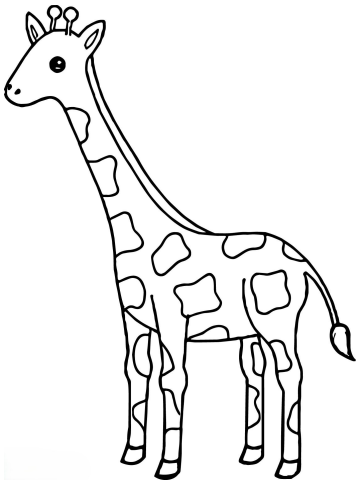
  Not selected answer: (a) Selected answer: (a)

- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- Multiple choice questions marked with [One Answer] should only be marked with one answer. All other multiple choice questions are [Select All That Apply], in which case any number of answers may be selected (**including none, one, or more**).
- For [Select All That Apply] questions, you will receive proportional credit for each option based on whether you get each "option" correct/incorrect. For example if there are 4 options, you will receive 0.25 points for each option that matches the solution.
- For each short answer question, please write your answer in the provided space.
- If you need to change an answer or run out of space, please very clearly indicate what your final answer is and what you would like graded. Responses where we cannot determine the selected option will be marked as incorrect.
- Please remain in your seats for the last 10 minutes of the exam. If you complete the exam before the last 10 minutes, you may turn in your exam to a TA.

These images are included only to cover the back of this page. They have no relation to the exam.
Comic from Natasha:



Giraffe for good luck :)



1. One Answer | 1 points

True/False: For a given model, irreducible error can be decreased by improving the model's complexity and increasing the amount of training data.

(a) True

(b) False

Correct answers: (b)

Explanation: You can't reduce irreducible error

2. One Answer | 1 points

You're training a classifier model using a neural network built from scratch in PyTorch. You are unable to decide on the depth of the neural network, so you decide to make the network as deep as possible. Despite achieving low training loss, your model performs poorly on the XOR test data. Why? Choose the most accurate explanation.

(a) The neural network is too complex and has too high of a bias squared error.

(b) The neural network is too complex and has too high of a variance error.

(c) The neural network is unable to learn non-linearities since XOR data is not linearly separable.

(d) We need to make the neural network even deeper to capture the complex relationship in the XOR dataset.

Correct answers: (b)

Explanation: The deep neural network is too complex and fits the training data too well (overfitting) resulting in a low bias squared error but fails to generalize as a result of high variance error.

3. 2 points

As dataset sizes increase, would you be more or less inclined to use leave-one-out cross-validation (LOOCV)? Provide reasoning to support your answer.

Answer: _____

_____

_____

Explanation: For larger datasets, leave-one-out cross validation becomes an extremely expensive process.

4. One Answer | 1 points

You are fine-tuning a model with parameters α, β, and γ, and have decided to perform 7-fold cross-validation to get the best set of hyperparameters. You have 5 candidate values for α, 3 candidate values for β, and 2 candidate values for γ. How many validation errors will you be calculating in total?

(a) Cannot be determined.
(b) 10
(c) 96
(d) 210
(e) 30

Correct answers: (d)

Explanation: 5 * 3 * 2 * 7 = 210

5. 3 points

You are analyzing the time until failure for a set of lightbulbs. The data represents the number of months each bulb lasted before failing and is given as follows: x_1, x_2, x_3, x_4. Assuming these times are modeled as being drawn from an exponential distribution. Derive the maximum likelihood estimate (MLE) of the rate parameter λ of this distribution. You must show your work.

Recall probability density function (PDF) for the exponential distribution is given by

$$f(x|\lambda) = \lambda e^{-\lambda x} \text{ for } x \geq 0$$

Hint: You should not have n in your final answer

Answer: $\lambda =$ _____

Explanation: The answer is $\boxed{\frac{4}{\sum_{i=1}^{4} x_i}}$.

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

$L(x_1, ..., x_n|\lambda) = P(x_1|\lambda) \cdot P(x_2|\lambda) \cdot ... \cdot P(x_n|\lambda) = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdot ... \cdot \lambda e^{-\lambda x_n}$

$= \lambda^n \cdot e^{-\lambda(x_1+x_2+...+x_n)}$

Now, we calculate the log-likelihood function:

$lnL(x_1, ..., x_n|\lambda) = ln(\lambda^n \cdot e^{-\lambda(x_1+x_2+...+x_n)}) = n \cdot ln(\lambda) - \lambda(x_1 + x_2 + ... + x_n)$

To find the argmax of λ (and thus the MLE) of this log-likehood expression, we need to take it's derivative with respect to λ and set it equal to 0.

$\frac{d}{d\lambda} lnL(x_1, ..., x_n|\lambda) = \frac{d}{d\lambda}(n \cdot ln(\lambda) - \lambda(x_1 + x_2 + ... + x_n)) = \frac{n}{\lambda} - (x_1 + x_2 + ... + x_n) = 0$

$\implies \lambda = \frac{n}{x_1+x_2+...+x_n} = \frac{n}{\Sigma_{i=1}^{n} x_i}$

Thus, the MLE here of λ is given by $\lambda = \frac{4}{\Sigma_{i=1}^{4} x_i}$.

6. Select All That Apply | 1 points

Which of the following can be convex?

(a) The intersection of non-convex sets
(b) The intersection of convex sets
(c) The union of non-convex sets
(d) The union of convex sets

Correct answers: (a), (b), (c), (d)

Explanation: The answers is all of them.

For the intersection of non-convex sets, the intersection of two five-pointed stars can be convex.
For the intersection of convex sets, just consider two circles that are on top of each other.
For the union of non-convex sets, just consider a circle that is split into two non-convex sets.
For the intersection of convex sets, just consider two circles that are on top of each other.

7. One Answer | 1 points

For convex optimization objectives, taking a gradient step using full-batch GD ensures that your loss shrinks.

(a) True
(b) False

Correct answers: (b)

Explanation: The answer is False.

Even for convex optimization objectives, if the learning rate is too high, there is a real probability of overshooting the global minima.

8. One Answer | 1 points

You are building a multi-class classifier using a deep neural network. You notice that your network is training slowly and that the gradients are diminishing quickly. Which activation function for the hidden layers of your network should you switch to, in order to avoid these issues?

(a) $f(x_i) = \frac{1}{1+e^{-x_i}}$

(b) $f(x_i) = \max(0, x_i)$

(c) $f(x_i) = x_i$

(d) $f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$

Correct answers: (b)

Explanation: Sigmoid ($f(x_i) = \frac{1}{1+e^{-x_i}}$) can cause vanishing gradients and hence can cause slow learning. ReLU ($f(x_i) = \max(0, x_i)$) avoids saturation. Having only linear layers reduces the network to a linear one. Softmax ($f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$) should be used in the output layer, but not the hidden layers of the network

9. One Answer | 1 points

If two neural networks differ only in the number of hidden layers, the deeper network will always achieve a lower training loss given the same training data.

(a) True

(b) False

Correct answers: (b)

10. Select All That Apply | 1 points

Snoopy is training a neural network to classify birds into "Woodstock" and "Not Woodstock". He has a plot of the training and validation accuracy for the neural network model during the training process.

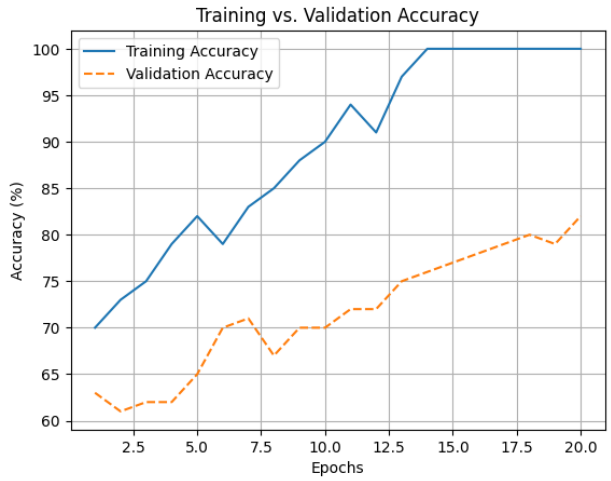


Figure 6: Snoopy's Training Plot

Which of the following actions could Snoopy take to help reduce the difference between training and validation accuracy?

(a) Increase the amount of training data

(b) Apply regularization techniques

(c) Reduce the complexity of the model (e.g., use fewer layers or units)

(d) Train for more epochs without making other changes

(e) Decrease the learning rate

Correct answers: (a), (b), (c)

11. One Answer | 1 points

Although both LASSO and PCA can be used for feature selection, they differ in their approach.

True/False: Specifically, LASSO sets some weight coefficients to 0 and selects a subset of the original features, whereas PCA selects features that minimize variance and creates a linear combinations of the original features.

(a) True
(b) False

Correct answers: (b)

Explanation: PCA selects features that capture the **most** variance, and produces a linear combination of the original features.

12. One Answer | 1 points

What is the minimization objective for logistic loss? Here $\hat{y}$ is the prediction, and y is the ground truth label.

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

Correct answers: (a)

Explanation: In this classification setting we are attempting to maximize the probability $P(y_i|x_i)$. Within the logistic regression problem setting, we set $P(y_i|x_i)$ to be equal to $\sigma(y_i w^\top x_i)$ where $\sigma(z)$ is the sigmoid function $\frac{1}{1+e^{-z}}$. If we are attempting to maximize the probability of $P(y_i|x_i)$, this is an equivalent objective to the minimization of the negative logprob. We therefore have the minimization objective become $-\log(\sigma(y_i w^\top x_i))$. Since $\hat{y}$ is our prediction, it is equivalent to $w^\top x_i$. Finally, using the definition of $\sigma(\cdot)$ and reducing $-\log(\sigma(y\hat{y}))$ gives us $\log(1 + e^{-y\hat{y}})$.

13. Select All That Apply | 1 points

The L-∞ norm is represented as $|| \cdot ||_\infty$ and is calculated for a vector $\mathbf{x} \in \mathbb{R}^d$ as $||\mathbf{x}||_\infty = \max_i(|x_i|)$. What happens to the parameters in w if we optimize for a standard linear regression objective with L-∞ regularization?

(a) There will be lots of parameters within w that are the same/similar absolute value.

(b) w will be very sparse.

(c) w will not be very sparse.

(d) Not enough information given to determine any of the above.

Correct answers: (a), (c)

Explanation: The L-∞ ball in parameter space is a square who's most protruding points are where the absolute values of the parameters are equivalent (corners of the square centered at the origin). Therefore A is correct. We know w will not be sparse because the protruding points of the L-∞ ball are not on the origin. Therefore C is also correct. Because A and C are correct, neither B nor D can be correct.

14. One Answer | 1 points

True/False: In k-means, increasing the value of k never worsens the model's performance on training data.

(a) True

(b) False

Correct answers: (a)

Explanation: Increasing k so that it is equal to n will make it so there is one cluster centroid per data point. This will perfectly fit the training data with zero training loss.

15. Select All That Apply | 1 points

Which of the following statements about PCA are true?

(a) The first principal component corresponds to the eigenvector of the covariance matrix with the smallest eigenvalue.

(b) If all the singular values are equal, PCA will not find a meaningful lower-dimensional representation.

(c) The principal components are the eigenvectors of the covariance matrix of the data.

(d) The reconstruction error of the recovered data points with a rank-q PCA strictly decreases as we increase q for all datasets.

Correct answers: (b), (c)

Explanation: A is false since the first principal component corresponds to the eigenvector with the largest eigenvalue. B is correct since if all the singular values are equal, the variance is equally distributed across all directions so PCA won't find a meaningful lower-dimensional representation. C is also correct since we find the eigenvalue decomposition of the covariance matrix. It isn't guaranteed that PCA will reduce the dimensionality, for example if all principal components are chosen.

16. Select All That Apply | 1 points

Consider the 2×2 matrix:

$$A = \begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix}$$

Let the Singular Value Decomposition (SVD) of A be given by:

$$A = U\Sigma V^T$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix containing the singular values of A. Which of the following statements are correct?

(a) The rank of A is 1.

(b) The nonzero singular value of A is 5.

(c) The columns of V must be $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

(d) The columns of V form an orthonormal basis for $\mathbb{R}^2$.

Correct answers: (a), (b), (d)

Explanation: (A) True: The rank of A is the number of nonzero singular values. Since the second row of A is entirely zero, its rank is **1**.

(B) True: The singular values of A are given by the square roots of the eigenvalues of A^TA:

$$A^TA = \begin{bmatrix} 9 & 12 \\ 12 & 16 \end{bmatrix}$$

The eigenvalues of this matrix are 25 and 0, so the singular values are $\sqrt{25} = 5$ and $\sqrt{0} = 0$.

(C) False: The columns of U are the **eigenvectors of AA^T**, not necessarily the standard basis vectors $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$.

(D) True: The matrix V is an orthogonal matrix, so its columns form an **orthonormal basis** for $\mathbb{R}^2$.

17. One Answer | 1 points

True/False: In kernel methods, we use a kernel function $k(x, x')$ to implicitly map input data into a feature space with different dimensions without explicitly computing the transformation. If we choose a linear kernel $k(x, x') = x^T x'$, then this is equivalent to mapping data into an infinite-dimensional feature space.

(a) True

(b) False

Correct answers: (b)

Explanation: Solution:

The statement is False because the linear kernel $k(x, x') = x^T x'$ does not map the data into an infinite-dimensional feature space. Instead, it corresponds to the original input space (i.e., the feature map $\phi(x)$ is simply x itself).

In contrast, nonlinear kernels, such as the Gaussian (RBF) kernel:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

correspond to an infinite-dimensional feature space. This is because an RBF kernel can be expressed as an infinite sum of polynomial terms in a Taylor expansion.

Thus, the error in the statement is that the linear kernel is incorrectly claimed to map data into an infinite-dimensional space, when in reality, it remains in the original finite-dimensional space.

18. Select All That Apply | 1 points

Which of the following statements about kernels is/are true?

(a) The kernel trick is a technique for computing the coordinates in a high-dimensional space.

(b) If the kernel matrix K is symmetric, it is always a valid kernel.

(c) Eigenvalues of a valid kernel matrix must always be non-negative.

(d) The kernel trick eliminates the need for regularization.

Correct answers: (c)

Explanation: (a) is false because the kernel trick's main benefit is that we do *not* have to compute in a high-dimensional space, ensuring efficiency. (b) is false because a kernel matrix must also be positive semi-definite. (c) is true because a valid kernel matrix must be positive semi-definite, meaning the eigenvalues will be non-negative. (d) is false because the kernel trick does not eliminate the need for regularization.

19. Suppose we are doing polynomial kernel regression with training dataset $X \in \mathbb{R}^{n \times d}$.

(a) [1 points]

Let $\mathbf{1} \in \mathbb{R}^n$ denote the vector of ones. Suppose we are using the polynomial kernel with degree up to 1, i.e., degree zero and degree one. Write the corresponding kernel matrix K in terms of X and $\mathbf{1}$.

$K =$ _____

(b) [1 points]

Now suppose we are using the polynomial kernel with degree up to k starting from degree zero. Let M be the corresponding kernel matrix. What is $M_{i,j}$ for row i and column j? Write your answer in terms of $K_{i,j}$.

$M_{i,j} =$ _____

Explanation: $K = XX^\top + \mathbf{1}\mathbf{1}^\top$. $M_{i,j} = (K_{i,j})^k$.
The computation of this matrix was done in the homework 3 (poly_kernel) using numpy.

20. Select All That Apply | 1 points

Which of the following statements about k-Nearest-Neighbors are true?

(a) The time complexity of the k-NN algorithm for a single query is $O(N \cdot d)$, where N is the number of training samples and d is the number of features.

(b) k-NN is highly efficient for large datasets because it has a low computational cost during the training phase.

(c) k-NN can suffer from the curse of dimensionality, where the effectiveness of the distance metric diminishes as the number of features increases.

(d) Scaling the features is crucial for k-NN performance, as it ensures that all features contribute equally to the distance computation.

(e) k-NN is inherently faster when the number of dimensions (features) is very high, because higher dimensions make the distance between data points more sparse.

Correct answers: (a), (c), (d)

Explanation: A single query involves iterating through all N data points and calculating a distance metric. Each distance calculation takes $O(d)$ time.
k-NN is not effecient for large datasets because N becomes infeasibly large.
The curse of dimensionality affects the distance metric of k-NN, making it less helpful in high-dimensional scenarios.
Scaling features is crucial because we want all the features to be the same scale for the distance calculation.
k-NN does not get faster as the dimensions of the data increases.

21. One Answer | 1 points

When choosing neural network architecture, we generally avoid overparameterization to prevent overfitting.

(a) True

(b) False

Correct answers: (b)

Explanation: In practice, overparameterized neural networks tend to generalize well, and overfitting is sometimes not entirely undesirable

22. One Answer | 1 points

When performing forward stagewise additive modeling, to compute a model at each iteration, we access:

(a) The most recently computed model
(b) The most recently computed ensemble
(c) All previously computed models
(d) All previously computed ensembles

Correct answers: (b)

Explanation: In forward stagewise additive modeling, at each iteration, the model accesses the most recently computed ensemble, which consists of the combination of all previous models.

23. Select All That Apply | 1 points

Select the following which is true for the K-means algorithm.

(a) The number of clusters (K) in K-means is a trainable parameter.
(b) The time complexity for running the K-means learning algorithm is agnostic to the number of data points.
(c) The time complexity for matching an unseen data point to k learned centroids is agnostic to the number of data points.
(d) K-means is a parametric model.
(e) K-means algorithm requires labeled data.
(f) K-means performs poorly on data with overlapping clusters.

Correct answers: (c), (d), (f)

Explanation: The number of cluster (K) is a hyperparameter and is not trained.
The time to learn k centroids scales with respect to the number of data points.
The time to match a data point to k centroids scales with respect to k.
The centroids in k-means are the learned "parameters". K-means is a Unsupervised Learning method which doesn't require explicit labeling of training data.
k-means performs poorly on overlapping clusters. GMMs are more suited for this problem.

24. Select All That Apply | 1 points

The following statements describe properties of K-means and Gaussian Mixture Models (GMM). Which of them are correct?

(a) K-means is a "hard clustering" method, while GMM is a "soft clustering" method.

(b) GMM can be used for both clustering and probability density estimation.

(c) Both GMM and K-means assume spherical/circular clusters.

(d) GMM cannot be used when clusters overlap significantly, as it assumes non-overlapping Gaussians.

(e) K-means is sensitive to the selection of initial centroids, which may lead to different clustering results.

Correct answers: (a), (b), (e)

Explanation: - K-means is a hard clustering method because each data point is assigned to exactly one cluster, while GMM is a soft clustering method where each point has a probability of belonging to multiple clusters.
- GMM is a probabilistic model that can be used not only for clustering but also for probability density estimation.
- GMM does not assume spherical clusters
- GMM can be used when clusters overlap. This is an advantage over K-means
- K-means is sensitive to the initial centroids chosen, and different initializations may lead to different clustering results.

25. 1 points

Suppose you are training a GMM with n Gaussians. How many parameters need to be learned?

Answer: _____

Explanation: $3n - 1$. We need to train n means μ, plus n covariances Σ, plus $n - 1$ mixing weights π. $3n$, $3n - 1$, $n(d^2 + d + 1)$ were accepted answers.

26. Select All That Apply | 1 points

Which of the following regarding bootstrapping are true?

(a) Bootstrapping is an approach for hyperparameter tuning.

(b) Bootstrapping can be applied to large datasets but is most accurate on small datasets.

(c) For a dataset of size N, there exists 2^N possible unique bootstrap datasets.

(d) Bootstrapping is commonly used to estimate the sampling distribution of a statistic, such as the mean or standard deviation, when the true distribution is unknown.

(e) Since we select N samples when creating the bootstrap dataset, each data point is guaranteed to be selected.

Correct answers: (d)

Explanation: Bootstrapping is used to calculate statistics of datasets (making option d correct), not for tuning hyperparameters. The representativeness of bootstrap statistics deteriorates as the size of the dataset decreases since the dataset is less representative of the true distribution. Since we randomly select N data points to create the bootstrapped dataset, there are multiple possible sets.

27. 2 points

Suppose you are the hiring manager at "Goggles" (a hypothetical tech giant) and you receive thousands of applicants for a role. Since you took CSE446, you decided to build a model and use past hiring data to automate the resume screening process, which has never been done before in the company. The dataset contains resumes and the labels are whether or not the owner of the resume passed the resume screening stage (previously done by humans). The benefit is two fold. You are able to distill the large pool of applicants quickly and you also eliminate human bias when screening resumes.
Explain why this approach can be problematic.

Answer: _____

Explanation: When the model is trained on biased data, the model can learn about the bias and perpetuate it, which doesn't eliminate human bias.

28. 2 points

Give an example of a task where we might expect a convolutional neural network to perform better than a deep neural network. Assume both models have roughly the same number of parameters.

Provide reasoning why the CNN might perform better in that setting.

Answer: _____

_____

_____

_____

_____

Explanation: Images are an example of a task. CNNs use shared parameters to learn filters that can be applied at any point in the image. So, if a cat occurs in the top left or top right corner, you can still recognize it.

29. Select All That Apply | 1 points

In the context of linear regression, general basis functions are used to:

(a) Minimize the computational complexity of linear regression models.

(b) Increase the speed of convergence in gradient descent optimization.

(c) Encourage sparsity in the learned weights.

(d) Transform the input data into a higher-dimensional space to capture non-linear relationships.

Correct answers: (d)

Explanation: A is incorrect because using basis functions with linear regression increases computational complexity. B is also wrong since basis functions don't directly affect the convergence rate. For C, sparsity isn't directly affected by using basis functions or not. D is the right answer as transforming input data to higher dimensional space is the exact purpose of basis functions.

30. Consider a neural network with 6 layers trained on a dataset of 600 samples with a batch size of 15.

a. [1 points]
How many forward passes through the entire network are needed to train this model for 8 epochs?

Answer: _____

b. [1 points]
How many forward passes through the entire network are needed to train this model for 5 epochs?

Answer: _____

Explanation: (a) 320
Since the batch size is 15, the number of forward passes for one epoch is $\frac{600}{15}$. Since the network is trained for 8 epochs, the total number of forwards passes is $\frac{600}{15} \cdot 8 = 320$
(b) 200
Since the batch size is 15, the number of forward passes for one epoch is $\frac{600}{15}$. Since the network is trained for 5 epochs, the total number of forwards passes is $\frac{600}{15} \cdot 5 = 200$

31. We define a two-layer neural network for a regression task as follows:

Let the input be:

$$x \in \mathbb{R}^d$$

The hidden layer applies a linear transformation followed by a ReLU activation:

$$h = \sigma(W_1 x + b_1), \quad \sigma(z) = \max(0, z), \quad h \in \mathbb{R}^m$$

Where:

- $W_1 \in \mathbb{R}^{m \times d}$ is the weight matrix for the hidden layer.
- $b_1 \in \mathbb{R}^m$ is the bias vector for the hidden layer.
- $\sigma(z)$ is the ReLU activation function, applied element-wise.
- $h \in \mathbb{R}^m$ is the hidden layer output.

The output layer applies a linear transformation without any activation:

$$\hat{y} = W_2 h + b_2, \quad \hat{y} \in \mathbb{R}$$

Where:

- $W_2 \in \mathbb{R}^{1 \times m}$ is the weight matrix for the output layer.
- $b_2 \in \mathbb{R}$ is the bias term for the output layer.
- $\hat{y} \in \mathbb{R}$ is the model prediction.

We use the mean squared error (MSE) as the loss function:

$$L = \frac{1}{2}(\hat{y} - y)^2$$

Where:

- $y \in \mathbb{R}$ is the true target value.
- $\hat{y}$ is the predicted output.

a. [3 points]
Find the gradient of L with respect to W_2.

$\frac{\partial L}{\partial W_2}$: _____

b. [3 points]
Find the gradient of L with respect to b_1.
Hint: Don't forget to take the gradient of the activation function!

$\frac{\partial L}{\partial b_1}$: _____

Explanation:

Part a

$$L = \frac{1}{2}(\hat{y} - y)^2$$

$$\frac{\partial L}{\partial \hat{y}} = (\hat{y} - y)$$

$$\hat{y} = W_2 h + b_2$$

$$\frac{\partial \hat{y}}{\partial W_2} = h^\top$$

$$\frac{\partial L}{\partial W_2} = (\hat{y} - y)h^\top$$

Final Answer:

$$\frac{\partial L}{\partial W_2} = (\hat{y} - y)h^\top$$

Part b

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h} \frac{\partial h}{\partial b_1}$$

$$\frac{\partial \hat{y}}{\partial h} = W_2^\top$$

$$h = \sigma(W_1 x + b_1) \quad \text{(ReLU activation)}$$

$$\frac{\partial h_i}{\partial b_1} = \begin{cases} 1 & \text{if } h_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial L}{\partial b_1} = (\hat{y} - y)W_2^\top \odot \mathbb{1}_{h>0},$$

where $\mathbb{1}_{\text{event}}$ is element wise indicator function, and $\odot$ means element wise multiplication. **Final Answer:**

$$\frac{\partial L}{\partial b_1} = \begin{cases} (\hat{y} - y)W_2^\top \odot 1 & \text{if } h > 0 \\ 0 & \text{otherwise} \end{cases}$$

32. 1 points

Suppose a dataset has n samples and d features. What is the maximum number of non-empty terminal nodes a decision tree built on this dataset can have? Assume you cannot split on the same feature more than once on any given path.

Answer: _____

Explanation: $\min\{n, 2^d\}$. In the worst case, we split on every feature on every path, which will result in 2^d terminal nodes. However there are only n data samples, so the number of non-empty terminal nodes is upperbounded by n.

33. 3 points

Prove K-means converges to a local minimum. An english proof (no explicit math) suffices.

Answer:

Explanation: The loss function L for k-means is the sum of squared distances between all points and their nearest cluster center. Note that this value for loss is non-negative.
With the assignment step of each iteration, the loss function cannot increase because every point is explicitly moved to the nearest centroid, which reduces or maintains the current total distance.

With the centroid update step of each iteration, the loss function cannot increase because the centroids are recalculated as the mean of the points in each cluster. The mean minimizes the squared distance between the points in that cluster and the centroid. The update step either reduces the loss or leaves it unchanged.

So at every iteration, either the loss is decreasing or staying the same. If it stays the same, then the cluster assignments haven't changed and the algorithm has converged. If it decreases, then there are a finite number of possible assignments to try (k^n). The algorithm will never revisit a cluster assignment because that means the loss function increases. So, in the worst case, the "last" possible assignment k-means finds, is the local minimum that it converges towards.

See lecture 16 slide 19.
https://courses.cs.washington.edu/courses/cse446/25wi/schedule/lecture-16/lecture_16.pdf

34. 1 points

Consider $M \in \mathbb{R}^{d \times d}$. Let λ be an eigenvalue of M. Suppose the eigenspace corresponding to λ equals $\mathbb{R}^d$. What is M in terms of λ?

$M =$ _____

Explanation: $M = \lambda I$. The eigenspace of λ equaling $\mathbb{R}^d$ means for any $v \in \mathbb{R}^d$, $Mv = \lambda v$. $M = \lambda I$ immediately follows.

END OF EXAM