CSE 446/546 Autumn 2024 Final Exam

December 11, 2024

Name ________________________________ **UW NetID** _____________________

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 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 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.

1. 2 points | One Answer

Imagine you are building a machine learning model to predict the stopping distance of cars based on their speed. You obtain a large dataset where each data point is a pair of observed speeds and stopping distances, and you decide to use a simple linear regression model to predict stopping distances from speed. However, in reality, the stopping distance increases quadratically with speed. As a result, your model consistently underestimates the stopping distance at higher speeds.

Compared to using a model that can model a quadratic relationship between stopping distance and speed, would your model have high or low bias?

○ a High bias

○ b Low bias

Correct answers: (a)

2. 2 points | One Answer

Follow the same car scenario as the above question. Compared to using a model that can model a quadratic relationship between stopping distance and speed, would your model have high or low variance?

○ a High variance

○ b Low variance

Correct answers: (b)

3. 2 points | One Answer

Follow the same car scenario as the above question. In reality, stopping distance is also affected by weather conditions, which your model does not capture. Which of these components of overall model error captures the error from not including weather conditions as a feature?

○ a Bias

○ b Variance

○ c Irreducible error

Correct answers: (c)

4. 4 points | Select All That Apply

Which of the following will generally help to reduce model variance?

ⓐ Increasing the size of the training data.

ⓑ Increasing the size of the validation data.

ⓒ Increasing the number of model parameters.

ⓓ Increasing the amount of regularization.

Correct answers: (a), (d)

Explanation:

a) The model has access to more information and thus is less likely to overfit to noise.

b) Increasing the size of the validation data does not help prevent the model from picking up noise in the training set.

c) This helps reduce bias not variance

d) Regularization helps prevent the model from overfitting to the training data.

5. 2 points | One Answer

For machine learning models and datasets in general, as the number of training data points grows, the prediction error of the model on unseen data (data not found in the training set) approaches 0.

ⓐ True

ⓑ False

Correct answers: (b)

Explanation: Even with infinite data, there may be noise in the data or inherent unpredictability in the relationship between input and output, which limits how low the prediction error can go.

6. 4 points | Select All That Apply

Which of the following statements about (binary) logistic regression is true? Recall that the sigmoid function is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$ for $x \in \mathbb{R}$.

○ a For any finite input $x \in \mathbb{R}$, $\sigma(x)$ is strictly greater than 0 and strictly less than 1. Thus, a binary logistic regression model with finite input and weights can never output a probability of exactly 0 or 1, and can never achieve a training loss of exactly 0.

○ b The first derivative of σ is monotonically increasing.

○ c There exists a constant value $c \in \mathbb{R}$ such that σ is convex when restricted to $x < c$ and concave when restricted to $x \geq c$.

○ d For binary logistic regression, if the probability of the positive class is $\sigma(x)$, then the probability of the negative class is $\sigma(-x)$.

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

Explanation:

a) True. $\sigma(x)$ has horizontal asymptotes at 0 and 1 and therefore is strictly bounded between those values. Because the output probability is the output of σ, this implies that the output probability is also strictly contained in $(0, 1)$. As it cannot output positive or negative labels with probability 1, it is therefore unable to reduce the training loss to exactly 0, though it can get arbitrarily close.

b) False. The gradient is highest around $x = 0$ and lowest at its asymptotes.

c) True. True for $c = 0$ and is apparent from visual inspection.

d) True. $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} = 1 - \sigma(-x)$.

7. 4 points | Select All That Apply

Consider performing Lasso regression by finding parameters $w \in \mathbb{R}^d$ that minimize $f(w) = \sum_{i=1}^{n}(y^{(i)} - x^{(i)\top}w)^2 + \lambda\|w\|_1$. Which of the following statements are true?

○ a Increasing λ will generally reduce the L_1 norm of the parameters w.

○ b Consider two models $w_1, w_2 \in \mathbb{R}^d$. Assume w_1 is more sparse, i.e., w_1 has strictly more zero coefficients than w_2. Then $\|w_1\|_1 < \|w_2\|_1$.

○ c Increasing λ generally increases model bias.

○ d Increasing λ generally increases model variance.

Correct answers: (a), (c)

Explanation:

a) True. Higher λ shrinks coefficients, encouraging sparsity.

b) False. Sparsity doesn't guarantee a smaller norm; non-zero coefficients' magnitudes matter.

c) True. Larger λ simplifies the model, leading to underfitting and higher bias.

d) False. Larger λ reduces flexibility, lowering variance.

8. 4 points Select All That Apply

Which of the following statements about ridge regression are true?

ⓐ When there are correlated features, ridge regression typically sets the weights of all but one of the correlated features to 0.

ⓑ Compared to unregularized linear regression, the additional computational cost of ridge regression scales with respect to the number of data points in the dataset.

ⓒ Ridge regression reduces variance at the expense of increasing bias.

ⓓ Using ridge and lasso regularization together (e.g., minimizing a training objective of the form $f(w) = \sum_{i=1}^{n}(y^{(i)} - x^{(i)\top}w)^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2)$ makes the training loss no longer convex.

Correct answers: (c)

Explanation:

a) False. This statement is more akin to Lasso regression. Ridge regression is more likely to somewhat equally decrease the weights of correlated features to each be smaller (as opposed to only keeping one large). See lecture 5 slide 37-38.

b) False. Ridge regression additional computational cost consists of calculating the L2-norm of all weights. This scales with respect to the number of features, not number of data points.

c) True. Ridge regression biases the model to have smaller weights and with the hope of being less likely to overfit—adding bias to reduce variance.

d) False. The sum of convex functions is also convex.

9. 5 points | Select All That Apply

Consider minimizing a function $f(x) : \mathbb{R} \to \mathbb{R}$.

Recall the following definitions:

- $x \in \mathbb{R}$ is a global minimum for f if $f(x') \geq f(x)$ for all $x' \in \mathbb{R}$.
- $x \in \mathbb{R}$ is a local minimum for f if there exists $\epsilon > 0$ such that $f(x') \geq f(x)$ for all $x' \in \mathbb{R}$ within ϵ distance of x, that is, $|x' - x| < \epsilon$.

Which of the following statements are true?

ⓐ All linear functions $f(x) = ax + b$, for some $a, b \in \mathbb{R}$, are both convex and concave.

ⓑ If f is convex, then it can have at most one global minimum. (That is, if $u, v \in \mathbb{R}$ are both global minima for f, then that implies $u = v$.)

ⓒ If f is convex, then all local minima are global minima.

ⓓ If f is convex and bounded below (i.e., there exists $c \in \mathbb{R}$ such that $f(x) \geq c$ for all $x \in \mathbb{R}$) then it must have at least one global minimum.

ⓔ If f is concave, then it must have no global minima.

Correct answers: (a), (c)

Explanation:

a) True. Linear functions are convex. Any of the tests we discussed in class apply, e.g., their second derivative (which is 0) is always greater than or equal to 0. If f is linear, then $-f$ is also linear and therefore convex, so f is also concave.

b) False. Consider the constant function $f(x) = 0$. Every point $x \in \mathbb{R}$ is a global minimum.

c) True. See class notes from lecture 7.

d) False. For example, $f(x)$ could be monotonically decreasing and asymptotically approaching 0 as x increases, so it is bounded below by 0 but has no global minimum.

e) False. Consider the same constant function $f(x) = 0$.

10. 2 points | One Answer

Let's say we want to standardize our data (i.e., normalizing the data to have zero mean and unit variance in each dimension) for the purposes of training and evaluating a ML model. Which of the following would be most appropriate?

○ a Split the dataset into the train/val/test splits, standardize the data separately for each split using the mean and variance statistics of that split.

○ b Split the dataset into the train/val/test splits, standardize the data for the training set, and use the mean and variance statistics of the training data to standardize the validation and test sets.

○ c Split the dataset into the train/val/test splits, standardize the training and validation sets separately using the mean and variance statistics of each split, then use the mean and variance statistics of the validation split to normalize the test set.

○ d Standardize the entire dataset (i.e., all splits combined) using the combined mean and variance statistics. Then, split the standardized data into train/val/test sets.

Correct answers: (b)

Explanation: We should do (b) to avoid leaking test set information to the training process. Other options may lead to overfitting to the validation or test data when picking hyperparameters.

11. 3 points | Select All That Apply

Which of the following statements about gradient descent are true? Recall that the gradient descent algorithm updates the weight parameter w at iteration t as follows: $w_{t+1} = w_t - \eta \nabla_w \ell(w)|_{w=w_t}$ (with η being the step size). For this question, we say that gradient descent has converged by iteration T if there is some iteration $t < T$ such that $\|\nabla_w \ell(w_t)\|_2^2 \leq \epsilon$ for some fixed $\epsilon > 0$.

○ a The gradient $\nabla_w \ell(w)$ points in the direction that maximizes the training loss.

○ b Assume $\ell(w)$ is convex. Then if gradient descent converges by iteration T for some fixed $\epsilon > 0$ and some step size η, it will converge in at most T iterations if we increase the step size η.

○ c Assume $\ell(w)$ is convex. Then if gradient descent converges by iteration T for some fixed $\epsilon > 0$ and some step size η, it will also eventually converge for all smaller step sizes $0 < \eta' < \eta$, given enough iterations.

Correct answers: (a), (c)

Explanation:

a) True. $\nabla_w \ell(w)$ points in the direction that maximizes the loss. Don't confuse this with the gradient descent update which steps in the "negative-gradient" direction.

b) False. large step size may cause the model to overshoot the optimum point, thus taking longer to converge.

c) True. With smaller step size, the model is likely to gradually approach the optimal point with less overshooting even if it takes more iterations.

12. $\boxed{\text{2 points}}$

Describe one advantage of mini-batch stochastic gradient descent over full-batch gradient descent.

Answer: ____

____

Explanation: One advantage is that mini-batch SGD is faster to compute over full-batch GD, while still offering an unbiased estimate of the gradient full-batch GD would compute. Another advantage is the variance of mini-batch SGD can lead to randomness that might help avoid local minima where full-batch GD might get stuck.

13. $\boxed{\text{2 points}}$

Describe one advantage of mini-batch stochastic gradient descent $(1 < B < n)$ over stochastic gradient descent with batch size $B = 1$ (e.g., updating the parameters at each iteration based only on one randomly sampled training point).

Answer: ____

____

Explanation: Possible answer: the update steps of mini-batch SGD will have less variance and might converge in fewer update steps.

More possible answers:

Noise Reduction: Mini-batches average the gradient over multiple samples, reducing the variance and leading to more stable updates.

Faster Convergence: By reducing noise, the algorithm can converge faster to a minimum.

Computational Efficiency: Mini-batches enable efficient use of parallelization on hardware like GPUs.

Better Generalization: Smoother updates can help the model generalize better.

Reduced Frequency of Parameter Updates: Fewer updates per epoch, which can improve training dynamics and efficiency.

parallelizability

14. 2 points

In a machine learning course, the distribution of final exam scores is approximately normal. However, an administrative error provided some students with prior access to practice materials closely resembling the exam, resulting in significant score increases for these students. Considering only the scores and without labeled information about who had access to the materials, what type of model would be most appropriate to estimate the likelihood that a given student had access to the practice materials?

Answer: ______________________

Explanation: This is an unsupervised learning problem because there are no labels indicating which students had access to the materials. The overall score distribution is a mixture of two Gaussian distributions:

1. Students without access: Their scores follow the original normal distribution.
2. Students with access: Their scores are higher on average, forming a second Gaussian with a higher mean.

A Gaussian Mixture Model (GMM) is the most suitable choice, as it models this bimodal distribution by combining multiple Gaussians. k-means clustering could also be used but is less effective, as it assumes spherical clusters and does not explicitly account for Gaussian distributions.

15. 4 points Select All That Apply

Assume we are given a fixed dataset $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ drawn i.i.d. (independently and identically distributed) from an underlying distribution $P(x)$. We use the bootstrap to draw bootstrap samples $\tilde{D} = \{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots\}$ from a bootstrap distribution $Q(x)$. Which of the following statements are true?

ⓐ The bootstrap samples in $\tilde{D}$ are drawn by sampling <u>with</u> replacement from D.

ⓑ The bootstrap samples in $\tilde{D}$ are drawn by sampling <u>without</u> replacement from D.

ⓒ The distribution of bootstrap samples in $\tilde{D}$ is always identical to the underlying data distribution P.

ⓓ The bootstrap samples in $\tilde{D}$ are independently and identically distributed.

Correct answers: (a), (d)

Explanation:

a) True. The bootstrap distribution is created by sampling with replacement from the fixed dataset.

b) False. Inverse of option (a)

c) False. Bootstrap samples are not guaranteed to be identical to population distribution.

d) True. By construction of the bootstrap method.

16. $\boxed{\text{2 points}}$

You are given a dataset with four data points $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)} \in \mathbb{R}$. The coordinates of these data points are:

$$\begin{aligned} x^{(1)} &= 0 \\ x^{(2)} &= 1 \\ x^{(3)} &= 5 \\ x^{(4)} &= 9. \end{aligned}$$

You run k-means on this dataset with $k = 3$ centroids, initialized at the first 3 data points: 0, 1, and 5. After k-means converges, what will be the new coordinates of these centroids? Give your answer as a sequence of 3 numbers in ascending order (e.g., "0, 1, 5").

Answer: ______________________

Explanation: $0, 1, 7$. In the first iteration, $x^{(1)}$ will be assigned to the first centroid, $x^{(2)}$ to the second centroid, and $x^{(3)}$ and $x^{(4)}$ to the third centroid. Thus the centroids will be updated to $0, 1, 7$ respectively. The centroid assignments will not change in subsequent assignments, so k-means will converge after one iteration.

Note that this clustering is not optimal (in the sense of L_2 distance from centroids); this is an example of how k-means can fail to find the globally optimal clustering.

17. 4 points | Select All That Apply

Which of the following statements are true about k-means?

○ a The output of k-means can change depending on the initial centroid positions.

○ b Assuming that the number of data points is divisible by k, k-means with k clusters always outputs clusters of equal sizes.

○ c If run for long enough, k-means will always find the globally optimal solution (as measured by the average L_2 distance between each point and its assigned cluster centroid).

○ d K-means will not converge unless all clusters in the underlying data distribution have equal, spherical variance.

Correct answers: (a)

Explanation:

a) True.

b) False. k-means is not guaranteed to produce clusters of equal sizes, it depends on where the distance between the points

c) False. k-means will converge when the cluster arrangement no longer changes. This may only be a local optimum, running longer would not help.

d) False. k-means will converge when the cluster arrangement no longer changes. This may only be a local optimum, running longer would not help.

18. 2 points

Should we initialize all the weights of a neural network to be the same small constant value (e.g., 0.001)? Why or why not?

Answer: ______________________

Explanation: No. It is important to break symmetry so that all neurons do not get the same gradient updates.

19. 4 points | Select All That Apply

In a neural network, the number of layers is an important hyperparameter. Which of these statements are true about adding layers to a neural network (keeping all other aspects of the model and training process the same)?

○ a Hyperparameters are independent, i.e., adding more layers will not affect the optimal choice of step size for gradient descent or the amount of regularization needed.

○ b We cannot use cross-validation to select hyperparameters that directly affect model architecture, such as the number of layers.

○ c Adding more layers generally decreases the training loss.

○ d Adding more layers generally increases the ability of the model to overfit the data.

Correct answers: (c), (d)

Explanation:

a) False. Adding layers can affect optimal learning rates and regularization needs.

b) False. Cross-validation can be used to select architecture-related hyperparameters like the number of layers.

c) True. More layers improve representational capacity, reducing training loss.

d) True. Deeper networks can overfit without proper regularization.

20. 4 points | Select All That Apply

Which of the following are advantages of Gaussian Mixture Models (GMMs) over K-means for a clustering application?

○ a GMMs are better suited if clusters have varying sizes and/or shapes.

○ b GMMs are better equipped to model overlapping clusters.

○ c GMMs are better suited to reason probabilistically about the data and the clusters.

○ d On a given dataset, a single iteration of the EM algorithm for fitting a GMM requires less computation than a single iteration of Lloyd's Algorithm for fitting K-means.

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

Explanation:

a) True. GMMs can model clusters with different sizes and shapes because they use a combination of Gaussian distributions, each with its own mean and covariance matrix.

b) True. GMMs can handle overlapping clusters by assigning probabilities to each data point for belonging to each cluster.

c) True. GMMs provide a probabilistic framework, giving the likelihood of each data point belonging to each cluster, which is useful for probabilistic reasoning.

d) False. The Expectation-Maximization (EM) algorithm used for fitting GMMs is generally more computationally intensive per iteration compared to Lloyd's Algorithm for K-means, due to the additional steps of calculating probabilities and updating the covariance matrices.

21. 2 points One Answer Kernel methods calculate the inner products of features in a transformed feature space, without explicitly computing the transformed features.

(a) True

(b) False

Correct answers: (a)

Explanation: A function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *kernel* for a map ϕ if $K(x, x') = \phi(x) \cdot \phi(x') = \langle \phi(x), \phi(x') \rangle$ for all x, x'. $\phi(x)$ doesn't need to be explicitly computed.

22. 2 points One Answer

Consider a fully connected neural network (MLP) with an input layer, a hidden layer, and an output layer. The input layer has n units, the hidden layer has h units, and the output layer has m units. Assume there are no bias units/terms. Which of the following statements about the number of trainable parameters is true?

(a) The total number of trainable parameters is $n \cdot h \cdot m$.

(b) The total number of trainable parameters is $n \cdot h + h \cdot m$.

(c) The total number of trainable parameters is $(n + 1) \cdot h + (h + 1) \cdot m$.

(d) The total number of trainable parameters is $n + h + m$.

Correct answers: (b)

Explanation: Connections between the input and the hidden layer: $n \cdot h$; connections between the hidden and the output layer: $h \cdot m$.

23. 6 points Select All That Apply

Consider a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{S}$ is an $r \times r$ diagonal matrix and $r = \text{rank}(\mathbf{A}) \leq \min(m, n)$. Which of the following statements are correct?

(a) The columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{A}^\top\mathbf{A}$.

(b) The columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^\top$.

(c) The columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{A}^\top\mathbf{A}$.

(d) The columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^\top$.

(e) The singular values in $\mathbf{S}$ are the square roots of the nonzero eigenvalues of $\mathbf{A}\mathbf{A}^\top$.

(f) The singular values in $\mathbf{S}$ are the square roots of the nonzero eigenvalues of $\mathbf{A}^\top\mathbf{A}$.

Correct answers: (b), (c), (e), (f)

Explanation: $\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{S}^2\mathbf{U}^\top$, implying that the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^\top$ with corresponding eigenvalues along the diagonal of $\mathbf{S}^2$. Similarly, $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{S}^2\mathbf{V}^\top$, implying that the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{A}^\top\mathbf{A}$ with corresponding eigenvalues along the diagonal of $\mathbf{S}^2$.

24. 3 points

Consider a dataset $X \in \mathbb{R}^{n \times p}$ with n observations and p features, and with corresponding covariance matrix Σ. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ be the eigenvalues of Σ in descending order. Express the total variance explained by the first k principal components (obtained by performing Principal Component Analysis (PCA) on X) as a fraction of the total variance in the original data.

Answer: Fraction of total variance explained = ______________________

Explanation: $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$. The fraction of total variance explained by the first k principal components in PCA can be expressed as the ratio of the sum of the first k eigenvalues to the sum of all eigenvalues of the covariance matrix Σ.

25. $\boxed{\text{3 points}}$

Consider a dataset $X \in \mathbb{R}^{n\times 2}$ with n observations and 2 features. Suppose Σ is the covariance matrix of the dataset:

$$\Sigma = \begin{bmatrix} 3 & \sqrt{3} \\ \sqrt{3} & 5 \end{bmatrix}$$

This covariance matrix has the following unit-norm eigenvectors u and v:

$$u = \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{bmatrix}, v = \begin{bmatrix} \frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{bmatrix}$$

Write the <u>second</u> principle component as a unit-length vector in vector form (i.e. $\begin{bmatrix} a \\ b \end{bmatrix}$).

Second principal component: ______________________________

Explanation: $\begin{bmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{bmatrix}$.

A vector x and value λ are defined to be an eigenvector-eigenvalue pair of A if $Ax = \lambda x$.

$\Sigma u = \begin{bmatrix} -\sqrt{3} \\ 1 \end{bmatrix} = 2 \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{bmatrix} = 2u$, so $\lambda_u = 2$.

$\Sigma v = \begin{bmatrix} 3 \\ \frac{6\sqrt{3}}{2} \end{bmatrix} = 6 \begin{bmatrix} \frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{bmatrix} = 6v$, so $\lambda_v = 6$.

Eigenvector-eigenvalue pairs of a covariance matrix represent pairs of principal components and the variance explained by that principal component. u's eigenvalue is less than v's, so it is the second principal component. u is already unit-length, so it is the final answer.

26. 4 points | Select All That Apply

You are applying PCA to a training dataset of $n = 1024$ grayscale images that are each 16×16 pixels (256 pixels per image). Consider reshaping each image into a vector $x_i \in \mathbb{R}^{256}$ and then composing a data matrix $X \in \mathbb{R}^{1024 \times 256}$, where the i^{th} row is $x_i^\top$. Let $\hat{x}_{i,k} \in \mathbb{R}^{256}$ be the PCA reconstruction of image x_i using the top k principal component directions in the data. Let $R(k)$ be the average reconstruction error on the training data using k principal components, $R(k) = \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x}_{i,k}\|_2^2$. Which of the following statements are true?

(a) $R(k)$ is monotonically decreasing as k increases, up to $k = 1024$. That is, if $0 < k_1 < k_2 \leq 1024$, then $R(k_1) > R(k_2)$. Note the strict equality.

(b) If $k < \text{rank}(X)$, then $R(k) > 0$.

(c) If $k \geq \text{rank}(X)$, then $R(k) = 0$.

(d) For $k \geq 1$, let $\delta(k) = R(k-1) - R(k)$ be the decrease in reconstruction error by going from $k-1$ to k principal components. (When $k = 0$, define the reconstruction of x_i to simply be the mean image $\bar{x}$.) Then, $\delta(k)$ is monotonically non-increasing as k increases.

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

Explanation:

a) False. The number of principal components cannot exceed the rank of X, which is $\min(n, 256)$. Since X is 1024×256, its rank is at most 256. Thus, $R(k)$ is only guaranteed to monotonically decrease for $k \leq 256$, not $k \leq 1024$.

b) True. The reconstruction error is non-zero when the number of principal components k is less than the rank of X, as there are remaining variations in X not captured by the top k components.

c) True. When k is greater than or equal to the rank of X, the PCA reconstruction captures all the variation in X, resulting in zero reconstruction error.

d) True. Each additional principal component explains the maximum remaining variance, so the decrease in reconstruction error ($R(k)$) diminishes as k increases, making $R(k)$ monotonically non-increasing.

27. 4 points | Select All That Apply

Which of the following is/are true about the k-Nearest Neighbors (k-NN) algorithm?

(a) Testing time (i.e., the amount of time it takes to produce an output for a new test point) increases with the number of training samples.

(b) The number of hyperparameters increases with the number of training samples.

(c) k-NN can learn non-linear decision boundaries.

(d) k-NN clusters unlabeled samples in a k-dimensional space based on their similarity.

Correct answers: (a), (c)

Explanation:

a) True. k-NN looks at the entire training dataset to classify points at testing time

b) False. The algorithm has a fixed number of hyperparameters (arguably, just k).

c) True. it can learn non-linear decision boundaries.

d) False. It is not a clustering algorithm.

28. 4 points | Select All That Apply

Which of the following statements about random forests and decision trees are true?

○ a Random forests are generally easier for humans to interpret than individual decision trees.

○ b Random forests reduce variance (compared to individual decision trees) by aggregating predictions over multiple decision trees.

○ c When constructing the individual trees in the random forest, we want their predictions to be as correlated with each other as possible.

○ d Random forests can give a notion of confidence estimates by examining the distribution of outputs that each individual tree in the random forest produces.

Correct answers: (b), (d)

Explanation:

a) False. Procedure is similar except random forest utilizes multiple decision trees

b) True. Aggregating predictions from multiple trees reduces sensitivity compared to a single tree.

c) False. Having as correlated trees as possible degenerates to a single tree, losing the benefits of a more complex forest.

d) True. Spread of decisions across different trees gives a confidence estimate.

29. 4 points | Select All That Apply

Which of the following is a correct statement about (mini-batch) Stochastic Gradient Descent (SGD)?

○ a The variance of the gradient estimates in SGD decreases as the batch size increases.

○ b Running SGD with batch size 1 for n iterations is generally slower than running full-batch gradient descent with batch size n for 1 iteration, because the gradients for each training point in SGD have to be computed sequentially, whereas the gradients in full-batch gradient descent can be computed in parallel.

○ c SGD is faster than full-batch gradient descent because it only updates a subset of model parameters with each step.

○ d SGD provides an unbiased estimate of the true (full-batch) gradient of the training loss.

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

Explanation:

a) True. In SGD, the gradient is estimated using a subset of the data. A sampled batch might not represent the entire dataset well. As the batch size increases, it becomes more representative of the entire dataset, reducing the variance in the gradient estimates.

b) True. In batch gradient descent, all gradients for the entire dataset are computed in one forward-backward pass, which can leverage parallel processing (e.g., on GPUs).

c) False. SGD does not update a subset of model parameters. It updates all parameters based on the gradient computed from a subset of the data. The faster convergence of SGD compared to full-batch gradient descent is due to the more frequent updates.

d) True. The gradient computed on a mini-batch is an unbiased estimate of the full gradient because the mini-batch is a random sample of the dataset. This randomness ensures that, on average, the mini-batch gradient equals the true gradient over the entire dataset.

30. $\boxed{\text{2 points}}$ $\boxed{\text{One Answer}}$

The probability density function for a gamma distribution with parameters $\theta > 0, k > 0$ is

$$f(x; \theta, k) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}},$$

where

$$\Gamma(x) = (x-1)!$$

Say we have a dataset D of n data points, $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, where each $x \in \mathbb{R}$. Assume that k is given to us and fixed. We would like to use D to find the maximum likelihood estimator for θ. What is the maximum likelihood estimator for θ in terms of k, n, and $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$?
Hint: The argmax of the logarithm of a function is the same as the argmax of the function.

- ○ a $\frac{1}{kn} \sum_{i=1}^{n} x^{(i)}$
- ○ b $\frac{n}{(k-1)!} \sum_{i=1}^{n} x^{(i)} e^{-\frac{x^{(i)}}{k}}$
- ○ c $\ln(\frac{1}{n} \sum_{i=1}^{n} x^{(i)}) - n(k-1)!$
- ○ d $\frac{\ln(k)-(k-1)!}{\frac{1}{k}}$

Correct answers: (a)

Explanation: To find the maximum likelihood estimator (MLE) for θ, we start with the likelihood function for a dataset $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$:

$$L(\theta) = \prod_{i=1}^{n} f(x^{(i)}; \theta, k) = \prod_{i=1}^{n} \frac{1}{\Gamma(k)\theta^k} (x^{(i)})^{k-1} e^{-\frac{x^{(i)}}{\theta}}.$$

The log-likelihood function is:

$$\ell(\theta) = \sum_{i=1}^{n} \ln f(x^{(i)}; \theta, k) = -n \ln \Gamma(k) - kn \ln \theta + (k-1) \sum_{i=1}^{n} \ln x^{(i)} - \frac{1}{\theta} \sum_{i=1}^{n} x^{(i)}.$$

To maximize $\ell(\theta)$, we differentiate with respect to θ and set the derivative to zero:

$$\frac{\partial \ell}{\partial \theta} = -\frac{kn}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x^{(i)} = 0.$$

Multiply through by θ^2 to simplify:

$$-kn\theta + \sum_{i=1}^{n} x^{(i)} = 0 \quad \Rightarrow \quad \theta = \frac{1}{kn} \sum_{i=1}^{n} x^{(i)}.$$

Thus, the maximum likelihood estimator for θ is:

$$\hat{\theta} = \frac{1}{kn} \sum_{i=1}^{n} x^{(i)}.$$

The correct answer is **(a)**.

31. 2 points | One Answer

Many ML algorithms, like the k-nearest neighbors (k-NN) algorithm, relies on distances between points. In high-dimensional spaces, distances can behave counterintuitively. This question illustrates one such example.

Consider two d-dimensional hypercubes S and T centered around the origin. S has side length 2, while T is contained within S and has side length 1:

$$S = \{x \in \mathbb{R}^d \ : \ \|x\|_\infty \leq 1\}$$
$$T = \{x \in \mathbb{R}^d \ : \ \|x\|_\infty \leq \frac{1}{2}\}.$$

Alternatively, we can write $S = [-1, 1]^d$, and $T = [-\frac{1}{2}, \frac{1}{2}]^d$. Let P be the uniform distribution of points in S. What is the probability of drawing a point $x \sim P$ such that $x \in T$, that is, x is contained within T? Give your answer in terms of d.

Answer: ____

Explanation: The volume of S is 2^d, while the volume of T is 1. Since x is uniformly distributed in S, the probability of $x \in T$ is the relative ratio of their volumes, which is $\frac{1}{2^d}$.

32. 3 points | Select All That Apply

Consider the following dataset of four points in $\mathbb{R}^2$:

$$\begin{aligned} x^{(1)} &= (0,0) & y^{(1)} &= -1 \\ x^{(2)} &= (0,1) & y^{(2)} &= +1 \\ x^{(3)} &= (1,0) & y^{(3)} &= +1 \\ x^{(4)} &= (1,1) & y^{(4)} &= -1. \end{aligned}$$

This is also known as a XOR problem because the labels y are the result of applying the XOR operation to the two components of x. For a given data point $x \in \mathbb{R}^2$, denote its first dimension as x_1 and its second dimension as x_2. For example, $x_1^{(2)} = 0$ and $x_2^{(2)} = 1$. Which of the following statements are true?

○ a There exists a linear model $w \in \mathbb{R}^3$, which predicts $+1$ if

$$w^\top \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} \geq 0$$

and -1 otherwise, that achieves 100% accuracy on this dataset.

○ b There exists a linear model $w \in \mathbb{R}^6$, which predicts $+1$ if

$$w^\top \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \\ 1 \end{bmatrix} \geq 0$$

and -1 otherwise, that achieves 100% accuracy on this dataset.

○ c Define a polynomial feature expansion $\phi(x)$ as any function $\phi(x) : \mathbb{R}^2 \to \mathbb{R}^d$ that can be written as

$$\begin{bmatrix} x_1^{a_1} x_2^{b_1} \\ x_1^{a_2} x_2^{b_2} \\ \vdots \\ x_1^{a_d} x_2^{b_d} \end{bmatrix}$$

for some integer $d > 0$ and integer vectors $a, b \in \mathbb{Z}^d$. Then there does not exist any polynomial feature expansion $\phi(x)$ such that a linear model w which predicts $+1$ if $w^\top \phi(x) \geq 0$, and -1 otherwise, achieves 100% accuracy on this dataset.

Correct answers: (b)

Explanation: a) There is no way to separate with linear features.

For option b) and example weight vector is $\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -4 \\ -1 \end{bmatrix}$

For option c) a counter example is feature expansion $\begin{bmatrix} x_1^0 x_2^0 \\ x_1^0 x_2^1 \\ x_1^1 x_2^0 \\ x_1^1 x_2^1 \end{bmatrix}$ with weight vector $\begin{bmatrix} -1 \\ 1 \\ 1 \\ -2 \end{bmatrix}$

33. 2 points | One Answer

Consider the following transfer learning setting. We have a large neural network $\phi : \mathbb{R}^d \to \mathbb{R}^p$ pretrained on ImageNet, and we would like to use this to learn a classifier for our own binary classification task for medical images. We decide to freeze the neural network ϕ and train a logistic regression classifier on top.

Formally, we are given n data points from our own medical imaging task $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$, where $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{-1, +1\}$. We train a classifier $\hat{w} \in \mathbb{R}^p$:

$$\hat{w} = \text{argmin}_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \log \left(1 + \exp\left(-y^{(i)} w^\top \phi(x^{(i)})\right)\right).$$

Which of the following statements is true?

(a) Learning $\hat{w}$ in this way is a convex optimization problem regardless of how complex ϕ is.

(b) Learning $\hat{w}$ in this way is a convex optimization problem if and only if ϕ is a convex function in each dimension. (Let $\phi = [\phi_1; \phi_2; \ldots; \phi_p]$; then we say ϕ is convex in each dimension if each of $\phi_1, \phi_2, \ldots, \phi_p$ is a convex function).

(c) Learning $\hat{w}$ in this way is a convex optimization problem if and only if ϕ is a linear function.

(d) Learning $\hat{w}$ in this way is a convex optimization problem if and only if ϕ is the identity function and $p = d$.

Correct answers: (a)

Explanation: Since we freeze ϕ and do not update it, this is equivalent to logistic regression with a fixed basis expansion. Thus, it is a convex optimization problem regardless of how complex ϕ is.

34. $\boxed{\text{2 points}}$

[This is an extra credit question that takes more time relative to the number of points awarded. We suggest you do not attempt it until you have finished the other questions in the exam.]

Recall from lecture that influence functions are used to approximate the effect of leaving out one training point, without actually retraining the model. Assume that we have a twice-differentiable, strongly convex loss function $\ell(x, y; w)$, and as usual, we train a model $\hat{w}$ to minimize the average training loss:

$$\hat{w} = \text{argmin}_w \ \frac{1}{n} \sum_{i=1}^{n} \ell_i(w),$$

where $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$ is our training set, and for notational convenience we define $\ell_i(w) = \ell(x^{(i)}, y^{(i)}; w)$. Let Δ_{-i} be the change in the parameters w after we remove training point $(x^{(i)}, y^{(i)})$ and retrain the model. The influence function approximation tells us that

$$\Delta_{-i} = \frac{1}{n} H(\hat{w})^{-1} \ \nabla_w \ell_i(w)\big|_{w=\hat{w}}$$

where the Hessian matrix $H(\hat{w})$ is defined as

$$H(\hat{w}) = \frac{1}{n} \sum_{i=1}^{n} \nabla_w^2 \ell_i(w)\big|_{w=\hat{w}}$$

Consider the following linear regression model $f_w(x) = w^\top x$, where $x, w \in \mathbb{R}^d$. We train with unregularized least squares regression to obtain $\hat{w}$. What is Δ_{-i} for this model, in terms of $\hat{w}$ and the training data points? Note: The symbols ℓ and H should not appear in your answer. Replace them by working out the appropriate loss.

Answer: ____

Explanation: For least squares regression, we have that $\ell_i(w) = \frac{1}{2}(y^{(i)} - w^\top x^{(i)})^2$. (The $\frac{1}{2}$ is for convenience;

we can leave it out without changing the final answer.) Thus, $\nabla_w \ell_i(w) = -(y^{(i)} - w^\top x^{(i)})x^{(i)}$, and $H(w) = \frac{1}{n}\sum_{i=1}^n x^{(i)}x^{(i)\top}$. Putting this together,

$$\begin{aligned}\Delta_{-i} &= -\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n x^{(i)}x^{(i)\top}\right]^{-1} \left(y^{(i)} - \hat{w}^\top x^{(i)}\right) x^{(i)} \\ &= -\left[\sum_{i=1}^n x^{(i)}x^{(i)\top}\right]^{-1} \left(y^{(i)} - \hat{w}^\top x^{(i)}\right) x^{(i)}.\end{aligned}$$

We accept both the simplified version (canceling $\frac{1}{n}$) and the unsimplified version.