CSE 446 Spring 2025 Final Exam

June 11, 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.

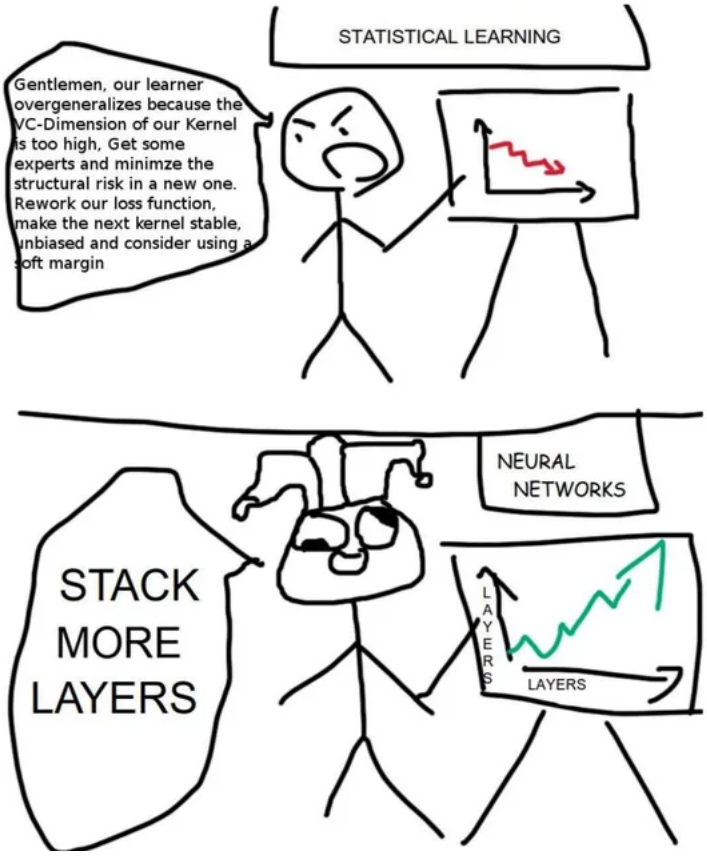
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

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

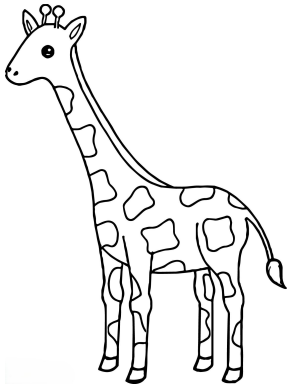
  Not selected answer: (a) Selected answer: (a)

- 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. Natasha's all-time favourite deep learning meme:



Giraffe for good luck :)



1. 1 points | One Answer

Which of the following is the cause/reason for irreducible error?

(a) Stochastic label noise

(b) Very few data points

(c) Nonlinear relationships in the data

(d) Insufficient model complexity

Correct answers: (a)

Explanation: A is correct. Stochastic label noise is what drives irreducible error. See lecture 4 slides. In essence, irreducible error comes from randomness that cannot be modeled since there is no deeper pattern to it. B and D are wrong because fewer data points and insufficient model complexity are responsible for *reducible* error. C is wrong because nonlinear relationships in the data don't have anything to do with irreducible error.

2. 1 points | One Answer

Saket unfortunately did not learn from the midterm and still has not attended lecture. He is now given the task of training 3 neural networks with increasing complexity on a regression task:

- Model A: 1 hidden layer with 10 neurons.
- Model B: 2 hidden layers with 50 neurons each.
- Model C: 10 hidden layers with 100 neurons each.

After training and evaluating these models on an appropriately split dataset with train and test splits, you find the following MSEs:

- Model A: train MSE = 2.5, test MSE = 2.6
- Model B: train MSE = 0.1, test MSE = 0.2
- Model C: train MSE = 0.01, test MSE = 1.3

Saket only knows about bias and variance, So based on the model architectures and train/test MSE losses, chose the best relative bias/variance estimates for each of the models.

Model	**Bias**				**Variance**			
A	Low	◯	High	◯	Low	◯	High	◯
B	Low	◯	High	◯	Low	◯	High	◯
C	Low	◯	High	◯	Low	◯	High	◯

Explanation: Correct answer: A => high bias, low variance; B => low bias, low variance; C => low bias, high variance

Due to the simpler architecture and high MSEs, A likely underfits. B achieves low but similar train/test MSEs so probably has a good balance. C has a low train MSE but a high test MSE so is probably overfitting, which matches the likely overcomplex architecture.

3. [2 points]

Explain one upside and one downside of using a high K in K-fold cross validation.

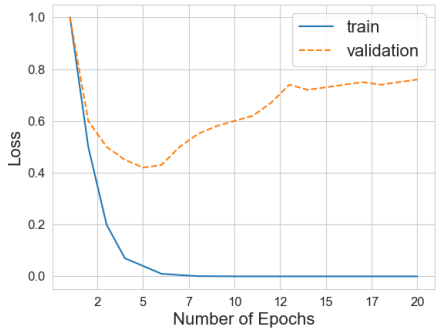
Upside:

Downside:

Explanation: Possible answer: **Upside:** You get a more accurate estimate of your test error, possibly making hyperparameter selection more accurate. **Downside:** A higher K means more folds and therefore much more compute/time needed to find the right hyperparameters. A higher K also means each validation set has fewer data points. This will result in higher variability in the results across different folds.

4. 1 points | Select All That Apply

You are training a model and get the following plot for your training and validation loss.



Which of the following statements are true?

(a) The model has high bias and low variance.

(b) The large gap between training and validation loss indicates underfitting.

(c) Training for more epochs will eventually decrease validation loss.

(d) The model might be too complex for the dataset.

(e) The model is likely memorizing the training data.

Correct answers: (d), (e)

Explanation: This is a classic example of overfitting, which is caused when we have too complex of a model and it ends up memorizing the training set. Overfitting means the model has low bias and high variance. Thus, the only correct options are D and E.

5. 1 points | Select All That Apply

Which of the following models that we studied in class use maximum likelihood estimation?

(a) Linear regression with Gaussian noise model

(b) Principal Components Analysis

(c) Gaussian Mixture Models

(d) Neural Network trained to do classification with softmax cross entropy loss

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

Explanation: a) True: is true because you maximize the likelihood of the data under a linear model which

assumes a gaussian distribution on errors b) False: PCA does not use MLE because it does not define a probalistic distribution for the data, it just uses linear algebra to find vectors that explain a lot of variance in the data c) True: Gaussian Mixture Models define a probability distribution that is a mixture of Gaussians and then find the parameters by maximizing likelihood under that model d) True: NNs with softmax define a probability distribution over the classification labels and try to maximize it with cross entroyp

6. 1 points | One Answer

Yann, a strict frequentist statistician, observes 5 flips of a possibly uneven coin. Here are the outcomes:

1. Heads
2. Tails
3. Heads
4. Heads
5. Tails

Based on these observations, Yann uses using maximum likelihood estimation to determine the most likely outcome of the next coin toss. What does he predict will happen?

(a) Heads

(b) Tails

(c) Both are equally likely

(d) It hits Marco in the head

Correct answers: (a)

Explanation: There were 3 Heads and 2 Tails. Based on these observations, the estimated probability of Heads is $3/5 = 0.6$, which is greater than the estimated probability of Tails ($2/5 = 0.4$). Therefore, Heads is the most likely outcome.

7. Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable everywhere, such that $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ for all $x, y \in \mathbb{R}^d$. Suppose there exists a unique $x_* \in \mathbb{R}^d$ such that $\nabla_x f(x_*) = 0$.

(a) [1 points] [One Answer]

x_* is a:

- (a) Minimizer of f
- (b) Maximizer of f
- (c) Saddle point of f
- (d) Not enough information to determine any of the above

(b) [1 points]

Suppose we are unable to solve for x_* in closed-form. Briefly outline a procedure for finding x_*.

Correct answers: (a)

Explanation: Part (a): f is convex, so x_* must be a minimizer of f.
Part (b): Gradient descent

8. 1 points | One Answer

Which of the following is true, given the optimal learning rate?
Clarification made during exam: "All options refer to convex loss functions that have a minimum bound / have a minimum value"

(a) For convex loss functions, stochastic gradient descent is guaranteed to eventually converge to the global optimum while gradient descent is not.

(b) For convex loss functions, both stochastic gradient descent and gradient descent will eventually converge to the global optimum.

(c) Stochastic gradient descent is always guaranteed to converge to the global optimum of a loss function.

(d) For convex loss functions, gradient descent with the optimal learning rate is guaranteed to eventually converge to the global optimum point while stochastic gradient descent is not.

Correct answers: (d)

Explanation: Due to the noisy updates of SGD, it is not guaranteed to converge at the minimum but for instance, cycle close to it whereas batch gradient descent alleviates this and is guaranteed to reach the minimum given appropriate step size.

9. 3 points

Imagine you are trying to find an optimal weight w for a simple model. You have a small dataset consisting of two data points, each influencing the overall loss:

- Data point 1: $(x_1, y_1) = (5, 4)$
- Data point 2: $(x_2, y_2) = (1, 3)$

You are using a squared error loss function for each individual data point, defined as

$$L_i(w) = (y_i - w \cdot x_i)^2$$

Your current weight parameter is $\mathbf{w_0 = 1}$. You will perform one iteration of Stochastic Gradient Descent (SGD) using a learning rate $\boldsymbol{\alpha = 0.1}$. You will process one "randomly" chosen data point to compute the gradient and update the weight. *For this exercise, you may choose which data point to process.*

a) "Randomly" select one data point for this iteration.

Circle one: **Data point 1** **Data point 2**

b) For your selected data point, calculate the loss at your current weight $\boldsymbol{w_0}$.

Loss at w_0: _____ (should be a real number)

c) For your selected data point, calculate the gradient of the loss with respect to w at $\boldsymbol{w_0}$.

Gradient at w_0: _____ (should be a real number)

d) Update the weight using the SGD rule to find $\boldsymbol{w_1}$.

Weight w_1: _____ (should be a real number)

Explanation: We are given:

- Data point 1: $(x_1, y_1) = (5, 4)$
- Data point 2: $(x_2, y_2) = (1, 3)$

Loss function: $L_i(w) = (y_i - w \cdot x_i)^2$
Initial weight: $w_0 = 1$
Learning rate: $\alpha = 0.1$

General Formulas: The loss function for a selected data point (x_i, y_i) is $L_i(w) = (y_i - w \cdot x_i)^2$. The gradient of the loss with respect to w is: $\nabla L_i(w) = \frac{d}{dw}(y_i - w \cdot x_i)^2 = 2(y_i - w \cdot x_i) \cdot (-x_i) = -2x_i(y_i - w \cdot x_i)$. The SGD update rule is: $w_{\text{new}} = w_{\text{old}} - \alpha \cdot \nabla L_i(w_{\text{old}})$.

—

Case 1: Student Selects Data Point 1 ($x_1 = 5, y_1 = 4$)

- **a) Selected Data Point:** Data point 1
- **b) Loss at $w_0 = 1$:** $L_1(1) = (y_1 - w_0 \cdot x_1)^2 = (4 - 1 \cdot 5)^2 = (4 - 5)^2 = (-1)^2 = \mathbf{1}$

- **c) Gradient at $w_0 = 1$:** $\nabla L_1(1) = -2x_1(y_1 - w_0 \cdot x_1) = -2(5)(4 - 1 \cdot 5) = -10(4-5) = -10(-1) = \mathbf{10}$
- **d) Weight w_1 after SGD update:** $w_1 = w_0 - \alpha \cdot \nabla L_1(w_0) = 1 - 0.1 \cdot (10) = 1 - 1 = \mathbf{0}$

—

Case 2: Student Selects Data Point 2 ($x_2 = 1, y_2 = 3$)

- **a) Selected Data Point:** Data point 2
- **b) Loss at $w_0 = 1$:** $L_2(1) = (y_2 - w_0 \cdot x_2)^2 = (3 - 1 \cdot 1)^2 = (3-1)^2 = (2)^2 = \mathbf{4}$
- **c) Gradient at $w_0 = 1$:** $\nabla L_2(1) = -2x_2(y_2 - w_0 \cdot x_2) = -2(1)(3 - 1 \cdot 1) = -2(3-1) = -2(2) = \mathbf{-4}$
- **d) Weight w_1 after SGD update:** $w_1 = w_0 - \alpha \cdot \nabla L_2(w_0) = 1 - 0.1 \cdot (-4) = 1 + 0.4 = \mathbf{1.4}$

10. 1 points | One Answer

Which of the following activation functions saturates, i.e. stops giving meaningful gradients for large positive inputs?

(a) ReLu

(b) Sigmoid

(c) Softmax

Correct answers: (b)

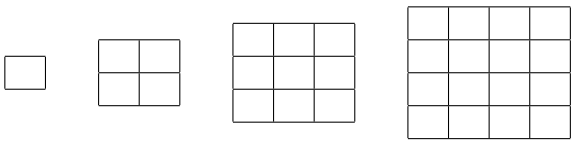
Explanation: The gradient for Sigmoid and Tanh approaches 0 as the magnitude of the input increases. Softmax is not an activation function.

11. 2 points

Consider the following matrix M and kernel filter F.

$$M = \begin{array}{|c|c|c|}\hline 9 & 7 & 8 \\ \hline 4 & 1 & 3 \\ \hline 2 & 6 & 4 \\ \hline \end{array} \qquad F = \begin{array}{|c|c|}\hline 1 & 0 \\ \hline 1 & 1 \\ \hline \end{array}$$

Apply the filter F to matrix M with padding = 0 and stride = 1, then perform a Max Pooling operation on the result with a 2x2 filter and stride 1. Write the resulting matrix below in the grid of the correct size. Only write answers in **one** matrix, otherwise the problem will be graded as incorrect.



Explanation: 14

After applying F to M, we get: $\begin{array}{|c|c|}\hline 14 & 11 \\ \hline 12 & 11 \\ \hline \end{array}$. Applying a Max Pool operation with a 2x2 filter just means taking the max of this matrix, since it's a 2x2, so we get the final answer of $\boxed{14}$

12. 2 points

What are the spatial dimensions of the output image if a 2×2 filter is convolved with a 3×3 image for paddings of 0, 1, and 2, and strides of 1 and 2? Fill in the dimensions below:

Padding	0	1	2
Stride 1	_____	_____	_____
Stride 2	_____	_____	_____

Explanation: **Stride 1**: Padding 0 (2x2) Padding 1 (4x4) Padding 2 (6x6)
Stride 2: Padding 0 (1x1) Padding 1 (2x2) Padding 2 (3x3)

13. 1 points | One Answer

Compared to Lasso, Ridge regression tends to be more stable in terms of which features are important to the model's predictions in high-dimensional cases because it doesn't drive correlated weights to 0.
Clarification made during exam: "Should read as "More stable in terms of which features are important to the model's predictions *as you increase the amount of regularization* in high-dimensional..."""

(a) True
(b) False

Correct answers: (a)

Explanation: This is true because Ridge "smoothly shrinks" all weights making it more stable to small changes in the data or noise.

14. 1 points | One Answer

For $X \in \mathbb{R}^{n \times d}$ and $y \in \{-1, 1\}^n$, if our data is linearly separable then the minimization problem

$$\arg\min_w \sum_{i=1}^{n} \log(1 + \exp(-y_i w^\top x_i))$$

does not have a unique solution.

(a) True

(b) False

Correct answers: (a)

Explanation: If our data is linearly separable we can push the magnitude of w to ∞ to push the objective to 0 but never actually reach 0, so there is no solution.

15. 1 points | One Answer

Suppose we have a matrix $M \in \mathbb{R}^{n \times m}$ and perform SVD on it to get 3 matrices U, S, V. If we take the first r singular vectors of U, V corresponding to the first r singular values in S (ordered highest to lowest), where $r = \min(n, m)$, then we can perfectly reconstruct M without any loss whatsoever.

(a) True

(b) False

Correct answers: (a)

Explanation: $r = \min(n, m) \geq \text{rank}(M)$. If we perform a rank r reconstruction on a matrix who's maximum rank is r, we will get a lossless reconstruction.

16. 1 points | Select All That Apply

Which of the following are equivalent to the principal components of a data matrix X? Assume X has already been de-meaned.

(a) Vectors that create a subspace which maximize the variance of X if X is projected onto that subspace.

(b) Vectors that create a subspace which minimize the variance of X if X is projected onto that subspace.

(c) The eigenvectors of $X^\top X$

(d) The right singular vectors of X

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

Explanation: A is correct because this is the definition of principal components. B is the opposite so it is false. The right singular vectors of X are also the eigenvectors of $X^\top X$, and both are equal to the principal components of X. Therefore C and D are correct.

17. 1 points | One Answer

In PCA, minimizing the reconstruction error is equivalent to minimizing the projected variance.

(a) True

(b) False

Correct answers: (b)

Explanation: Minimizing the reconstruction error is equivalent to maximizing the variance.

18. 1 points | Select All That Apply

You apply PCA on a dataset of 100 features and get 100 principal components. Which of the following are good reasons to chose only the top q principal components instead of all 100? Assume $q < 100$.

(a) To remove noise by discarding the highest variance components.

(b) To reduce redundant features in the dataset.

(c) To reduce the computational cost of working with the data.

(d) To make a beautiful plot.

Correct answers: (c), (d)

Explanation: If we chose the top q components, those would be the ones with highest variance, so A is incorrect. B is incorrect as we have 100 features and 100 PCs in this case, so we are not reducing redundant features here; all of them are meaningful features. C is correct, because by only picking the top q PCs, we are reducing the dimensionality of the dataset and thus reducing computational cost. D is correct, as it has been mentioned numerous times in lecture before: it helps us create a beautiful plot.

19. 1 points | One Answer

Generally, tree-based methods have:
Clarification made during exam: "It should be 'decision trees' instead of 'tree-based methods."

(a) Low bias, low variance

(b) Low bias, high variance

(c) High bias, low variance

(d) High bias, high variance

Correct answers: (b)

Explanation: Tree-based methods usually have low bias and high variance.

20. 1 points | Select All That Apply

Forrest just trained a decision tree for predicting whether a person will like a song based on features like its genre, key, length, etc. He notices an extremely low training error, but an abnormally large test error. He also notices that a regularized multi-class logistic regression model performs much better than his tree. What could be the cause of his problem?

(a) Learning rate too high

(b) Decision tree is too deep

(c) There is too much training data

(d) Decision tree is overfitting

Correct answers: (b), (d)

Explanation: He is observing overfitting which could be caused by a complex/deep tree.

21. 2.5 points | One Answer

Match each of the modeling problems (which include a description of the data and desired criteria for the model) with the best machine learning method for the job. **Use each model type once.**

Modeling problems:

- Problem A: You are training a model for a medical setting where you have a small number of categorical input features, and the ability to be able to interpret your model's decisions is important.
- Problem B: You have a small dataset (small n), continuous Y labels, but many features. You want an interpretable model that you can regularize to give you information about which features are more important.
- Problem C: You have a large dataset (large n) of images.
- Problem D: You have a lot of data (large n) in a small dimensional feature space (small d), and you assume that your labels y change smoothly with changes in the feature space.
- Problem E: Your data has a relatively small number of categorical features and you want to win a Kaggle competition.

Machine learning methods:

- k-Nearest Neighbours (kNN)
- Decision Tree (DT)
- Random Forest (RF)
- Convolutional Neural Network (CNN)
- Linear Regression (LR)

Problem	**Machine Learning method**				
A	◯ k-NN	◯ DT	◯ RF	◯ CNN	◯ LR
B	◯ k-NN	◯ DT	◯ RF	◯ CNN	◯ LR
C	◯ k-NN	◯ DT	◯ RF	◯ CNN	◯ LR
D	◯ k-NN	◯ DT	◯ RF	◯ CNN	◯ LR
E	◯ k-NN	◯ DT	◯ RF	◯ CNN	◯ LR

Explanation: Problem A is decision tree because they're good for categorical features and interpretable. Problem B is linear regression because it works for small datasets and continuous labels. Problem C is convolutional neural networks. Problem D is kNN. Problem E is Random Forests

22. [1 points] [One Answer]

You are training a decision tree to perform classification of into labels $Y \in 0, 1$. Your tree sorts the labels into the following leaves. What is the entropy $H(X)$ for each of the following sets X:

a) $X = 1, 1, 1, 1$: _____

b) $X = 1, 1, 0, 0$: _____

c) $X = 0, 0, 0, 0$: _____

Explanation: $H(X) = -\sum_i p(i) * \log p(i)$ Answers: a) 0.0, $(H = -1 * \log(1) = 0)$ b) 1.0, $(H = [0.5 \log(0.5) + 0.5 \log(0.5)] = [20.5(1)] = 1)$ c) 0.0, $(H = -1 * \log(1) = 0)$

23. [1 points] [Select All That Apply]

You are applying the kernel method to n data points, where each data point $x_i \in \mathbb{R}^d$. Which of the following statements are true.

(a) The kernel method performs computations on a high dimensional feature space $\phi(x_i) \in \mathbb{R}^p$, where $p >> d$.

(b) A function K is a kernel for a feature map ϕ if $K(x, x') = \phi(x)^T \phi(x')$.

(c) The kernel trick relies on the fact if $p >> n$, then the data spans at most a d-dimensional subspace of $\mathbb{R}^p$.

(d) Kernel methods can be considered non-parametric because they require retaining the training data for making predictions about new points.

Correct answers: (b), (d)

Explanation: a) is not correct because it avoids actually performing computations in the p-dimensional feature space b) correct, defn of kernel c) incorrect, it should be an n-dimensional subspace d) correct.

24. Consider data matrix $X \in \mathbb{R}^{n \times d}$ and feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$, for some p. Let K be the corresponding kernel matrix.

(a) [1 points]
Let $\phi(X)$ denote X with ϕ applied to each data point. Write K in terms of $\phi(X)$.

$K =$ _____

(b) [1 points] [One Answer]
The ith entry on the diagonal of K is:

(a) $||\phi(x_i)||_1$

(b) $||\phi(x_i)||_2$

(c) $||\phi(x_i)||_2^2$

(d) None of the above

Correct answers: (c)

Explanation: part (a): $K = \phi(X)\phi(X)^\top$.
part(b): $K_{ii} = \phi(x_i)^\top \phi(x_i) = ||\phi(x_i)||_2^2$

25. [1 points] [Select All That Apply]
Natasha is trying to train a k-Nearest Neighbors model, and she encounters the "curse of dimensionality". This refers to the fact that as the dimensionality of her feature space d increases...

(a) Distances between points become less meaningful, since all points are far apart.

(b) She has too much data making computation too expensive to perform on a single machine.

(c) The amount of data required to cover the space increases exponentially.

(d) Thinking in more than three dimensions is hard so we should use PCA to make a 2D plot.

Correct answers: (a), (c)

Explanation: a-c are all correct statements of the same idea. d is a joke

26. 1 points | Select All That Apply

You want to cluster this data into 2 clusters. Which of the these algorithms would work well?

(a) Spectral clustering

(b) K-means

(c) GMM clustering

Correct answers: (c)

Explanation: Only GMM takes the Gaussian distributions of the two clusters into account even when they overlap

27. 1 points | One Answer

Which of the following statements is true about K-means clustering?

(a) K-means clustering works effectively in all data distributions.

(b) K-means is guaranteed to converge.

(c) K-means clustering is a supervised learning algorithm.

(d) The accuracy of K-means clustering is not affected by the initial centroid selections.

Correct answers: (b)

Explanation: A is false since K-means doesn't work well in all distributions, such as non-spherical clusters. B is true, since K-means will always converge (see lecture notes for proof). C is false, since K-means is unsupervised. D is false, since the accuracy of the classifier is influenced by the initial centroid selections.

28. 1 points | One Answer

Suppose a Gaussian Mixture Model (GMM) with k components/clusters is used to model a dataset of dimensionality d. Which value does the total number of parameters in the GMM primarily scale with respect to?

(a) $O(k \cdot d)$

(b) $O(k \cdot d^2)$

(c) $O(d)$

(d) $O(d^2)$

(e) $O(k)$

(f) $O(n)$

(g) $O(\frac{d}{n})$

Correct answers: (b)

Explanation: The parameters of a GMM are the mixture weights, the means, and the covariance matrices. There are k mixing weights, each $\in \mathbb{R}$. There are k means, each $\in \mathbb{R}^d$. There are k covariance matrices, each $\in \mathbb{R}^{d \times d}$. Since the covariance matrices have the most parameters, the k covariance matrices are the "determining factor". So the answer is $O(k \cdot d^2)$ Can we minimize the number of options to maybe 4?

29. 1 points | One Answer

Because bootstrap sampling randomly draws data points with replacement, the size of the original dataset does not affect accuracy of the estimated statistics produced by bootstrapping.

(a) True

(b) False

Correct answers: (b)

Explanation: Smaller datasets will not be as representative of the true dataset, yielding less accuracte statistics.

30. [1 points] [One Answer]

Suppose you are working with a dataset that includes demographic information (e.g., age, gender, race) to predict loan approval. You notice that your model not only performs significantly worse on some groups, but it is more likely to reject underrepresented minorities for a loan. Which of the following is the best way to address this bias? Choose the *best* answer.

(a) Remove the demographic information altogether, forcing the model to not rely on demographic information.

(b) Over-sample underrepresented groups to balance the dataset and reduce bias.

(c) Include fairness constraints such as ensuring that the type II error (probability of rejecting someone for a loan when they deserved it) is balanced across groups.

(d) Collect more historical data about loan approvals for underrepresented groups and re-train your model.

Correct answers: (c)

Explanation: a) Demographic info is often highly correlated with other features so removing them wouldn't entirely help. b) Balancing data can help but doesn't help with the underlying issue of biases. c) This is the current state-of-the-art approach d) this doesn't necessarily work because the historical data is still biased

31. [1 points] [One Answer]

I've trained a linear regression model on my dataset and learned weights w_i for each of my d features. I notice that $w_i > w_j$, so I can conclude feature i is more important than feature j.

(a) True

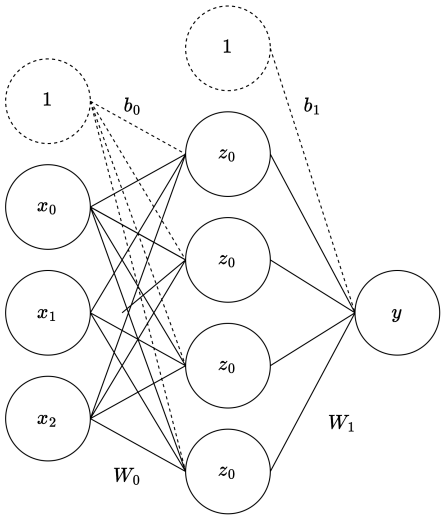
(b) False

Correct answers: (b)

Explanation: no, the features could have different scales, such as square feet vs. number of bathrooms.

32. 4 points

Consider the following network:



The forward pass for the hidden layer is $z = \sigma(W^{(0)}x + b^0)$, where σ refers to the sigmoid activation function. The output layer is $y = W^{(1)}z + b^1$. Derive the partial derivatives with respect to $W^{(1)} \in \mathbb{R}^{1\times h}$, $b^{(1)} \in \mathbb{R}$, $W^{(0)} \in \mathbb{R}^{h\times d}$, and $b^{(0)} \in \mathbb{R}^h$, where $d = 3$ and $h = 4$.

Clarification made during exam: "Typo: $b^0 = b_0 = b^{(0)}$. They all refer to the same thing."

a) $\frac{\partial y}{\partial W^{(1)}}$: _____

b) $\frac{\partial y}{\partial b^{(1)}}$: _____

c) $\frac{\partial y}{\partial W^{(0)}}$: _____

d) $\frac{\partial y}{\partial b^{(0)}}$: _____

Explanation:

a) $\frac{\partial y}{\partial W^{(1)}} = z$

b) $\frac{\partial y}{\partial b^{(1)}} = 1$

c) $\frac{\partial y}{\partial W^{(0)}} = [{W^{(1)}}^T \odot z \odot (1-z)] * x^T$

This problem is very similar to the question from section 8. First, to make the math simpler, we can compute $\frac{\partial y}{\partial W_i^{(0)}}$, where W_i^0 is the i-th of W^0. Computing the derivatives w.r.t. to $W_i^{(0)}$ necessitates chain rule; we can rewrite it as $\frac{\partial y}{\partial W_i^{(0)}} = \frac{\partial y}{\partial z_i} * \frac{\partial z_i}{\partial W_i^{(0)}} = W_i^{(1)} * \frac{\partial z_i}{\partial W_i^{(0)}}$. From here, the derivative of z_i w.r.t. $W_i^{(0)}$ can be computed using the derivative of the sigmoid function ($\sigma * (1-\sigma)$). Doing so, we get $z_i * (1-z_i) * x^T$, where the x^T comes from applying chain rule. Putting everything together, we get $\frac{\partial y}{\partial W_i^{(0)}} = W_i^{(1)} * z_i * (1-z_i) * x^T$. Note that this is a column vector, with the derivatives for a single row. To generalize this and get the derivative of y w.r.t. to the entirety of $W^{(1)}$, we repeat the same process for all rows of $W^{(1)}$, which we can denote using the elementwise operator. Thus, we get $[{W^{(1)}}^T \odot z \odot (1-z)] * x^T$. Note we need to transpose $W^{(1)}$ in order multiply it elementwise with $z \odot (1-z)$.

d) $\frac{\partial y}{\partial b^{(0)}} = {W^{(1)}}^T \odot z \odot (1-z)$

This derivation is very similar to the one above, except we don't have x^T since only the weights matrix is multiplied with the data vector. So we get: $\frac{\partial y}{\partial b_i^{(0)}} = W_i^{(1)} * \frac{\partial z_i}{\partial b_i^{(0)}} = W_i^{(1)} * z_i * (1-z_i) \rightarrow {W^{(1)}}^T \odot z \odot (1-z)$.

33. Transitioning to electric cars can help fight climate change, but electric cars cause such a strain on the electrical grid that if several people on the same block all buy an electric car within a few weeks or months of each other, it can actually cause the grid to go down!

2 points

You've been hired by the electric company to build a cool new machine learning model to help predict which houses will start charging electric cars next. You've been handed several messy files of data. The first contains high-level information about n different houses, including whether they have an electric vehicle or not, each house's location, square footage, value, household income, results of the last election in the house's zipcode, public school ratings in the zip code, etc. But, you can also get detailed electricity data for each house, including daily electricity consumption going back at least 3 years. **Describe the feature engineering or data preprocessing steps you would take to prepare to use this data to train a machine learning model:**

2 points

Now, you must use the data you prepared to train a machine learning model that can tell you which houses are likely to get an electric car in the next year. **Please describe the machine learning model you will use for this problem**. You will be graded on how well you can **justify why your model is a good choice for this problem**, by explaining how the properties of your model suit the problem.

Explanation: The criteria for grading this are do they find ways to mention real things about machine learning models they learned in class. Like "I will use a random forest because it's good for categorical data but has lower variance than a tree".

Valid explanations include: **Feature engineering steps.** Find some way to reduce the daily electricity data into something more manageable. Could use something like PCA, or manually extract features.

Can mention separating into train, validation, and test.

Could mention normalizing the features to be on the same scale.

Give mega bonus points if they mention propagating features about the neighbors' recent adoption of electric cars into the feature space for a house.

Neural network. don't need to do much feature engineering, throw the daily electricity data for each house into the features for each house, end up with a huge d, run gradient descent, see if it works

Trees. Good for categorical data like political affiliations, public school ratings. Use a random forest to reduce variance. Doesn't work with a ton of features (high d, so should only be used in conjunction with feature engineering.

Logistic regression. They might mention this is a classification problem so they want to use this. Maybe they mention it's more interpretable, so the power grid company can inspect the results.

kNN. If they reduce the feature space small enough this could be a good pick. They could say it's best to figure out who will adopt a car based on whether other similar people adopted a car. Maybe in terms of literal distance. But this is not a good answer if they use all the daily electricity data, because then the feature space would be too large.

CNN. Not a great answer because no images.

34. Bonus Question | 4 points

This is a bonus question. You can get extra points for completing it, but you will not lose points if you do not get the right answer.

Let $f, g : \mathbb{R}^d \to \mathbb{R}$ be convex. Use the epigraph definition of convexity to prove that $h(x) = \max\{f(x), g(x)\}$ is convex.

Hint 1: You may use that for any convex sets $A, B \subset \mathbb{R}^d$, $A \cap B$ is convex.

Hint 2: You may use that for any $a, b, c \in \mathbb{R}$, $c \geq a \wedge c \geq b$ if and only if $c \geq \max\{a, b\}$.

Explanation: *Proof.* Denote $\text{epi}(f) := \{(x, t) \in \mathbb{R}^{d+1} : t \geq f(x)\}$, with $\text{epi}(g)$, $\text{epi}(h)$ defined similarly. By the epigraph definition of convexity, we know that the sets $\text{epi}(f)$ and $\text{epi}(g)$ are convex. Note that for any $(x, t) \in \mathbb{R}^{d+1}$ we have by hint 2 that $t \geq f(x)$ and $t \geq g(x)$ if and only if $t \geq \max\{f(x), g(x)\} = h(x)$. Thus we have that $(x, t) \in \text{epi}(f) \cap \text{epi}(g)$ if and only if $(x, t) \in \text{epi}(h)$. It follows that $\text{epi}(h) = \text{epi}(f) \cap \text{epi}(g)$. Since $\text{epi}(f)$ and $\text{epi}(g)$ are convex, by hint 1, $\text{epi}(h)$ must be convex. So by the epigraph definition of convexity, h is convex. □

END OF EXAM