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.

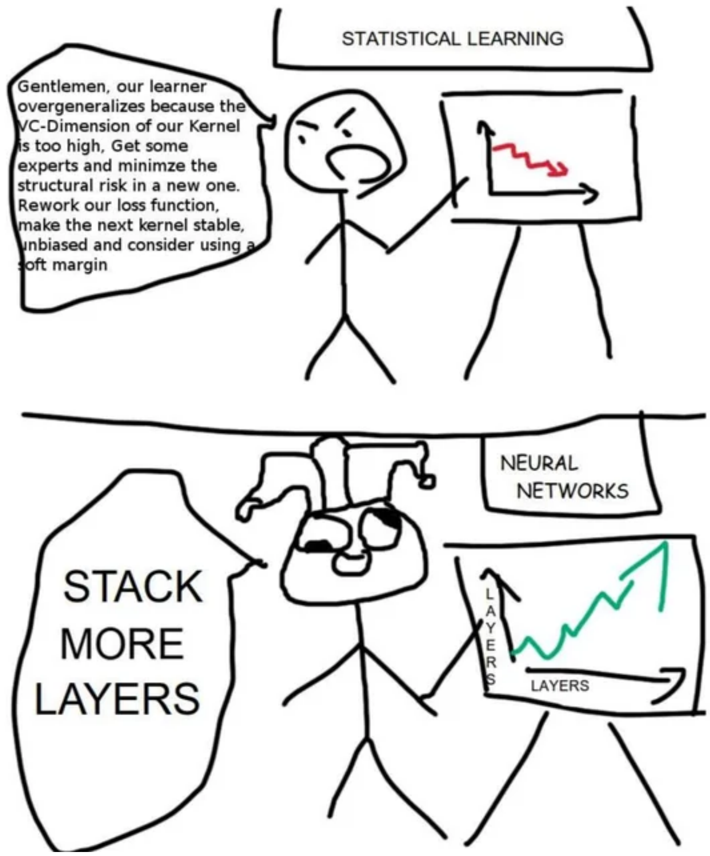
Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

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

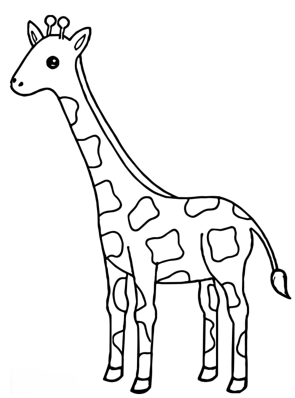
  Not selected answer: (a) Selected answer: (a)

- 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

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 ◯

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

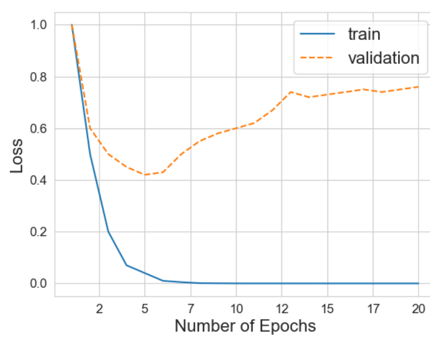
Explain one upside and one downside of using a high K in K-fold cross validation.

Upside:

Downside:

4. $\boxed{\text{1 points}}$ $\boxed{\text{Select All That Apply}}$

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.

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

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

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_*.

   ___

   ___

   ___

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.

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

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 $\mathbf{w_0} = \mathbf{1}$. You will perform one iteration of Stochastic Gradient Descent (SGD) using a learning rate $\boldsymbol{\alpha} = \mathbf{0.1}$. 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)

10. $\boxed{\text{1 points}}$ $\boxed{\text{One Answer}}$

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

(a) ReLu

(b) Sigmoid

(c) Softmax

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

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.

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

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	________	________	________

13. $\boxed{\text{1 points}}$ $\boxed{\text{One Answer}}$

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

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

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

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

17. 1 points | One Answer

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

(a) True
(b) False

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.

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

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

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

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) $X = 0, 0, 0, 0$: ________________________________

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.

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

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.

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

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.

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

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

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.

31. 1 points One Answer

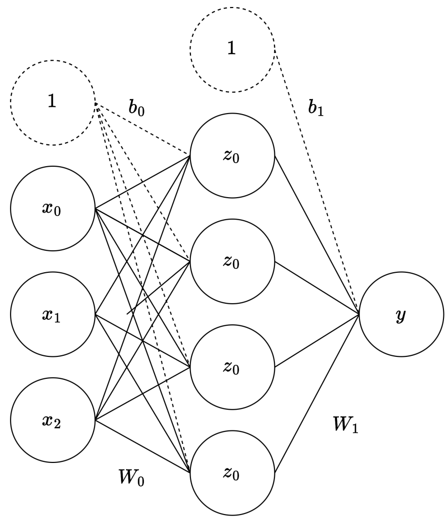
I've trained a linear regression model on my dataset and learned weights w_i for each of my d features. I notice that $w_i > w_j$, so I can conclude feature i is more important than feature j.

(a) True

(b) False

32. $\boxed{\text{4 points}}$

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

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!

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

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:**

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

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.

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 $c \geq \max\{a, b\}$.

END OF EXAM