CSE 446 Spring 2024 Final Exam

June 5, 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 (multiple choice, short answer).

- Each question is worth 1 point unless noted otherwise.
- For each multiple-choice question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- Multiple choice questions marked with $\boxed{\text{One Answer}}$ should only be marked with one answer. All other multiple choice questions are marked $\boxed{\text{Select All}}$, in which case they are "select all that apply" and 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 and note sheet by handing them to a TA.

1. One Answer Let $L_i(w)$ be the loss of parameter w corresponding to a sample point X_i with label y_i. The update rule for stochastic gradient descent with step size η is

   (a) $w_{new} \leftarrow w - \eta \nabla_{X_i} L_i(w)$

   (b) $w_{new} \leftarrow w - \eta \sum_{i=1}^{n} \nabla_{X_i} L_i(w)$

   (c) $w_{new} \leftarrow w - \eta \nabla_{w} L_i(w)$

   (d) $w_{new} \leftarrow w - \eta \sum_{i=1}^{n} \nabla_{w} L_i(w)$

   Correct answers: (c)

2. One Answer Suppose data $x_1, ..., x_n$ is drawn from an exponential distribution $\exp(\lambda)$ with PDF $p(x|\lambda) = \lambda \exp(-\lambda x)$. Find the maximum likelihood for λ ?

   (a) $\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$

   (b) $\lambda = \sum_{i=1}^{n} x_i$

   (c) $\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$

   (d) $\lambda = \log(\sum_{i=1}^{n} x_i)$

   Correct answers: (a)

   Explanation: $log(p(data|\lambda)) = nlog(\lambda) - \lambda \sum_{i=1}^{n} x_i$
   $d(log(p(data|\lambda))) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$

3. One Answer Aman and Ed built a model on their data with two regularization hyperparameters λ and γ. They have 4 good candidate values for λ and 3 possible values for γ, and they are wondering which λ, γ pair will be the best choice. If they were to perform five-fold cross-validation, how many validation errors would they need to calculate?

   (a) 12

   (b) 17

   (c) 24

   (d) 60

   Correct answers: (d)

4. One Answer Which of the following is most indicative of a model overfitting?

(a) Low bias, low variance.

(b) Low bias, high variance.

(c) High bias, low variance.

Correct answers: (b)

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

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

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

(c) It reduces computational time.

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

Correct answers: (b)

6. One Answer Two realtors are creating machine learning models to predict house costs based on house traits (i.e. house size, neighborhood, school district, etc.) trained on the same set of houses, using the same model hyperparameters. Realtor A includes 30 different housing traits in their model. Realtor B includes 5 traits in their model. Which of the following outcomes is most likely?

(a) Realtor B's model has higher variance and lower bias than Realtor A's model.

(b) Realtor A's model has higher variance than Realtor B's model and without additional information, we cannot know which model has a higher bias.

(c) Realtor A's model has higher variance and lower bias than Realtor B's model.

(d) Realtor A's model has higher variance and higher bias than Realtor B's model.

Correct answers: (b)

7. Select All Suppose we have N data points $x_1, x_2, \ldots, x_N$ that $x_i \in \mathbb{R}^d$. Define $X \in \mathbb{R}^{n\times d}$ such that $X_{i,j} = (x_i)_j$, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, and $\mathbf{1}_N = (1, 1, \ldots, 1)^\top \in \mathbb{R}^N$. Which of the following are true about principal components analysis (PCA)?

(a) The principal components are eigenvectors of the centered data matrix $X - \mathbf{1}_N\bar{x}^\top$.

(b) The principal components are right singular vectors of the centered data matrix.

(c) The principal components are eigenvectors of the sample covariance matrix $\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\top$.

(d) Applying a rigid rotation matrix Q (i.e., $QQ^\top = Q^\top Q = I$) to X will not change the principal components' directions.

Correct answers: (b), (c)

Explanation: (a) They are eigenvectors of $(X - \mathbf{1}_N\bar{x}^\top)^\top(X - \mathbf{1}_N\bar{x}^\top)$. (d) The directions change by Q.

8. Select All In the context of singular value decomposition (SVD) $A = U\Sigma V^\top$, which of the following statements are correct?

(a) The columns of U are called left singular vectors and form an orthonormal basis for the range of A, while the columns of V are called right singular vectors and form an orthonormal basis for the range of $A^\top$.

(b) For any A that is real and symmetric, we have $U = V$.

(c) For a square matrix A, the singular values of A are the absolute values of the eigenvalues of A.

(d) Singular values are always non-negative real numbers.

Correct answers: (a), (d)

Explanation: (c) Consider $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, then eigenvalues are $\lambda_1 = \lambda_2 = 1$. However, $AA^\top = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, then singular values are $\sigma_{1,2} = \sqrt{(3 \pm \sqrt{5})/2}$.

9. Select All Which of the following statements about matrix completion are correct?

(a) It may not perform well when the real-world data is not inherently low-rank or when the pattern of missing observations is not random.

(b) The purpose of matrix completion is to estimate missing entries in a partially observed matrix.

(c) Matrix completion is only applicable for square matrices.

Correct answers: (a), (b)

Explanation: (c) No such restriction.

10. $\boxed{\text{One Answer}}$ Consider the feature map $\phi : \mathbb{R}^2 \to \mathbb{R}^4$ defined as $\phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_2 x_1 \end{bmatrix}$.

What is the corresponding kernel function K for ϕ?

(a) $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ and $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^2$.

(b) $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $K(x, x') = x^4 + x'^4 + 2x^2 x'^2$.

(c) $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ and $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$.

(d) $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $K(x, x') = x^2 + x'^2$.

Correct answers: (a)

11. $\boxed{\text{One Answer}}$ In the context of kernel methods, what does the "kernel trick" refer to?

(a) Adding an extra kernel layer to the end of a neural network.

(b) A technique for explicitly computing the coordinates in a high-dimensional space.

(c) A method for computing the inner products in a high-dimensional feature space without explicitly mapping data to that space.

(d) A technique for speeding up the convergence of gradient descent.

Correct answers: (c)

12. One Answer When using a kernel method to solve a regression problem with training set $\{(\mathbf{x_i}, y_i)\}_{i=1}^n$ and $\mathbf{x_i} \in \mathbb{R}^d$, we first prove that there exists an $\alpha \in \mathbb{R}^n$ such that the weight vector $\hat{\mathbf{w}} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x_i})$, where $\phi : \mathbb{R}^d \to \mathbb{R}^p$ is a feature map transforming $\mathbf{x_i}$ into a very high dimensional space $\mathbb{R}^p$ with $p \gg d$. Then, solving the problem is equivalent to finding $\hat{\alpha} = \text{argmin}_\alpha \sum_{i=1}^n (y_i - \sum_{j=1}^n \alpha_j K(\mathbf{x_i}, \mathbf{x_j}))^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(\mathbf{x_i}, \mathbf{x_j})$. After we computed the value of $\hat{\alpha}$, given an input $\mathbf{x'} \in \mathbb{R}^d$ in the test set, how can we make the prediction?

(a) Because $\hat{\mathbf{w}} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x_i})$, we can compute the value of $\hat{\mathbf{w}}$, and then applying $\hat{y} = \hat{\mathbf{w}}^\top \mathbf{x'}$.

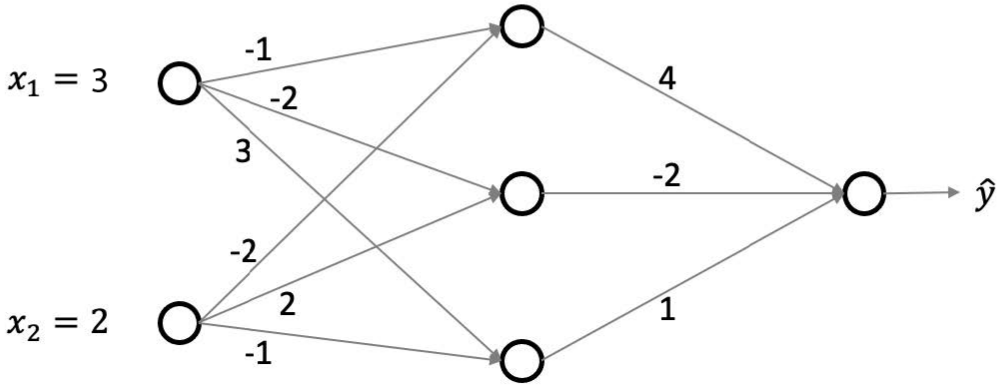
(b) Because $\hat{y} = \hat{\mathbf{w}}^\top \mathbf{x'} = \sum_{i=1}^n \alpha_i \mathbf{x_i}^\top \mathbf{x'}$, we can compute the values of $\mathbf{x_i}^\top \mathbf{x'}$ and then get the value of $\hat{y}$.

(c) Because $\hat{y} = \hat{\mathbf{w}}^\top \phi(\mathbf{x'}) = \sum_{i=1}^n \alpha_i \phi(\mathbf{x_i})^\top \phi(\mathbf{x'}) = \sum_{i=1}^n \alpha_i K(\mathbf{x_i}, \mathbf{x'})$, we can compute the values of $K(\mathbf{x_i}, \mathbf{x'})$ and then get the value of $\hat{y}$.

(d) Because $\hat{\mathbf{w}} = \sum_{i=1}^n \alpha_i \phi(\mathbf{x_i})$, we can compute the value of $\hat{\mathbf{w}}$, and then applying $\hat{y} = \hat{\mathbf{w}}^\top \phi(\mathbf{x'})$.

Correct answers: (c)

13. Consider the following neural network with weights shown in the image below. Every hidden neuron uses the ReLU activation function, and there is no activation function on the output neuron. Assume there are no bias terms. What is the output of this network with the input $x = (3, 2)$?



Answer: ____________________________

Explanation: Answer is 7

14. How many parameters does the neural network shown in the previous problem have?
Answer: ______________________________

Explanation: Answer is 9

15. One Answer Which of the following defines the correct ordering of steps needed to perform backpropagation in PyTorch?

(a) (1) compute loss, (2) compute gradients, (3) take step, (4) zero the gradient buffers

(b) (1) compute loss, (2) zero the gradient buffers, (3) compute gradients, (4) take step

(c) (1) compute loss, (2) take step, (3) zero the gradient buffers, (4) compute gradients

(d) (1) zero the gradient buffers, (2) compute gradients, (3) take step, (4) compute loss

Correct answers: (b)

16. Consider a convolutional neural network (CNN) layer with the following parameters:
- Input image size: $3 \times 32 \times 32$ (channels, height, width)
- Number of filters: 16
- filter size: 3×3
- Stride: 1
- Padding: 1

What will be the shape of the output after applying this convolutional layer (in the order of channels, height, width)?

Answer: ______________________________

Explanation: $16 \times 32 \times 32$. $\frac{32-3+2\times 1}{1} + 1 = 32$

17. One Answer Which of the following best describes the purpose of pooling layers in a convolutional neural network (CNN)?

(a) To increase the resolution of the feature maps.

(b) To reduce the spatial dimensions of the feature maps, thereby reducing the computational load and the number of parameters.

(c) To convert the feature maps into a fully connected layer.

(d) To normalize the feature maps by scaling them to a fixed range.

Correct answers: (b)

18. Given the following setup in a simple recurrent neural network (RNN):
Input at time t: x_t
The RNN has one hidden layer with the following parameters:
    - Input to hidden state weights: W_{xh}
    - Hidden state to hidden state weights: W_{hh}
    - Hidden state to output weights: W_{hy}
    - Bias for the hidden state: b_h
    - Bias for the output: b_y

    The activation function for the hidden state is ReLU.

    Given the following parameter values:

$$W_{xh} = \begin{bmatrix} 0.5 & 0.1 \\ 0.3 & 0.2 \end{bmatrix}, \quad W_{hh} = \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.5 \end{bmatrix}, \quad W_{hy} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b_h = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}, \quad b_y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

    Calculate the hidden state h_1 after processing the first input $x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Assume the initial hidden state h_0 is a zero vector.

    Answer: ______________________________

    Explanation: $relu(W_{xh}x_1 + W_{hh}h_0 + b_h) = \begin{bmatrix} 0.6 \\ 0.5 \end{bmatrix}$

19. One Answer Which of the following statements about the k-means clustering algorithm is true?

    (a) It guarantees convergence to the global optimum.

    (b) It is robust against the initialization of cluster means.

    (c) It may converge to a local optimum depending on the initial placement of cluster means.

    Correct answers: (c)

20. One Answer Why might a Gaussian Mixture Model (GMM) be preferred over K-means in cases where the data contains mixed or overlapping clusters?

    (a) GMM exclusively uses hard assignments which are better for mixed clusters.

    (b) GMM utilizes soft assignments, allowing points to belong to multiple clusters with varying probabilities.

    (c) GMM always converges faster than k-means.

    Correct answers: (b)

21. One Answer In k-means clustering, the algorithm is executed several times on the same dataset, each time with a fresh random initialization of cluster centers and the same number of clusters. If these multiple runs yield widely varying cluster outcomes, what might this suggest about the algorithm's sensitivity to initial conditions?

(a) The choice of k is optimal.

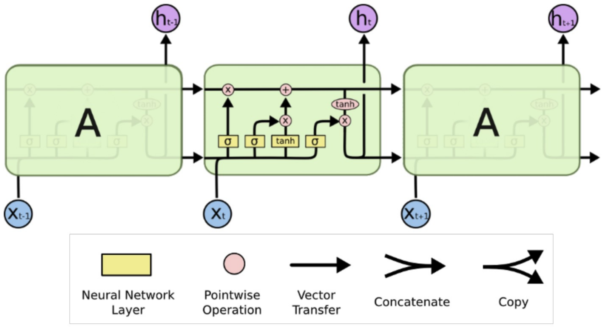
(b) The dataset is perfectly clustered.

(c) The initialization of centers might be influencing the results.

(d) The algorithm is not suitable for clustering.

Correct answers: (c)

22. One Answer Consider an LSTM (Long Short-Term Memory) network with the following characteristics: a forget gate, an input gate, a memory cell, and an output gate. Which of the following statements correctly describes the function of the forget gate in an LSTM?



(a) The forget gate decides which information should be discarded from the cell state.

(b) The forget gate calculates the output of the current position at each time step.

(c) The forget gate extracts useful information from the input to update memory.

(d) The forget gate calculates the next hidden state based on the current input.

Correct answers: (a)

Explanation: The forget gate in an LSTM outputs a value between 0 and 1 for each number in the cell state c_{t-1}, where 1 represents "completely keep this" and 0 represents "completely forget this". It is used to remove information that is no longer needed from the cell state.

23. One Answer In the context of neural machine translation, which key benefit does the attention mechanism provide over the standard RNN models?

(a) It significantly reduces the computational complexity of the model.

(b) It uses convolutional layers to handle long-term dependencies.

(c) It relies entirely on recurrent layers for processing sequences.

(d) It solves the bottleneck problem and long-term dependency issues by focusing on specific parts of the input sequence.

Correct answers: (d)

Explanation: The attention mechanism addresses the bottleneck problem and long-term dependency issues present in standard Seq2Seq models by enabling the model to focus on specific parts of the input sequence during the decoding process.

24. One Answer What is the main purpose of using positional encoding in the Transformer architecture?

(a) It introduces non-linearity in the model.

(b) It helps in maintaining long-term dependencies.

(c) It provides information about the order of the input sequence.

(d) It reduces the computational complexity.

Correct answers: (c)

Explanation: Positional encoding is used in the Transformer architecture to provide information about the position of each element in the input sequence, which is necessary because the self-attention mechanism in Transformers is order-invariant and does not inherently capture the order of the sequence.