ECE 47300: Midterm 1 (Spring 2023), Version: C

Instructions

Please fill out this cover page but do NOT start until instructed to do so.

What is your full name and Purdue ID (10-digit number)?

Name (please PRINT, no cursive): Paloma Arellano

Purdue ID (10 digits): 0031926985

Honor Pledge Signature: PA

- **Total number of points**: 100 (Roughly 50 questions or subquestions)
- **Question format notes**
  - Circular checkboxes are for multiple choice (i.e., a single correct answer)
  - Square checkboxes are for select all questions (i.e., possibly multiple correct answers).
  - Only put your answer in the boxes like [ ] .
- **No partial credit**: Every (sub-)question is all or nothing credit. Thus, you must get the answer exactly right to get credit for the question (including SELECT ALL questions). No partial credit will be given.
- **Number Format**: When giving numbers as short answers, please give in standard decimal notation with preceding "0." for decimals if needed but no trailing 0s (e.g., "0.15", "2.9", "0.001", "100" but NOT "0.15000" NOR ".15" NOR ".001" NOR "6.0").
- **Honor Pledge**: I assert that I have not received any information about this exam and will not share any exam content with anyone else, even after the exam. I understand that any violation of this will result in a failing grade for the whole class (not just this exam).

1. (2 points) Select all the reasons that dimensionality reduction may be helpful:
   - [x] Noise reduction via reconstruction
   - [x] Visualization
   - [x] Lower computational costs

2. (2 points) What was the main strength(s) of the first wave of AI called "Handcrafted Knowledge" according to DARPA's perspective (e.g. chess, TurboTax, rule-based scheduling systems)? Select all that apply.
   - [ ] Perceiving
   - [ ] Abstracting
   - [x] Reasoning
   - [ ] Learning

3. (2 points) Assuming that A, B, C, D are 3 x 3 matrices, which of the following statements are ALWAYS true? Select all that apply.
   - [ ] $A(B+C) = AB + CA$
   - [ ] $C(A+B) = CA + CB$
   - [ ] $(AB)^T = A^T B^T$
   - [x] $ABC + ABD = (AB)(D+C)$
   - [x] A is not singular if and only if $\text{rank}(A) = 3$

4. (2 points) Numpy has two types of 1D arrays: column vectors and row vectors.
   - ○ True
   - ⊗ False

5. (2 points) If $X = \begin{bmatrix} 1 & -2 \\ -3 & 1 \end{bmatrix}$ what is $\|X\|_F^2$?

   Answer: 15

   $(1)^2 + (-2)^2 + (-3)^2 + 1^2$
   $= 1 + 4 + 9 + 1 = 15$

6. (2 points) What is the ℓ_∞-norm of the vector $\mathbf{x} = [-1, -2, 3, -5, -4, 3, -9, 1]$ (i.e., what is $\|\mathbf{x}\|_\infty$)?

   Answer: 9

7. (2 points) If $X_c = USV^T$ is the SVD decomposition of the centered matrix and $\Sigma_x = \frac{1}{n} X_c^T X_c$ is the covariance matrix whose eigendecomposition is $Q\Lambda Q^T$, which of the following are solutions of the principal component vectors for the dataset X_c (i.e., W^*)?
   - [ ] $U_{1:k}$
   - [ ] $S_{1:k}$
   - [x] $V_{1:k}$
   - [ ] $\Sigma_{1:k}$
   - [x] $Q_{1:k}$
   - [ ] $\Lambda_{1:k}$

8. (2 points) Suppose $x = [10, 5]^T$ and $w = [\frac{3}{5}, \frac{4}{5}]^T$ (note $\|w\|_2 = 1$). What is the solution to $\arg\min_z \|x - zw\|_2^2$?

$(10)(\frac{3}{5}) + (5)(\frac{4}{5}) = 6 + 4 = 10$

- ○ $z^* = 5$
- ○ $z^* = 8$
- ○ $z^* = 11$
- ☒ None of the above / Many possible solutions

9. (2 points) Suppose $X_c \in \mathbb{R}^{100 \times 10}$ and X_c has a rank of 8. If we run PCA with $k = 6$ latent dimensions, what will be the PCA reconstruction error?

- ☒ Greater than 0
- ○ 0
- ○ Less than 0

10. (2 points) Cross validation provides a good estimate of a model's performance even if the real-world test distribution of unseen data is different from the training dataset.

- ☒ True
- ○ False

11. (2 points) The KNN algorithm uses less computation during training rather than test/inference time.

- ○ True
- ☒ False

12. (2 points) Let $\theta = [1, 5, 1, 0]^T$ be the parameters of a 3 dimensional linear regression model f_θ (where θ_4 is the intercept term). What is $f_\theta(x + [1, 0, 3, 7]^T) - f_\theta(x)$?

$f_\theta(x+\vec{w}) = 1(x_1+1) + 5(x_2+0) + 1(x_3+3) + 0(x_4+7)$
$= x_1 + 1 + 5x_2 + x_3 + 3 = 4$
$f_\theta(x) = x_1 + 5x_2 + x_3$

- ○ 1
- ○ 2
- ☒ 3
- ☒ 4
- ○ Impossible to know given the available information

13. (2 points) Deep learning methods will always perform better than linear methods because they can model non-linear functions.

- ○ True
- ☒ False

14. (2 points) Gradient descent will not converge to a local optimum for any fixed step size.

- ☒ True
- ○ False

15. (2 points) Stochastic gradient descent computes the full gradient as in gradient descent but then explicitly adds noise to make the gradient stochastic.

- ☒ True
- ☒ False

16. (2 points) If $\lambda \geq 0$, R is a regularizer, θ is the model parameters, and $\mathcal{D}$ is the training dataset, what is the regularized optimization problem?

- ○ $\min_\theta \mathcal{L}(\theta, \mathcal{D}) - \lambda R(\theta)$
- ○ $\min_\theta \mathcal{L}(\theta, \mathcal{D}) - \lambda R(\mathcal{D})$
- ⊗ $\min_\theta \mathcal{L}(\theta, \mathcal{D}) + \lambda R(\theta)$
- ○ $\min_\theta \mathcal{L}(\theta, \mathcal{D}) + \lambda R(\mathcal{D})$

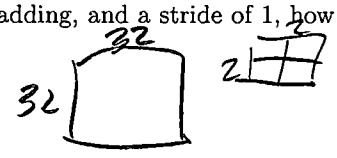
17. (2 points) Which optimization problem is most likely to have exact zeros in the solution for θ?

- ○ $\arg\min_\theta \|y - X\theta\|_2^2$
- ⊗ $\arg\min_\theta \|y - X\theta\|_2^2 + 10\|\theta\|_1^2$
- ○ $\arg\min_\theta \|y - X\theta\|_2^2 + 10\|\theta\|_2^1$
- ○ $\arg\min_\theta \|y - X\theta\|_2^2 + 10\|\theta\|_\infty^3$

18. (2 points) Suppose f is a complex non-linear function (e.g., a deep model), and you know that the unique solution of $\arg\min_z f(z)$ is $z^* = 5$, i.e., $\arg\min_z f(z) = 5$. What is the solution of $\arg\min_v g(v)$, where $g(v) = 3f(4v + 3) - 3$?

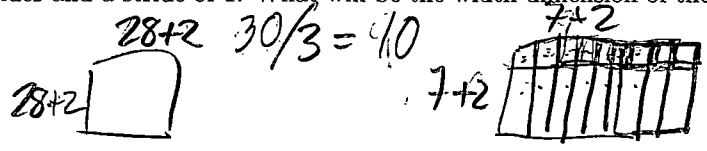
- ○ $v^* = -0.5$ 5
- ⊗ $v^* = 0$
- ○ $v^* = 0.5$
- ○ Knowing the solution given the information in the question is impossible.

19. (2 points) Suppose the input tensor is an CIFAR-10 image (32x32 size with 3 channels) and we apply a "2×2" convolution with 9 filters, no padding, and a stride of 1, how many channels will the output of the convolution have?



Answer: 3

20. (2 points) Suppose the input is a MNIST image (28x28 size with only 1 channel) and we perform a 3x3 convolution with a padding of 1 on all sides and a stride of 1. What will be the width dimension of the output?



Answer:

21. (2 points) If we perform a 1D convolution where the input length is 6 and the filter/kernel is size 2 with no padding and a stride of 2, what will the length of the output be?

Answer: 3

22. You are given a Python function to classify tweets as "spam", "not-spam", or "toxic". Provide this function's predictions for each Tweet below.

```
def classify_tweet(tweet):
    spam_words = ['free', 'buy', 'cheap', 'discount', 'limited', 'offer', 'sale']
    toxic_words = ['hate', 'killer', 'attack', 'terrific', 'abuse', 'bully', 'hurt', 'racist']

    spam_word_count = 0

    for word in tweet:
        if word in spam_words:
            spam_word_count += 1
        if word in toxic_words:
            return 'toxic'

    if spam_word_count > 1:
        return 'spam'
    else:
        return 'not spam'
```

(a) (1 point) It's a limited-time offer, so buy now!
(True label = spam, Predicted label = spam)

(b) (1 point) Best deal with a huge discount on the purchase of this iPhone.
(True label = spam, Predicted label = not spam)

(c) (1 point) That was a killer performance.
(True label = not spam, Predicted label = toxic)

(d) (1 point) They did a terrific job painting my house.
(True label = not spam, Predicted label = toxic)

(e) (3 points) What words should be added or removed (if any) from the above lists to make the classifier accurate on all the samples above? (One or more may be left blank.)

- Add purchase to spam_words
- Add [ ] to [ ]
- Remove killer from toxic_words
- Remove terrific from toxic_words

$m \times p \sim p \times n$

$m \times n$

23. Recall the following equation from the Power Iteration method for calculating PCA of sparse matrces.
$v^{(i)} = X^T(Xv^{(i-1)}) - \mu(1^T(Xv^{(i-1)})) - (X^T 1)(\mu^T v^{(i-1)}) + \mu(1^T 1)(\mu^T v^{(i-1)})$
We initialize the variables in the following manner:

```
# X is of shape (n, d)
v = np.random.randn(X.shape[1])
one_vec = np.ones(X.shape[0])
mu = np.array(np.mean(X, axis=0)).squeeze()
```

→ (d,)

Write the shapes in terms of "n" and "d" where the format of the shape should be "(a, b)" for 2D arrays or "(a,)" for 1D arrays.

(a) (1 point) Shape of v: (d,)

(b) (1 point) Shape of one_vec: (n,)

(c) (1 point) Shape of mu: (d,)

(d) (2 points) Shape of $\mu(1^T 1)(\mu^T v^{(i-1)})$: (n, d)

(e) (2 points) Shape of $X(\mu(1^T 1)(\mu^T v^{(i-1)})X^T 1)$: (n, d)

24. Suppose the following are matrices and their shapes.
$X : (100, 50), Y : (50, 1), Z : (100, 1)$
If the operation is valid, write the shape of the output in the format "(a,b)". If the operation is not valid, write "Invalid".

(a) (2 points) $X^T ZY$: invalid → (50,100)·(100,1)(50,1) → (50,1)·(50,1)

(b) (2 points) $X^T(ZY^T)$: (50, 50) → (50,100)·((100,1)(1,50)) → (50,100)·(100,50)

(c) (2 points) $XYZ^T X^T(ZY^T)$: invalid → (100,50)(50,1)(1,100)(50,100) = ((100,1)(1,5

(d) (2 points) $XYZ^T X(ZY^T)$: invalid ↳ (100,1)(1,100)(50,100)·(100,50)

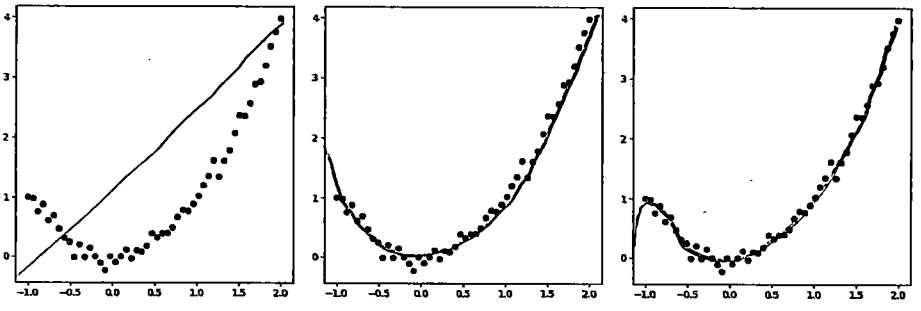
↳ (100,100) (50,100)

↳ (100,100)(100,50)·(100,50)

↳ (100,50)(100,50)

25. We want to fit the given data with varying degrees of the polynomial, i.e., the prediction $\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + ... + \theta_d x^d$.

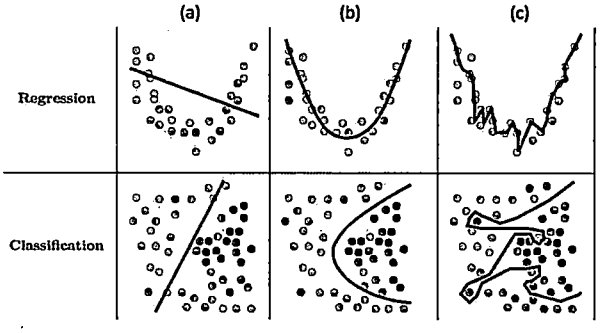
(a) (6 points) For each box below, draw the approximate fitted polynomial for the given polynomial degree.



(a) Polynomial Degree = 0 (b) Polynomial Degree = 1 (c) Polynomial Degree = 2

(b) (1 point) Which degree will likely generalize the best to new data? $d =$ 2

26. We have learned model generalization. Solve the questions below referring to the following figure.



(a) (2 points) Select one that correctly diagnoses (a)-(c)

- ○ (a) overfitting, (b) overfitting, (c) underfitting
- ⊗ (a) underfitting, (b) just right (c) overfitting
- ○ (a) just right, (b) overfitting (c) overfitting
- ○ (a) underfitting, (b) overfitting (c) overfitting

(b) (2 points) Select *all* that correctly describe (a)

- ☒ High training error
- ☐ Low training error
- ☒ High testing error
- ☐ Low testing error

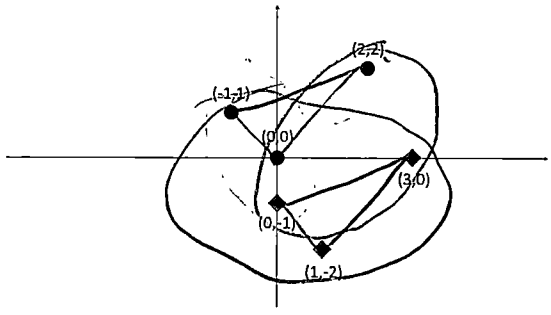
(c) (2 points) Select *all* that correctly describe (c)

- ☐ High training error
- ☒ Low training error
- ☐ High testing error
- ☒ Low testing error

train low = overfitting

27. Let red circles denote class 1 and let blue diamonds denote class 2. In this question, we would like to select the best value of k using cross validation where the number of folds is equal to the number of points (i.e., leave-one-out cross validation). The CV accuracy for $k = 3$ neighbors is 3/6. You will have to compute the CV accuracy for $k = 1$ and $k = 5$ in the next questions. (Note that all points are at integer locations.)

(0,0), (2,2), (-1,1)



(a) (4 points) Compute the CV accuracy for $k = 1$ neighbors in KNN:

- ○ 0
- ⊗ 1/6
- ○ 2/6
- ○ 3/6
- ○ 4/6
- ○ 5/6
- ○ 1

(b) (4 points) Compute the CV accuracy for $k = 5$ neighbors in KNN:

- ○ 0
- ○ 1/6
- ⊗ 2/6
- ○ 3/6
- ○ 4/6
- ○ 5/6
- ○ 1

(c) (1 point) What is the best k:

- ○ $k = 1$
- ⊗ $k = 3$
- ○ $k = 5$

28. Suppose we want to train a deep learning model by using 1000 training samples. Choose the appropriate training setting.

(a) (1 point) Given batch size of 10, what is the number of mini batches?

- ○ 10
- ⊗ 100
- ○ 1,000
- ○ 10,000
- ○ 100,000

$\frac{1000}{10} = 100$

(b) (1 point) If we want to train the model for 100 epochs, how many times the model will go over each training sample?

- ○ 10
- ○ 100
- ○ 1,000
- ⊗ 10,000
- ○ 100,000

$100 \times 100 = 10{,}000$

29. Assume that the goal of the code below is to compute the covariance matrix correctly. However, the current code below contains a bug.

```
# Assume X is a 2D numpy array
n, d = X.shape
x_center = X.mean(axis=0)                                    # (a)
zero_centered_data = X - x_center                            # (b)
cov = (1/n) * np.dot(zero_centered_data.T, zero_centered_data)   # (c)
```

(a) (2 points) Which line will raise an error?

- ○ (a)
- ○ (b)
- ○ (c)

(b) (2 points) Which line should be changed to correctly compute the covariance matrix?

- ○ (a)
- ○ (b)
- ○ (c)

30. Suppose we want to train a 1D linear regression model using a mean squared error (MSE) objective. For simplicity, we will assume that the model does not include a shift/bias term and is simply $f_\theta(x) = \theta x$, where both θ and x are scalars. The training data is two points (note that X only has 1 column since this is a 1D problem):

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 2.0 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.0 \\ 5.0 \end{bmatrix} \tag{1}$$

(a) (2 points) Assuming that $\theta = 1$, what is the gradient of the parameters for sample 1?

$$\nabla_\theta \ell(y_1, f_\theta(x_1)) = \boxed{1.0} \tag{2}$$

(b) (2 points) Assuming that $\theta = 1$, what is the gradient of the parameters for sample 2?

$$\nabla_\theta \ell(y_2, f_\theta(x_2)) = \boxed{3.0} \tag{3}$$

(c) (2 points) Assuming that $\theta = 1$ at the current iteration and the step size is $\eta = 0.2$, what is the updated θ after one step of gradient descent?

$$\theta^{\text{new}} = \boxed{0.8}$$

(d) (2 points) Assuming that $\theta = 1$ at the current iteration, the step size is $\eta = 0.1$, and the batchsize is 1 (where the first sample is in the first batch), what is the updated θ after one step of stochastic gradient descent?

$$\theta^{\text{new}} = \boxed{2.7}$$

$\arg\min_\theta \|y - X\theta\|_2^2$

$2.0 - 1.0 = (1.0)^2 = \sqrt{1} = 1$

$2.0 - 5.0 = (-3.0)^2 = 3$

$1 - 0.2(1) = 0.8$

$3.0 - \frac{1}{1}\ell(3.0)(0.1) = 0.3$

$3 - 0.3 = 2.7$