

CJAIO Part 1 - Practice Set 2

Questions

Time	60 minutes	Total	100 points
Single Choice	12 x 2 = 24	Multiple Choice	6 x 4 = 24
Short Answer I	4 x 5 = 20	Short Answer II	4 x 8 = 32

Instructions: For Questions 1-12, select one answer. For Questions 13-18, select all correct answers. For Questions 19-26, show enough work or code to justify your answer. Multiple-choice questions each have five options.

Section A - Single-Choice Questions (24 points)

1. What is printed by the following code?

```
nums = [3, 8, 5, 2, 9]
print(nums[1:4])
```

- A. [8, 5, 2]
- B. [3, 8, 5]
- C. [8, 5, 2, 9]
- D. [5, 2, 9]
- E. An error occurs.

2. Which task is most naturally a regression problem?

- A. Predicting whether an email is spam
- B. Predicting tomorrow's temperature in degrees Celsius
- C. Predicting whether a photo contains a bicycle
- D. Grouping customers into three clusters
- E. Predicting whether a transaction is fraudulent

3. What is printed by the following code?

```
s = "DATA"
result = ""

for i in range(len(s)):
    if i % 2 == 0:
        result += s[i]
    else:
        result += str(i)

print(result)
```

- A. D1T3
- B. DA13
- C. D1A3
- D. DT
- E. DATA

4. What is printed by the following program?

```
data = ["A", "B", "A", "C", "B", "A"]
freq = {}

for x in data:
    freq[x] = freq.get(x, 0) + 1

print(freq["A"] * freq["B"] - freq["C"])
```

- A. 3
- B. 4
- C. 5
- D. 6
- E. 7

5. A bag contains 4 red balls and 3 blue balls. Two balls are selected without replacement. What is the probability that both selected balls are red?

- A. $1/7$
- B. $2/7$
- C. $3/7$
- D. $1/2$
- E. $4/7$

6. A sensor dataset records some temperatures in Celsius and others in Fahrenheit in the same column. What is the best first step before computing summary statistics?

- A. Delete every Fahrenheit record
- B. Convert all measurements to one consistent unit
- C. Replace all values by the mean
- D. Sort the column from smallest to largest
- E. Convert the temperature column into text

7. What is printed by the following code?

```
a = [2, 5, 8, 12, 15, 21]
target = 11
left = 0
right = len(a) - 1

while left <= right:
    mid = (left + right) // 2
    if a[mid] < target:
        left = mid + 1
    else:
        right = mid - 1

print(left)
```

- A. 2
- B. 3
- C. 4
- D. 5
- E. 6

8. What is the final value of total?

```
total = 0

for i in range(1, 6):
    for j in range(i):
        if (i + j) % 3 == 0:
            total += i
```

- A. 14
- B. 15
- C. 17
- D. 19
- E. 20

9. What is printed?

```
a = [1, 2]
b = [a, a[:]]

b[0].append(3)
b[1][0] = 9

print(a, b)
```

- A. [1, 2] [[1, 2, 3], [9, 2]]
- B. [1, 2, 3] [[1, 2, 3], [9, 2]]
- C. [9, 2, 3] [[9, 2, 3], [9, 2]]
- D. [1, 2, 3] [[1, 2, 3], [9, 2, 3]]
- E. An error occurs.

10. A K-nearest-neighbors classifier uses Euclidean distance and $k = 3$. A new point is $P = (2, 2)$. Which class is predicted?

Point	Class
-------	-------

(1, 2)	A
(2, 4)	B
(3, 2)	B
(0, 0)	A
(4, 3)	B

- A. Class A
- B. Class B
- C. A tie between A and B
- D. It cannot be determined without retraining
- E. KNN cannot be used with two features

11. A decision-tree split produces two children of equal size. The left child contains 4 positive and 0 negative examples. The right child contains 1 positive and 3 negative examples. Using $Gini = 1 - p_1^2 - p_2^2$, what is the weighted Gini impurity after the split?

- A. 0
- B. 1/8
- C. 3/16
- D. 1/4
- E. 3/8

12. What is printed by the following sorting program?

```
nums = [5, 2, 4, 1]
swaps = 0

for i in range(len(nums) - 1):
    for j in range(len(nums) - 1 - i):
        if nums[j] > nums[j + 1]:
            nums[j], nums[j + 1] = nums[j + 1], nums[j]
            swaps += 1

print(swaps, nums[1] + nums[-1])
```

- A. 4 7
- B. 5 6
- C. 5 7
- D. 6 7
- E. 6 9

Section B - Multiple-Choice Questions (24 points)

Select ALL correct answers. Each question has one or more correct choices.

13. Which statements about Python are correct?

- A. `list(range(1, 6, 2))` produces `[1, 3, 5]`.
- B. Dictionary keys must be unique.
- C. A Python set keeps duplicate copies of the same value.
- D. `"abc"[::-1]` evaluates to `"cba"`.
- E. If `b = a` for a list `a`, appending to `b` can never affect `a`.

14. Which actions may be appropriate when cleaning a real-world dataset?

- A. Check for duplicate records.
- B. Investigate why values are missing.
- C. Replace every unusual numerical value with the mean automatically.
- D. Convert measurements to consistent units.
- E. Delete every row containing any outlier.

15. For the dataset 2, 4, 6, 8, which statements are correct?

- A. The median is 4.
- B. The mean is 5.
- C. The range is 6.
- D. Adding 3 to every value leaves the range unchanged.
- E. Multiplying every value by 2 leaves the mean unchanged.

16. Which statements about algorithms are correct?

- A. Standard binary search requires the data to be ordered appropriately.
- B. In the worst case, linear search may inspect every element.
- C. Selection sort always makes exactly n^2 comparisons on n elements.
- D. A queue normally follows first-in-first-out behavior.
- E. A recursive solution is always faster than an iterative solution.

17. Which statements about machine learning are correct?

- A. Feature scaling can affect KNN because KNN uses distances.
- B. K-Means requires a class label for every training example.
- C. A decision tree can overfit its training data.
- D. 100% training accuracy guarantees 100% test accuracy.
- E. Linear regression is normally used to predict numerical values.

18. A student is building a classifier from data containing missing values, duplicate rows, and an imbalanced target. Which practices are appropriate?

- A. Split into training and test sets before fitting preprocessing steps such as scaling or imputation.
- B. Fit the scaler on the full dataset because scaling does not use class labels.
- C. Check that duplicate copies of the same record do not appear in both training and test sets.
- D. Use metrics beyond accuracy when one class is much more common than the other.
- E. Repeatedly choose the model that scores highest on the test set until the score is satisfactory.

Section C - Short Answer I (20 points)

Questions 19-22 are worth 5 points each.

19. What is printed by the following code?

```
data = [6, 3, 8, 5, 10]
total = 0

for i in range(len(data)):
    if data[i] % 2 == 0:
        total += data[i] - i

print(total)
```

20. Complete the two blanks so that the program counts how many values are strictly greater than the average. Then give the final value printed.

```
nums = [4, 7, 10, 5, 9]

average = _____
count = 0

for x in nums:
    if _____:
        count += 1

print(count)
```

21. Trace the binary search. Find the final values of left, right, and checks.

```
a = [2, 5, 8, 12, 16, 23, 38]
target = 15
left = 0
right = len(a) - 1
checks = 0

while left <= right:
    mid = (left + right) // 2
    checks += 1

    if a[mid] == target:
        break
    elif a[mid] < target:
        left = mid + 1
    else:
        right = mid - 1
```

22. Two linear-regression models are evaluated on the data below. Calculate the predictions and the mean squared error (MSE) for each model, then state which model has the lower MSE.

Data: $(x, y) = (1, 3), (2, 4), (3, 8)$

Model A: $\hat{y} = 2x + 1$

Model B: $\hat{y} = 2.5x$

Use $MSE = [\text{sum of } (y - \hat{y})^2] / n$.

Section D - Short Answer II (32 points)

Questions 23-26 are worth 8 points each. Partial credit may be awarded for correct intermediate work.

23. K-Means uses two initial centers $C1 = (1, 1)$ and $C2 = (7, 5)$ for the five points:

$(1, 1), (2, 1), (6, 4), (7, 5), (5, 4)$

(a) Assign each point to its nearest center using Euclidean distance. You may compare squared distances. [4 marks]

(b) Calculate the new center $C1$. [2 marks]

(c) Calculate the new center $C2$. [2 marks]

24. A logistic-regression model computes

$$z = 0.8x_1 - 0.6x_2 + 0.2$$

$$p = 1 / (1 + e^{(-z)})$$

For $x_1 = 2$ and $x_2 = 1$, use $e^{(-1.2)}$ approximately 0.301.

(a) Calculate z . [2 marks]

(b) Calculate p to three decimal places. [3 marks]

(c) What class is predicted using threshold 0.5? [1 mark]

(d) What class is predicted using threshold 0.8? [2 marks]

25. The list below contains repeated values. Write Python code that finds the most frequent value. If two or more values have the same highest frequency, print the smallest such value. You may use loops, dictionaries, and conditionals, but not `collections.Counter`.

`data = [4, 2, 4, 3, 2, 2, 4, 5, 5]`

Your program must print the correct result for this data. [8 marks]

26. A KNN classifier has two features with known ranges:

$$0 \leq x_1 \leq 10$$

$$0 \leq x_2 \leq 1000$$

Use min-max scaling: $x_{\text{scaled}} = (x - \min) / (\max - \min)$.

The new point is $P = (5, 500)$. The labeled points are:

$A = (4, 700)$, Class A

$B = (9, 510)$, Class B

$C = (6, 520)$, Class B

$D = (2, 480)$, Class A

(a) Scale P and the four labeled points. [3 marks]

(b) Using $k = 3$ and squared Euclidean distance on the scaled features, identify the three nearest neighbors and predict the class of P . [3 marks]

(c) If the same 3-NN calculation were done on the unscaled features, what class would be predicted? Briefly explain why the result changes. [2 marks]