

CJAIO Part 1 - Practice Set 1

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 = [6, 1, 8, 3, 7]
print(nums[2:5:2])
```

- A. [6, 8]
- B. [8, 7]
- C. [8, 3, 7]
- D. [1, 3]
- E. An error occurs.

2. Which task is most naturally a regression problem?

- A. Predicting whether a message is spam
- B. Predicting the number of visitors to a website tomorrow
- C. Predicting whether an image contains a dog
- D. Grouping songs into clusters
- E. Predicting whether a payment is fraudulent

3. What is printed by the following program?

```
s = "PYTHON"
result = ""

for i in range(1, len(s), 2):
    result += s[i]

print(result)
```

- A. PT0
- B. YHN
- C. PYT
- D. YH0
- E. THON

4. What is printed by the following program?

```
data = ["red", "blue", "red", "green", "blue", "red", "green"]
freq = {}

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

print(freq["red"] + 2 * freq["green"])
```

- A. 5
- B. 6
- C. 7
- D. 8
- E. 9

5. A bag contains 3 white balls and 5 black balls. Two balls are selected without replacement. What is the probability that exactly one selected ball is white?

- A. $3/14$
- B. $5/14$
- C. $15/28$
- D. $1/2$
- E. $5/8$

6. For the dataset 4, 5, 7, 8, 26, which statement is correct?

- A. Mean = 7 and median = 10
- B. Mean = 10 and median = 7
- C. Mean = 10 and median = 8
- D. Mean = 7 and median = 8
- E. Mean = 8 and median = 10

7. What is printed by the following code?

```
a = [3, 6, 9, 14, 18, 22]
target = 13
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. 1
- B. 2
- C. 3
- D. 4
- E. 5

8. What is the final value of total?

```
total = 0

for i in range(1, 7):
    for j in range(1, i + 1):
        if (i * j) % 4 == 0:
            total += 1
```

- A. 6
- B. 7
- C. 8
- D. 9
- E. 10

9. What is printed?

```
a = [[1, 2], [3, 4]]
b = a[:]
```

```
b[0][1] = 9
b.append([5, 6])
```

```
print(a, b)
```

- A. [[1, 2], [3, 4]] [[1, 9], [3, 4], [5, 6]]
- B. [[1, 9], [3, 4]] [[1, 9], [3, 4], [5, 6]]
- C. [[1, 9], [3, 4], [5, 6]] [[1, 9], [3, 4], [5, 6]]
- D. [[1, 2], [3, 4], [5, 6]] [[1, 9], [3, 4], [5, 6]]
- E. An error occurs.

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

Point	Class
(2, 2)	A
(4, 2)	B
(3, 4)	B
(1, 2)	A
(3, 1)	A

- A. Class A
- B. Class B
- C. A tie between A and B
- D. It cannot be determined
- E. KNN cannot be used with two features

11. A decision-tree split produces a left child with 3 positive and 0 negative examples, and a right child with 2 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. $3/20$
- C. $1/4$
- D. $3/10$
- E. $12/25$

12. What is printed by the following program?

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

for i in range(1, len(nums)):
    key = nums[i]
    j = i - 1
    while j >= 0 and nums[j] > key:
        nums[j + 1] = nums[j]
        moves += 1
        j -= 1
    nums[j + 1] = key

print(moves, nums[0] + nums[-1])
```

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

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(2, 9, 3))` produces `[2, 5, 8]`.
- B. `"AI" * 2` evaluates to `"AIAI"`.
- C. Dictionary keys are unique.
- D. `len(set([2, 2, 3]))` is 3.
- E. If `b = a.copy()` for a list `a`, appending to `b` also appends to `a`.

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

- A. Check for duplicate records.
- B. Investigate why some values are missing.
- C. Automatically replace every outlier with the mean.
- D. Convert measurements to consistent units.
- E. Standardize inconsistent date formats.

15. For the dataset 1, 3, 5, 7, 9, which statements are correct?

- A. The median is 6.
- B. The mean is 5.
- C. The range is 8.
- D. Adding 4 to every value leaves the range unchanged.
- E. Multiplying every value by 2 leaves the mean unchanged.

16. Which statements about algorithms and data structures are correct?

- A. Binary search can discard roughly half of the remaining search range after a comparison when the data is ordered.
- B. A queue normally follows first-in-first-out behavior.
- C. A stack removes the oldest item first.
- D. The number of pairs (i, j) satisfying $0 \leq i < j < n$ is $n(n - 1)/2$.
- E. Linear search is always faster than binary search on a sorted list.

17. Which statements about machine learning are correct?

- A. Feature scaling can affect KNN because KNN uses distances.
- B. K-Means is a supervised classification algorithm.
- C. A decision tree can overfit its training data.
- D. The sigmoid function in logistic regression can output a negative probability.
- E. After the sigmoid function, logistic regression produces a value between 0 and 1.

18. A student is building a classifier from a dataset with missing values, duplicate records, and a highly imbalanced target. Which practices are appropriate?
- A. Split the data before fitting preprocessing steps such as imputation or scaling.
 - B. Repeatedly use the test set to choose model settings.
 - C. Check whether duplicate copies of the same record appear in both training and test sets.
 - D. Use metrics beyond accuracy when one class is much more common than the other.
 - E. Removing one sensitive feature always guarantees a fair model.

Section C - Short Answer I (20 points)

Questions 19-22 are worth 5 points each.

19. What is printed by the following code?

```
values = [2, 7, 4, 9, 6]
score = 0

for i in range(len(values)):
    if values[i] % 2 == 0:
        score += values[i] // 2
    else:
        score -= i

print(score)
```

20. Complete the two blanks so that the program removes duplicates while preserving the first occurrence of each value. Then give the final value printed.

```
data = [5, 3, 5, 8, 3, 9]
seen = set()
result = []

for x in data:
    if _____:
        result.append(x)
        _____

print(sum(result))
```

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

```
a = [1, 5, 9, 13, 17, 21, 25, 30]
target = 20
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 each model's three predictions and its mean squared error (MSE). Which model has the lower MSE?

Data: $(x, y) = (1, 3), (2, 6), (3, 8)$

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

Model B: $\hat{y} = 2.5x + 0.5$

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 $C1 = (0, 1)$ and $C2 = (7, 5)$ for the five points:

$(0, 1), (1, 1), (5, 4), (6, 4), (7, 5)$

(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

$$z = 0.9x_1 - 0.5x_2 - 0.4$$

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

For $x_1 = 2$ and $x_2 = 1$, use $e^{(-0.9)}$ approximately 0.407.

(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 0.75? [2 marks]

25. The list below contains several increasing runs:

`data = [3, 5, 7, 2, 4, 6, 8, 1]`

Write Python code using loops and conditionals that prints the length of the longest contiguous strictly increasing run. For example, 2, 4, 6, 8 is one such run. Your program must print the correct result for the given data. [8 marks]

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

$$0 \leq x_1 \leq 20$$

$$0 \leq x_2 \leq 2000$$

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

The new point is $P = (10, 500)$. The labeled points are:

$A = (8, 900)$, Class A

$B = (18, 520)$, Class B

$C = (12, 100)$, Class A

$D = (4, 490)$, Class B

(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]