

2026-27 CAIO National Qualifier Practical Competition

Part 1 — AI Fundamentals Challenge (Paper B)

Total: 100 marks

Instructions

- Answer all 35 questions.
- Questions 1-15 are single-choice questions (2 marks each). Choose the one best answer.
- Questions 16-25 are multiple-select questions (2 marks each). Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected; otherwise you receive zero for that question.
- Questions 26-35 are fill-in-the-blank questions (5 marks each). Each has exactly five blanks; write your answer for each blank clearly.
- A non-programmable calculator may be used. Total: 100 marks.

Do not open or begin until instructed by the proctor.

Question 1

Data Preparation

2 Points

Consider the following Python code:

```
import numpy as np
a = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
mask = a > 4
b = a[mask]
b[0] = 99
c = a.flatten(order='F')
print(a[1, 2], b[0], c[1], mask.sum())
```

What is the printed output?

- A. 99 99 4 5
- B. 6 99 5 5
- C. 6 99 4 5
- D. 6 5 4 5
- E. 99 99 5 5

Question 2

Data Preparation

2 Points

Consider the following Python code:

```
import pandas as pd
left = pd.DataFrame({'id': [1, 2, 3], 'score': [80, 90, 70]})
right = pd.DataFrame({'id': [2, 3, 4], 'club': ['A', 'B', 'C']})
m = left.merge(right, on='id', how='left')
print(len(m), m['club'].isna().sum(), int(m['score'].sum()))
```

What is the printed output?

- A. 3 1 240
- B. 2 0 170
- C. 4 1 240
- D. 3 0 240
- E. 3 1 170

Question 3

Supervised Learning

2 Points

For binary logistic regression the cross-entropy loss is

$$L(w, b) = -(1/n) \sum_i [y_i \log p_i + (1-y_i) \log(1-p_i)], \quad p_i = \sigma(w \cdot x_i + b)$$

The gradient of L with respect to w is:

- A. $-(1/n) \sum_i (p_i - y_i) x_i$
- B. $(1/n) \sum_i (p_i - y_i) x_i$
- C. $(1/n) \sum_i (y_i - p_i)$
- D. $(2/n) \sum_i (p_i - y_i) x_i$
- E. $(1/n) \sum_i p_i (1 - p_i) x_i$

Question 4

Supervised Learning

2 Points

You train k-nearest-neighbors classifiers on the same dataset with different values of k.

Which statement is TRUE?

- A. $k = 1$ gives the smoothest decision boundary.
- B. A larger k always improves test accuracy.
- C. kNN with $k = n$ (all training points) has the lowest bias.
- D. Increasing k typically increases bias and decreases variance.
- E. kNN needs no feature scaling because Euclidean distance is scale-invariant.

Question 5

Supervised Learning

2 Points

Batch gradient descent minimizes $J(w) = w^2$ from $w_0 = 4$ with the update $w := w - \alpha \cdot J'(w)$.

For which learning rate α does the sequence diverge ($|w| \rightarrow \infty$)?

- A. 0.1
- B. 0.5
- C. 1.5
- D. 0.9
- E. 1.0

Question 6

Supervised Learning

2 Points

A decision-tree node holds 10 samples (6 of class A, 4 of class B). A candidate split sends 5 samples left (4 A, 1 B) and 5 right (2 A, 3 B).

The weighted Gini impurity after the split is:

- A. 0.48
- B. 0.40
- C. 0.32
- D. 0.50
- E. 0.44

Question 7

Evaluation of ML Models

2 Points

A 3-class classifier has per-class precision and recall: class 1: $P=1.00$, $R=0.50$; class 2: $P=0.50$, $R=1.00$; class 3: $P=0.80$, $R=0.80$.

The macro-averaged F1 score is:

- A. 0.67
- B. 0.80
- C. 0.73
- D. 0.71
- E. 0.77

Question 8

Evaluation of ML Models

2 Points

You are building a fraud-detection model on a dataset with 1% fraudulent transactions.

Which evaluation choice is most appropriate?

- A. Accuracy, because it is the simplest and most intuitive metric.
- B. Precision-recall analysis (e.g., F1 or PR-AUC), because accuracy can be misleading on imbalanced data.
- C. Training accuracy, because it reflects the model's full capacity.
- D. ROC-AUC must never be used on imbalanced data.
- E. Mean squared error computed on the predicted class labels.

Question 9

Unsupervised Learning

2 Points

PCA is applied to centered data whose covariance matrix has eigenvalues 6, 3, 1.

What fraction of the total variance is retained by projecting onto the first two principal components?

- A. 0.60
- B. 0.30
- C. 0.90
- D. 1.00
- E. 0.67

Question 10

Kernel Methods

2 Points

A kernel is valid (Mercer) if it equals an inner product $\langle \phi(x), \phi(z) \rangle$ for some feature map ϕ .

Which of the following is NOT a valid Mercer kernel?

- A. $k(x,z) = (x \cdot z + 2)^2$
- B. $k(x,z) = \exp(-2\|x - z\|^2)$
- C. $k(x,z) = (x \cdot z)^3$
- D. $k(x,z) = -(x \cdot z)^2$
- E. $k(x,z) = x \cdot z$

Question 11

Kernel Methods

2 Points

A hard-margin SVM has support vectors $x_1=(1,1)$ with $y_1=+1$, $\alpha_1=1$ and $x_2=(-1,-1)$ with $y_2=-1$, $\alpha_2=1$.

The weight vector w of the SVM equals:

- A. (2, 2)
- B. (0, 0)
- C. (1, 1)
- D. (2, -2)
- E. (1, -1)

Question 12

AI Search

2 Points

Greedy best-first search expands nodes in order of increasing heuristic value $h(n)$, ignoring the path cost so far.

Which statement is TRUE?

- A. Greedy best-first search always finds the optimal path.
- B. Greedy best-first search is optimal whenever the heuristic is admissible.
- C. A* with an admissible but inconsistent heuristic never re-expands a node.
- D. Greedy best-first search always expands fewer nodes than A* with the same heuristic.
- E. On a finite graph, greedy best-first search with a closed (visited) set is complete, though not optimal.

Question 13

Logical Reasoning

2 Points

Knowledge base $KB = \{ \neg P \vee Q, \neg Q \vee R, P \}$.

Which formula is logically entailed by KB?

- A. $\neg R$
- B. R
- C. $\neg P$
- D. $Q \wedge \neg R$
- E. $\neg Q$

Question 14

Constraint Satisfaction Problems

2 Points

Variables X, Y with domains $D(X) = \{1, 2, 3\}$, $D(Y) = \{1, 2, 3\}$ and a single constraint $X < Y$.

After enforcing arc consistency, the domains are:

- A. $D(X) = \{1, 2, 3\}$, $D(Y) = \{1, 2, 3\}$
- B. $D(X) = \{1, 2\}$, $D(Y) = \{1, 2, 3\}$
- C. $D(X) = \{1, 2\}$, $D(Y) = \{2, 3\}$
- D. $D(X) = \{1\}$, $D(Y) = \{3\}$
- E. $D(X) = \{2, 3\}$, $D(Y) = \{1, 2\}$

Question 15

Reinforcement Learning

2 Points

An agent in state s takes action a , receives reward $r = 4$, and moves to s' . Currently $Q(s, a) = 2$. In s' , $Q(s', a_1) = 1$, $Q(s', a_2) = 5$, $Q(s', a_3) = 3$. Learning rate $\alpha = 0.5$, discount $\gamma = 0.8$.

After one Q-learning update, $Q(s, a)$ becomes:

- A. 5.0
- B. 5.5
- C. 4.2
- D. 8.0
- E. 3.0

Question 16

Data Preparation

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about handling missing data are TRUE?

- A. Filling missing values with the column mean reduces that column's variance compared with the complete data.
- B. Dropping every row that contains a missing value can bias results if the missingness depends on unobserved values.
- C. Forward-fill is always the best imputation strategy for time-series data.
- D. Imputation statistics (e.g., the mean) should be computed on the training set only, then applied to the test set.
- E. A missing-value indicator column can never help a model.

Question 17

Data Preparation

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about feature encoding and scaling are TRUE?

- A. One-hot encoding a categorical feature with k levels creates k binary columns.
- B. Label-encoding (0, 1, 2, ...) a nominal feature like "color" can mislead a linear model into assuming an ordering.
- C. Standardization (zero mean, unit variance) is required before training a decision tree.
- D. Min-max scaling to [0, 1] keeps kNN and RBF-kernel SVM from being dominated by large-range features.
- E. Target encoding never causes leakage because it only uses the target column.

Question 18

Supervised Learning

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about logistic regression are TRUE?

- A. The decision boundary of binary logistic regression is linear in the feature space.
- B. The cross-entropy loss for logistic regression is convex, so gradient descent converges to the global optimum.
- C. Logistic regression outputs can be interpreted as estimated class probabilities.
- D. Logistic regression cannot handle more than two classes.
- E. Adding L2 regularization always increases training accuracy.

Question 19

Supervised Learning

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about Naive Bayes are TRUE?

- A. Naive Bayes assumes the features are conditionally independent given the class.
- B. Laplace (add- α) smoothing prevents zero probabilities for unseen feature-class combinations.
- C. Gaussian Naive Bayes models each continuous feature with a Gaussian per class.
- D. Naive Bayes can only be used with binary features.
- E. Violating the independence assumption always makes Naive Bayes worse than logistic regression.

Question 20

Unsupervised Learning

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about PCA are TRUE?

- A. Principal components are eigenvectors of the covariance matrix, ordered by decreasing eigenvalue.
- B. The first principal component is the direction of maximum variance in the data.
- C. PCA is unaffected by the scale of the features.
- D. Projecting onto all principal components loses information.
- E. PCA can denoise data by dropping low-variance components.

Question 21

Evaluation of ML Models

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about model evaluation methodology are TRUE?

- A. The test set should be used only once, for the final performance estimate.
- B. Hyperparameters should be tuned on the validation set, not the test set.
- C. k-fold cross-validation reduces the variance of the performance estimate compared with a single train/validation split.
- D. Stratified k-fold is only useful for regression problems.
- E. Computing preprocessing statistics (e.g., the scaler mean) on the full dataset before splitting is harmless.

Question 22

Kernel Methods

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about SVMs and kernels are TRUE?

- A. The kernel (Gram) matrix of a valid kernel is always positive semidefinite.
- B. In the soft-margin SVM, a larger C penalizes margin violations more heavily.

- C. The polynomial kernel $(x \cdot z + c)^d$ corresponds to a finite-dimensional feature space.
- D. Support vectors are exactly the training points lying strictly inside the margin.
- E. The RBF kernel can be written as a finite polynomial in the input features.

Question 23

AI Search

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about search are TRUE?

- A. Iterative deepening DFS combines the memory efficiency of DFS with the optimality (for unit step costs) of BFS.
- B. If $h_1(n) \leq h_2(n) \leq h^*(n)$ for all n , A^* with h_2 expands no more nodes than A^* with h_1 (up to tie-breaking).
- C. Tree search can revisit the same state many times, while graph search avoids this with a closed set.
- D. A consistent heuristic can overestimate the true cost to the goal.
- E. Depth-limited search is complete in infinite state spaces.

Question 24

Recommender Systems

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about recommender systems are TRUE?

- A. Matrix factorization learns latent user and item vectors whose dot product predicts ratings.
- B. Implicit feedback (clicks, views) can train a recommender, but it does not directly reveal negative preference.
- C. Popularity bias can cause a recommender to over-recommend already-popular items.
- D. Matrix factorization requires explicit star ratings and cannot use implicit feedback.
- E. Precision@k counts how many relevant items appear anywhere in the full catalog.

Question 25

Logical Reasoning

2 Points

Select ALL correct options. You receive full marks only if every correct option — and no incorrect option — is selected.

Which statements about propositional logic are TRUE?

- A. Modus ponens: from $P \rightarrow Q$ and P , infer Q .
- B. Affirming the consequent — from $P \rightarrow Q$ and Q , infer P — is a valid inference rule.

- C. The contrapositive $\neg Q \rightarrow \neg P$ is logically equivalent to $P \rightarrow Q$.
- D. De Morgan's law: $\neg(P \wedge Q)$ is equivalent to $\neg P \vee \neg Q$.
- E. A formula that is true under every assignment is called satisfiable.

Question 26

Data Preparation

5 Points

Movie Ratings Pipeline (fill in the 5 blanks)

movies.csv columns: title, year, genre, rating (float, may be NaN).

Clean the titles, drop duplicate (title, year) rows, fill missing ratings with the genre median, then rank genres with at least 10 movies by mean rating.

```
import pandas as pd
df = pd.read_csv('movies.csv')
df['title'] = df['title']_____
df = df._____(subset=['title', 'year'], keep='first')
df['rating'] = df.groupby('genre')['rating']._____(lambda s:
s.fillna(s.median()))
g = df.groupby('genre')['rating']._____(['mean', 'count'])
g = g[g['count'] >= 10]._____('mean', ascending=False)
```

Write your answer for each blank below:

- (a) _____
- (b) _____
- (c) _____
- (d) _____
- (e) _____

Question 27

Data Preparation

5 Points

Outlier Removal with NumPy (fill in the 5 blanks)

Standardize the columns of a numeric CSV, drop rows containing any $|z| \geq 3$ outlier, then find the 5 largest values in column 0.

```
import numpy as np
X = np.loadtxt('data.csv', _____=',')
mu = X._____(axis=0)
sd = X.std(axis=0)
Z = (X - mu) / sd
ok = np._____(Z).max(axis=1) < 3
clean = X[_____]
k = 5
```

```
idx = np._____(clean[:, 0])[-k:]
print(round(float(clean[idx][-1, 0]), 3), len(clean))
```

Write your answer for each blank below:

- (a) _____
- (b) _____
- (c) _____
- (d) _____
- (e) _____

Question 28

Supervised Learning

5 Points

Polynomial Ridge with Cross-Validation (fill in the 5 blanks)

Fit a degree-3 polynomial ridge regression with 5-fold cross-validation. X, y are already loaded.

```
from sklearn.preprocessing import PolynomialFeatures, StandardScaler
from sklearn.linear_model import Ridge
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import cross_val_score
model = make_pipeline(
    PolynomialFeatures(degree=_____, include_bias=_____),
    StandardScaler(),
    Ridge(alpha=_____)
)
scores = cross_val_score(model, X, y, cv=_____, scoring=_____)
print(round(-scores.mean(), 3), round(scores.std(), 3))
```

Write your answer for each blank below:

- (a) _____
- (b) _____
- (c) _____
- (d) _____
- (e) _____

Question 29

Supervised Learning

5 Points

Naive Bayes by Hand (fill in the 5 blanks)

Spam filter, binary word features, Laplace smoothing $\alpha=1$.

Training: 10 emails — 4 spam, 6 ham.

“offer” appears in 3 spam, 1 ham; “meeting” in 1 spam, 4 ham.

Classify a new email containing “offer” but NOT “meeting”. Round to 2–3 decimals.

$P(\text{spam}) = \underline{\hspace{2cm}}$ # prior
 $P(\text{offer} \mid \text{spam}) \approx \underline{\hspace{2cm}}$ # Laplace $\alpha=1$
 $P(\text{meeting} \mid \text{spam}) \approx \underline{\hspace{2cm}}$
 unnormalized score(ham) $\approx \underline{\hspace{2cm}}$
 predicted class: $\underline{\hspace{2cm}}$

Write your answer for each blank below:

- (a) $\underline{\hspace{2cm}}$
- (b) $\underline{\hspace{2cm}}$
- (c) $\underline{\hspace{2cm}}$
- (d) $\underline{\hspace{2cm}}$
- (e) $\underline{\hspace{2cm}}$

Question 30

Unsupervised Learning

5 Points

PCA Eigenvalues by Hand (fill in the 5 blanks)

2-D centered data has covariance matrix $\begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$.

Hint: $\sqrt{17} \approx 4.123$. Round eigenvalues to 2 decimals.

covariance = $\begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$
 trace = $\underline{\hspace{2cm}}$
 det = $\underline{\hspace{2cm}}$
 λ_1 (larger eigenvalue) $\approx \underline{\hspace{2cm}}$
 λ_2 (smaller eigenvalue) $\approx \underline{\hspace{2cm}}$
 variance fraction explained by PC1 $\approx \underline{\hspace{2cm}}$

Write your answer for each blank below:

- (a) $\underline{\hspace{2cm}}$
- (b) $\underline{\hspace{2cm}}$
- (c) $\underline{\hspace{2cm}}$
- (d) $\underline{\hspace{2cm}}$
- (e) $\underline{\hspace{2cm}}$

Question 31

AI Search

5 Points

Uniform-Cost Search Trace (fill in the 5 blanks)

Uniform-cost search from S to G.

Edges: $S \rightarrow A$ (3), $S \rightarrow B$ (1), $A \rightarrow C$ (1), $B \rightarrow D$ (4), $B \rightarrow C$ (5), $C \rightarrow G$ (2), $D \rightarrow G$ (1).

1st node expanded: _____

2nd node expanded: _____

3rd node expanded: _____

cost of the path found to G: _____

the optimal path is: _____ (e.g. $S \rightarrow B \rightarrow D \rightarrow G$)

Write your answer for each blank below:

(a) _____

(b) _____

(c) _____

(d) _____

(e) _____

Question 32

Logical Reasoning

5 Points

Resolution Proof (fill in the 5 blanks)

Convert each formula to CNF clauses, then refute the negated goal.

Prove $P \rightarrow R$ from $KB = \{ P \rightarrow Q, Q \rightarrow R \}$ by resolution.

clause (1): _____

clause (2): _____

clause (3): _____

clause (4): _____

resolve (3)+(1) $\rightarrow Q$; resolve with (2) $\rightarrow R$; resolve with (4)

yields the _____ clause

Write your answer for each blank below:

(a) _____

(b) _____

(c) _____

(d) _____

(e) _____

Question 33

Constraint Satisfaction Problems

5 Points

Arc Consistency and Counting (fill in the 5 blanks)

Variables X, Y, Z , each with domain $\{1, 2, 3, 4\}$.Constraints: $X < Y$, $Y \neq Z$.

value removed from $D(X)$ by arc consistency: _____
 value removed from $D(Y)$ by arc consistency: _____
 $D(X)$ after arc consistency: _____ (e.g. 1,2,3)
 number of complete assignments satisfying all constraints: _____
 assign $X=2$; after forward checking, $D(Y) =$ _____

Write your answer for each blank below:

- (a) _____
 (b) _____
 (c) _____
 (d) _____
 (e) _____

Question 34

Kernel Methods

5 Points

Polynomial Kernel by Hand (fill in the 5 blanks)

Polynomial kernel $k(x,z) = (x \cdot z + 2)^2$. Compute by hand.

$k(x, z) = (x \cdot z + 2)^2$
 $x=(1,2), z=(3,-1): x \cdot z = \underline{\hspace{2cm}} ; k(x,z) = \underline{\hspace{2cm}}$
 $x=(2,0), z=(1,1): x \cdot z = \underline{\hspace{2cm}} ; k(x,z) = \underline{\hspace{2cm}}$
 dimension of the explicit feature map ϕ for degree 2 in 2-D inputs: _____

Write your answer for each blank below:

- (a) _____
 (b) _____
 (c) _____
 (d) _____
 (e) _____

Question 35

Supervised Learning

5 Points

One Gradient-Descent Step by Hand (fill in the 5 blanks; 2 decimals)

Do one batch gradient-descent step by hand.

```
points (1,1), (2,2), (3,2); w = 0, b = 0, α = 0.1
J(w,b) = (1/2m) Σi (wxi + b - yi)2
∂J/∂w = _____ (2 decimals)
∂J/∂b = _____
after one GD step: w = _____ ; b = _____
new mean squared error = _____
```

Write your answer for each blank below:

- (a) _____
- (b) _____
- (c) _____
- (d) _____
- (e) _____