Boğaziçi University

Institute for Data Science and Artificial Intelligence

Graduate Science Exam Questions

Booklet A

The exam consists of four sections, with a total of 40 questions:

- **Probability and Statistics for Data Science and Artificial Intelligence** (Questions 1-10)
- **Mathematics for Data Science and Artificial Intelligence** (Questions 11-20)
- **Introduction to Programming with Python** (Questions 21-30)
- **Python for Data Science and Artificial Intelligence** (Questions 31-40)

Each question has one correct answer. FOUR INCORRECT ANSWERS CANCEL OUT ONE CORRECT ANSWER.

The duration of the exam is **120 minutes.**

1. What is the expected value of a fair 6-sided die roll?

   A. 3
   B. 3.5
   C. 4
   D. 4.5
   E. 5

2. Let $X \sim N(\mu, \sigma^2)$. Which transformation yields a standard normal variable? *(Assume μ is the mean, σ is the standard deviation)*

   A. $X - \mu$
   B. $\frac{X}{\sigma}$
   C. $\frac{X-\mu}{\sigma}$
   D. $\frac{X+\mu}{\sigma}$
   E. $\frac{X-\mu}{\sigma^2}$

3. Two events A and B are independent if:

   A. $P(A \cup B) = P(A) + P(B)$
   B. $P(A \cap B) = P(A)P(B)$
   C. $P(A \mid B) = P(B \mid A)$
   D. $P(A \mid B) = 1$
   E. $P(A \cup B) = P(A)P(B)$

4. Let $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \cap B) = 0.1$. What is $P(A \cup B)$?

   A. 0.6
   B. 0.7
   C. 0.8
   D. 0.9
   E. 1.0

5. The Central Limit Theorem applies to:

   A. Only normal populations

   B. Small samples from skewed distributions

   C. The sum (or mean) of i.i.d. (independent and identically distributed) random variables as $n \to \infty$

   D. Variance of normal samples

   E. Correlated variables

6. In MLE (Maximum Likelihood Estimation), the estimator is found by:

   A. Minimizing the loss function

   B. Maximizing the posterior distribution

   C. Maximizing the log-likelihood function

   D. Solving the Bayes rule

   E. Taking the expected value of the likelihood

7. Let X be a discrete random variable with the following probability mass function (pmf):

$$P(X = -1) = 0.3, \quad P(X = 0) = 0.5, \quad P(X = 1) = 0.2.$$

   What are the expected value $E(X)$ and variance $\mathrm{Var}(X)$ of X?

   A. $E(X) = 0.33, \quad \mathrm{Var}(X) = 0.61$

   B. $E(X) = -0.1, \quad \mathrm{Var}(X) = 0.49$

   C. $E(X) = 0.33, \quad \mathrm{Var}(X) = 0.50$

   D. $E(X) = -0.1, \quad \mathrm{Var}(X) = 1.0$

   E. $E(X) = 0.5, \quad \mathrm{Var}(X) = 0.61$

8. In a two-sided hypothesis test with significance level $\alpha = 0.05$, the p-value is calculated to be 0.03. What should be the decision?

   A. Do not reject the null hypothesis

   B. Reject the null hypothesis

   C. Increase the sample size

   D. Reduce the significance level

   E. Repeat the experiment

9. Which of the following best describes the Law of Large Numbers in probability and statistics?

   A. The probability of an event changes as more trials are conducted.
   B. As the number of trials increases, the average of the results will get closer to the expected value.
   C. The outcome of each trial affects the outcome of the next trial.
   D. As the number of trials decreases, the average becomes more accurate.
   E. The sample average converges to the population mean with probability 1.

10. The uniform distribution $\mathcal{U}(0,1)$ has the probability density function (pdf):

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

    If $X \sim \mathcal{U}(0,1)$, what is $P(X < 0.2)$?

    A. 0.1
    B. 0.2
    C. 0.5
    D. 1
    E. Cannot be determined

11. If $A \in \mathbb{R}^{n \times n}$ is invertible, which of the following statements is not necessarily true?

    A. It is a full-rank matrix.
    B. All its rows are linearly independent.
    C. All its columns are linearly independent.
    D. All its eigenvalues are positive.
    E. Its determinant is nonzero.

12. For $A \in \mathbb{R}^{2 \times 2}$, we have $|A| = -4$ and $\text{trace}(A) = 0$. What is the larger eigenvalue?

    A. -2
    B. 0
    C. 2
    D. 4
    E. It cannot be determined based on the given information.

13. Which of the following matrix is always symmetric?

    A. The matrix $B = A + A^T$ where $A \in \mathbb{R}^{n \times n}$.

    B. The matrix $B = A - A^T$ where $A \in \mathbb{R}^{n \times n}$.

    C. Any invertible matrix.

    D. Any upper triangular matrix.

    E. Any lower triangular matrix.

14. Which of the following is not a vector space over $\mathbb{R}$?

    A. The set of all real-valued polynomials.

    B. The set of all real $n \times n$ matrices.

    C. The set of all solutions to a homogeneous linear system $A\mathbf{x} = \mathbf{0}$.

    D. The set of all positive real numbers.

    E. The set of all real-valued continuous functions.

15. Consider the following statements about basis $\mathbf{B}$ of a vector space $\mathbf{V}$:
    i-) $\mathbf{B}$ must contain infinitely many vectors.
    ii-) Every vector in $\mathbf{V}$ can be uniquely written as a linear combination of basis vectors.
    iii-) A basis must include the zero vector.
    iv-) Different bases of $\mathbf{V}$ can have different numbers of vectors.
    v-) $\mathbf{B}$ is unique.
    How many of the above statements are correct?

    A. 1

    B. 2

    C. 3

    D. 4

    E. 5

16. What is the rank of the given matrix?

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 0 & 1 & 1 \end{bmatrix}$$

A. 0
B. 1
C. 2
D. 3
E. 4

17. If a matrix A is negative definite, which of the following is true?

A. Zero is among the eigenvalues of A.
B. All eigenvalues of A are positive.
C. All eigenvalues of A are negative.
D. A must be a triangular matrix.
E. A must be a singular matrix.

18. Which of the following sets is convex?

A. The set $S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$
B. The set $S = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0\}$
C. The set $S = \{(x, y) \in \mathbb{R}^2 : x > 0 \text{ or } y > 0\}$
D. The set $S = \{(x, y) \in \mathbb{R}^2 : xy = 1\}$
E. The set $S = \{(x, 0) \in \mathbb{R}^2 : x \in \mathbb{R}\} \cup \{(0, y) \in \mathbb{R}^2 : y \in \mathbb{R}\}$

19. Which of the following statements about gradient descent and its variants is correct?

A. In batch gradient descent, the update at each step is based on a single randomly chosen data point.
B. Stochastic gradient descent (SGD) uses the full dataset to compute the gradient at every iteration.
C. Mini-batch gradient descent uses a subset of the dataset to compute an approximate gradient.
D. Adding momentum to gradient descent generally slows down convergence.
E. In gradient descent with momentum, the update depends only on the current gradient, not on previous steps.

20. Which of the following optimization problems is a convex optimization problem?

    A. Minimize $x^2 + y^2$ subject to $x^4 + y^4 \leq 1$

    B. Minimize $x^4 + y^4$ subject to $x^2 + y^2 \geq 1$

    C. Minimize $-x^2 - y^2$ subject to $x + y = 2$

    D. Minimize $\sin(x)$ subject to $0 \leq x \leq 1$

    E. Minimize xy subject to $x \geq 0, y \geq 0$

21. What is the output of the following code?

```
x = 5
y = "5"
print(x + int(y))
```

    A. 55

    B. 10

    C. TypeError

    D. "55"

    E. None of the above

22. Which of the following is a correct way to define a function in Python?

    A. `function myFunc():`

    B. `def myFunc:`

    C. `def myFunc():`

    D. `function = myFunc()`

    E. `func myFunc():`

23. Which data type is *immutable* in Python?

    A. list

    B. dictionary

    C. set

    D. tuple

    E. bytearray

24. What is the result of the following expression?

```
bool([]) or bool(0)
```

A. `True`
B. `False`
C. `Error`
D. `None`
E. `0`

25. What does this list comprehension return?

```
[i**2 for i in range(4) if i % 2 == 0]
```

A. `[1, 4, 9]`
B. `[0, 1, 4, 9]`
C. `[0, 4]`
D. `[0, 2]`
E. `[2, 4]`

26. Which statement is true about Python variable scope?

A. A variable defined inside a function definition is also accessible outside the function.
B. Variables in the global scope can override local variables.
C. The `nonlocal` keyword is used to access variables in the global scope.
D. A variable defined in an outer function definition can be used in an inner function definition with `nonlocal`.
E. Local variables are always immutable.

27. What will this code print?

```
def foo(val, lst=[]):
    lst.append(val)
    return lst

print(foo(1))
print(foo(2))
```

A. `[1]` and `[2]`

B. `[1]` and `[1, 2]`

C. `[1]` and `[1]`

D. `[1, 2]` and `[1, 2]`

E. `Error`

28. What will be the output of the following code using default arguments?

```
def add_to_list(val, lst=None):
    if lst is None:
        lst = []
    lst.append(val)
    return lst

print(add_to_list(1))
print(add_to_list(2))
```

A. `[1]` and `[2]`

B. `[1]` and `[1, 2]`

C. `[1, 2]` and `[1, 2]`

D. `[1]` and `[2, 1]`

E. `[1]` and `[1]`

29. What does this code do?

```
def make_counter():
    count = 0
    def counter():
        nonlocal count
        count += 1
        return count
    return counter

c1 = make_counter()
print(c1(), c1(), c1())
```

A. Prints `0 0 0`

B. Prints `1 2 3`

C. Prints `1 1 1`

D. `Error`

E. Prints `0 1 2`

30. What will be the output of this code involving unpacking and starred expressions?

```
a, *b, c = [1, 2, 3, 4, 5]
print(b)
```

A. `[1, 2, 3, 4, 5]`

B. `[2, 3, 4]`

C. `[1, 2, 3]`

D. `[2, 3, 4, 5]`

E. `[1, 2, 3, 4]`

31. The cell below computes the dot-product between two vectors of length one million using five different techniques. Which technique (**A**–**E**) will typically run in the shortest wall-clock time on a modern CPU?

```
import numpy as np, time, math
n = 1_000_000
u = np.random.rand(n)
v = np.random.rand(n)

# ----- five variants -----
# Technique A
A = np.dot(u, v)
# Technique B
B = (u * v).sum()
# Technique C
C = sum(float(u[i]*v[i]) for i in range(n))
# Technique D
D = math.fsum(u[i]*v[i] for i in range(n))
# Technique E
acc = 0.0
for a, b in zip(u, v):
    acc += a * b
E = acc;
```

A. Technique A

B. Technique B

C. Technique C

D. Technique D

E. Technique E

32. A DataFrame orders has the schema below (five rows shown for illustration):

```
   order_id customer  price  timestamp
0       101     C001  39.90 2024-06-01
1       102     C002  24.50 2024-06-01
2       103     C001  15.75 2024-06-02
3       104     C003  42.00 2024-06-02
4       105     C002  30.10 2024-06-03
```

Which one-liner returns a `Series` whose index is `customer` and whose values are the *mean* order price?

A. `orders.groupby('customer').mean()['price']`

B. `orders.groupby('customer')['price'].mean()`

C. `orders.groupby('price')['customer'].mean()`

D. `orders.pivot(index='customer', columns='order_id', values='price').mean()`

E. `orders.set_index('customer').mean('price')`

33. Consider this column with missing values:

```
import pandas as pd, numpy as np
s = pd.Series([2.0, np.nan, 5.0, 3.0, np.nan, 4.0])
```

Which line (A-E) fills the NaN slots with the median of the non-missing values and puts the result back into s ?

A. `s = s.fillna(s.mean())`

B. `s.fillna(s.median(), inplace=True)`

C. `s = s.interpolate('median')`

D. `s = s.fillna(np.median(s))`

E. `s = s.dropna().median()`

34. We are training a decision tree classifier on a dataset. The baseline tree (default parameters) achieves training accuracy = 1.00 and validation accuracy = 0.72—a clear over-fit. Five alternative initialisations are proposed:

```
from sklearn.tree import DecisionTreeClassifier
A = DecisionTreeClassifier(max_depth=None,  min_samples_leaf=1)
B = DecisionTreeClassifier(max_depth=25,   min_samples_leaf=1)
C = DecisionTreeClassifier(max_depth=6,    min_samples_leaf=5)
D = DecisionTreeClassifier(max_depth=3,    min_samples_leaf=25)
E = DecisionTreeClassifier(max_depth=None, min_samples_leaf=10)
```

Which model (A-E) is most likely to improve validation accuracy by introducing sensible, moderate regularisation while avoiding extreme under-fitting? *Hint: the default tree has* ***max_depth=None*** *and* ***min_samples_leaf=1****.*

A. Model A

B. Model B

C. Model C

D. Model D

E. Model E

35. On a credit-card-fraud validation set the classifier produced this confusion matrix:

	Pred Fraud	Pred Not Fraud
Actual Fraud	**73**	7
Actual Not Fraud	33	**2887**

The bank's priority is to catch as many fraudulent transactions as possible while keeping the follow-up workload (false positives) manageable.

Below are five code snippets from the from sklearn.metrics library that could be used to report performance. Which snippet (A-E) computes the metric that best matches the stated objective?

A. `score = accuracy_score(y_true, y_pred)`

B. `score = roc_auc_score(y_true, y_scores)`

C. `score = recall_score(y_true, y_pred)`

D. `score = precision_score(y_true, y_pred)`

E. `score = f1_score(y_true, y_pred)`

36. Fill in the single missing line (marked ???) so that gradients do **not** accidentally accumulate across mini-batches:

```
for xb, yb in dataloader:
    # ???
    preds = model(xb)
    loss  = criterion(preds, yb)
    loss.backward()
    optimizer.step()
```

A. `optimizer.zero_grad()`

B. `loss = loss.detach()`

C. `torch.cuda.empty_cache()`

D. `optimizer.step()`

E. `model.zero_grad()`

37. For a class that inherits `torch.utils.data.Dataset`, which set of methods must you implement so that the object works correctly with DataLoader?

A. __iter__

B. __len__

C. __len__ **and** __getitem__

D. __getitem__ **and** __iter__

E. None

38. We are using the PyTorch library for training a neural network and using Weights & Biases library for tracking experiment results. However our results show small accuracy differences each time the script is executed. Insert one line before model training to stabilise the results:

A. `np.random.seed(42)`

B. `torch.manual_seed(42)`

C. `random.shuffle(my_dataset)`

D. `torch.cuda.empty_cache()`

E. `wandb.config.seed = 42`

39.
```python
import pandas as pd

df = pd.DataFrame({
    'id':    [1, 2, 3, 4],
    'color': ['red', 'blue', 'red', 'green'],
    'value': [10, 15, 7, 12]
})

# ???
print(df)
```

Replace the comment ??? with one line (A-E) that produces a new DataFrame df where the color column is one-hot encoded into binary indicator columns prefixed with color, and the original color column is removed.

A. `df = pd.get_dummies(df, columns=['color'])`

B. `df = df.join(pd.get_dummies(df['color'], prefix='color'))`

C.
```python
df = pd.concat([df.drop(columns=['color']),
                pd.get_dummies(df['color'], prefix='color')], axis=1)
```

D. `df = df.drop(columns=['color']).get_dummies()`

E. `df = df.assign(pd.get_dummies(df['color']))`

40. Which line creates a 2×3 tensor of zeros *on the GPU* and *does not track gradients*? Choose exactly one.

```python
import torch
# select ONE of A to E ↓↓↓
```

A. `t = torch.zeros((2, 3), device='cuda')`

B. `t = torch.zeros((2, 3), device='cuda').requires_grad_()`

C. `t = torch.zeros((2, 3), device='cuda', requires_grad=False)`

D. `t = torch.zeros((2, 3), device='cuda').detach()`

E. `t = torch.zeros((2, 3)).to('cuda')`