F78DS 2024/2025 Practice Exam

Instructions

This is a practice exam. In the actual exams, you are to:

Answer all questions.

- *20 1-mark questions from Section A, and*
- *10 3-mark questions from Section B.*

In this practice exam:

Answer all questions.

- **10 1-mark questions from Section A, and**
- **5 3-mark questions from Section B.**

No due date

Available until 27 Apr at 23:59

Points 25

Questions 15

Time limit 60 Minutes

Allowed attempts Unlimited

SECTION A

Multiple choice questions: This section is worth 20 marks ***(In this practice exam, it is worth 10 marks)***. Each question is worth 1 mark. Identify the choice that **best** completes the statement or answers the question. There is only one best answer to each question. Sometimes two answers may appear feasible, but you are to pick the one you believe is the best.

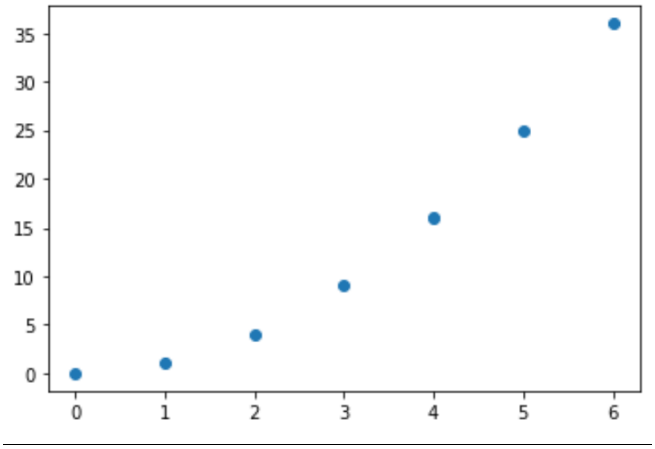
Marking Scheme for Multiple Choice Questions:

- 1 mark for a correct answer
- 0 marks for the wrong answer
- 0 marks for no answer

Q1 [1pt] What is the primary purpose of data serialisation?

- To encrypt data for security
- To convert data into a format suitable for storage or transmission
- To compress data for efficient storage
- To analyse data for patterns

Q2 [1pt] Which of the following will produce a plot that looks like below:



For reference, the code prior to the plot is

```
import pandas as pd

import matplotlib.pylab as plt

df = pd.DataFrame({ 'X' : [0,1,2,3,4,5,6], 'Y' : [0,1,4,9,16,25,36]})
```

- plt.plot(df['Y'])
- plt.hist(df['Y'])
- plt.boxplot(df['Y'])
- plt.scatter(df['X'], df['Y'])

Q3 [1pt] Which of these tasks might a data scientist typically perform?

- Funding the project
- Approving the project
- Pitching project ideas
- Acquiring tools for the project

Q4 [1pt] Which of the following choices is **NOT** a method to address missing data in a dataset?

- Mean imputation
- Deletion of rows with missing values
- Feature scaling
- Predictive imputation (using machine learning)

Q5 [1pt] Which data type best represents the weight of a person?

- Categorical-Nominal
- Categorical-Ordinal
- Numeric-Discrete
- Numeric-Continuous

Q6 [1pt] You have applied a clustering algorithm (e.g., *k*-means) to a dataset and you want to evaluate the quality of the resulting clusters (without having access to any pre-existing labels). Which of the following metrics would be most appropriate for this purpose?

F1-Score

Silhoutte Score

RMSE

Lift

Q7 [1pt] What is the primary purpose of a "Model Staleness Test" in a machine learning deployment?

- To evaluate the model's computational efficiency during training
- To determine if the model's predictions are still accurate and relevant over time
- To measure the model's ability to handle missing data
- To assess the model's robustness against adversarial attacks

Q8 [1pt] A grocery store manager wants to understand which items are frequently purchased together to optimise product placement. Which of the following techniques would be most suitable for the grocery store manager's goal?

- Classification
- Regression
- Clustering
- Associative Rule Mining

Q9 [1pt] A machine learning model is consistently underfitting the training data, exhibiting high error on both the training and testing sets. What does this indicate about the model's bias and variance?

- High bias, low variance
- Low bias, high variance
- High bias, high variance
- Low bias, low variance

Feedback

While high bias is present, high variance would cause the model to perform differently on different training sets. Since it is underfitting consistently, the variance is low.

Q10 [1pt] Which of the following best describes underfitting in a machine learning model?

- The model performs well on the training data but poorly on unseen data
- The model captures noise in the training data, leading to poor generalization
- The model fails to capture the underlying patterns in the training data, resulting in poor performance on both training and unseen data
- The model is perfectly aligned with the training data, leading to optimal performance

SECTION B

Short Answer Questions: This section is worth 30 marks ***(In this practice exam, it is worth 15 marks)*** and each question is worth 3 marks. Your answer should be written in clear, simple English and should be complete enough in addressing the question. Extensive prose is not required. Structured bullet points are acceptable.

Q11 [3pts] Give an example of implicit data that reveals personal information about a user. Describe the regular data that lies behind the implicit data and then describe the implicit data and why it is implicit.

Regular data: website browsing history (pages visited, search history, etc.).

Implicit data: Inferring pregnancy based on frequent visits to maternity/baby product pages (or if you have read the Target story, you can elaborate on that).

Why Implicit: The daughter (from the Target story) never said "I'm pregnant" but their actions suggest it

Q12 [3pts] Assuming that 'results.csv' is a CSV file that contains 3 columns, which are the students' identity, the course code and course's respective marks. Explain the code and explain what the output of the following Python code would be?

```
import pandas as pd
df = pd.read_csv('results.csv', sep=',')
df.groupby(['course'])['mark'].max()
```

1. import pandas as pd: This line imports the pandas library, which is a data manipulation and analysis tool in Python.
2. df = pd.read_csv('results.csv', sep=','): This line reads the CSV file named 'results.csv' into a DataFrame called df. The sep=',' argument specifies that the values in the CSV file are separated by commas.
3. df.groupby(['course'])['mark'].max(): This line performs a groupby operation on the DataFrame df. df.groupby(['course']): This groups the DataFrame by the 'course' column. This means that all rows with the same course code are grouped together. ['mark']: This selects the 'mark' column from the grouped data. '.max()': This calculates the maximum value in the 'mark' column for each group (i.e., for each course).

The output of this code will be an array where the index is the course code and the values are the highest marks obtained in each course.

Q13 [3pts] What is the difference between training and testing datasets? In what circumstances would you normalise the dataset?

- Training Dataset (usually 80% of the dataset): Used to train the model, containing features and labels for learning patterns.
- Testing Dataset (usually 20% of the dataset): Used to evaluate the model's performance on unseen data.
- Normalization: Scales features to a standard range (ex: 0 to 1) to improve convergence, enhance performance and ensure consistency. We need to normalize when features have different scales or units, especially for algorithms sensitive to feature magnitude.

Q14 [3pts] Briefly explain how the Decision Tree algorithm works.

The Decision Tree algorithm splits data into subsets based on the most informative features, creating a hierarchical structure of decision nodes and leaf nodes. It recursively selects the best feature to split the data until a stopping criterion is met, making it interpretable and easy to understand for both classification and regression tasks.

Q15 [3pts] Briefly explain the differences between regression and clustering, giving an example case. Which one can be used to predict a salary based on the age and job title of a person?

- Regression is a supervised learning technique used to predict a continuous output variable based on input features. For example, predicting house prices based on size and location.
- Clustering is an unsupervised learning technique used to group similar data points together without predefined labels. For example, segmenting customers based on purchasing behaviour.
- To predict a salary based on age and job title, regression would be used, as it involves predicting a continuous variable (salary) from given input features (age and job title).