WRITTEN EXAMINATION

Course: Intro to Data Science A1N

Examination

Course code: IT803A

Credits for written examination: 7.5

Date: 2025-03-14

Examination time: 08:15 – 12:30

Examination responsible: Addi Ait-Mlouk

Teachers concerned

Aid at the exam/appendices

Other

Instructions

☐ Take a new sheet of paper for each teacher.

☐ Take a new sheet of paper when starting a new question.

☐ Write only on one side of the paper.

☒ Write your name and personal ID No. on all pages you hand in.

☒ Use page numbering.

☒ Don't use a red pen.

☒ Mark answered questions with a cross on the cover sheet.

Grade points: Each question is graded 0–10 points. To pass the exam, you need a minimum of 5 points on each question (more details on the next page).

Examination results should be made public within 18 working days

Good luck!

Total number of pages

Questions

- The exam has five questions, one for each course objective.
- Each question has sub-questions (1, 2, 3, ...)
- Each question is graded with up to 10 points.
- To pass a question, you need to have at least 5 points on the question.
- To pass the exam, you need to pass all the questions.
- The maximum number of points on the exam is 50.

Grading

If your score on any question is below 5 points, your grade will be U (Fail). If you have at least 5 points on each question, your grade is determined using the sum of points as follows:

Points	Grade	Percentage
45-50	A	90-100
40-44	B	80-89
35-39	C	70-79
30-34	D	60-69
25-29	E	50-59
0-24	F	0-49

A (Excellent), B (Very good), C (Good), D (Satisfactory), E (Sufficient) or F (Fail)

Don't forget to motivate all your answers!

Good luck!

Question 1

[Course objective: extensively describe and problematize the state of the art of the field of Data Science, and discuss fundamental application areas for Data Science]

1. Which step is often the most time-consuming in the data science project lifecycle?
2. In the context of preprocessing a dataset, data can be referred to be in a *wide format* or in a *long format*, respectively. Explain the concepts of wide and long format together with examples of when to use which format.
3. In the context of data preprocessing, what are *missing values* and why might they be a problem for data science applications?
4. Describe three different approaches to handle missing values (see also Q 3.5).
5. Modeling using `scikit-learn` follows a generic pattern (or template) that is independent of the model being used. Describe the major steps in this pattern.

Question 2

[Course objective: extensively exemplify and contrast different perspectives on central foundations, principles, methods and theories within the field]

1. In the context of model evaluation and critique, thoroughly explain the concepts of *true positives*, *true negatives*, *false positives* and *false negatives*.
2. What is a *density plot*?
3. In the context of graphical perception, order the following visual attributes from which allows for *more accurate judgements* to which allows for *more generic judgements*: Area, length, color shading (lightness) and direction.
4. In the context of data and information visualization, thoroughly explain the concept of *small multiples,* together with a concrete example.
5. The performance of clustering models/algorithms depends on choosing the right *similarity measure* (distance function). Describe three different similarity measures together with concrete examples for which they are suitable.

Question 3

[Course objective: account for syntax and semantics for programming languages that are particularly suited for Data Science]

1. In Python, what would be the output of running

   ```
   print(False == 0, [] == False, () == False, None == True, “” == False)
   ```

   and

   ```
   print(bool({}), bool(“”), bool(-5), bool(“False”), bool(None))
   ```

   respectively?
2. Explain why the *functional programming* paradigm is of interest to data science?
3. Describe how Python supports functional programming.
4. Define a Python function that given two lists $[x_1, x_2, ..., x_n]$ and $[y_1, y_2, ..., y_m]$ uses *list comprehension* to return a list of all combinations of pairs (x_i, y_j).
5. Provide Python code for three different approaches to handle missing values (see Q 1.4).

Question 4

[Course objective: present and discuss ethical and societal issues in connection to Data Science and its applications]

1. What rights does GDPR give individuals, and how do these affect data collection?
2. How do quasi-identifiers threaten privacy in data analysis?
3. What is k-anonymity, and how does it protect privacy using quasi-identifiers?
4. How can bias in language models cause unfair outcomes and stereotypes?
5. How do collaborative and federated learning protect privacy while training models?

Question 5

[Course objective: critically reflect on and describe requirements and issues for programming within the area; and independently develop computer programs within the area]

1. What are the key steps in building a model service using REST APIs, and how can Flask-RESTful or FastAPI help?
2. What programming skills are needed for linear regression, and how can scikit-learn help?
3. What architecture should you consider when building neural networks for regression, and how do you choose activation and loss functions?
4. How can Docker manage dependencies and ensure reproducibility, and what are the benefits?
5. Critically examine the challenges associated with deploying machine learning models on edge devices.