UNIVERSITY
OF SKÖVDE

School of informatics

WRITTEN EXAMINATION

Course: Data mining A1N

Examination

Course code: IT734A

Credits for written examination: 4.5

Date: 2025-01-10

Examination time: 14:15 - 17:30

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 (a, b, c, ...)
- 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 have passed all 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: critically reflect and describe utility, problems and limitations of data mining]

a. List and explain the main stages of the data mining lifecycle
b. In what ways can missing data affect the performance of data mining algorithms
c. Identify three industries where data mining has significantly impacted operations. Provide examples to illustrate its applications
d. How do overfitting and underfitting impact data mining models, and what strategies can be employed to address them?
e. Describe the challenges associated with mining high-dimensional datasets and provide potential solutions.

Question 2

[Course objective: critically reflect and describe data mining algorithms within the classification, association analysis and cluster analysis, with respect to application and structure]

a. Explain the role of clustering in data mining. How does it differ from classification?
b. Describe the Apriori algorithm for association rule mining. How does its structure affect its scalability in large datasets?
c. Discuss the role of support and confidence thresholds in association rule mining. How do they affect the discovery of meaningful patterns?
d. How does k-means clustering work, and what are its limitations when applied to datasets with varying densities or shapes? Suggest possible alternatives.
e. How can clustering techniques be combined with classification models to enhance the accuracy of predictions? Provide a practical example

Question 3

[Course objective: implement and explain basic data mining algorithms]

a. How does the curse of dimensionality affect classification and clustering algorithms, and what techniques can mitigate its impact?
b. Explain the difference between supervised and unsupervised learning in the context of classification and clustering. How do their evaluation metrics differ?

c. What role does feature selection play in improving the performance of classification algorithms?
d. Elaborate on the concept of Stochastic Gradient Descent (SGD). What role does it play in optimizing machine learning models,
e. Provide an overview of the workings of Convolutional Neural Networks (CNNs). Support your explanation with a relevant example, highlighting the key components and their roles in image recognition.

Question 4

[Course objective: identify and describe problems where data mining is relevant]

- Given the five following data mining problems, classify them as classification, regression or clustering problems. Motivate your answer.

a. Estimating the number of daily visitors to a theme park based on weather, ticket prices, and holidays.
b. Categorizing customer support tickets into predefined topics like billing, technical issues, or general inquiries.
c. Identifying fraudulent transactions in a dataset of credit card payments.
d. Dividing an online retail store's customers into distinct purchasing patterns without predefined categories.
e. Predicting the monthly electricity consumption of a household based on historical data and external factors like temperature.

Question 5

[Course objective: select suitable data mining algorithms for solving such problems and analyze, compare and evaluate results]

a. What are the common preprocessing steps in text mining?
b. When is it best to use a decision tree algorithm over a neural network?
c. What evaluation metrics would you use to assess the accuracy of a classification?
d. How do you interpret the results of a clustering algorithm?
e. What is the difference between Bag of Words (BOW) and Word Embedding, motivate your answer by providing examples.