UNIVERSITY
OF SKÖVDE

School of informatics

WRITTEN EXAMINATION

Course: Data mining A1N

Examination

Course code: IT734A

Credits for written examination: 4.5

Date: 2023-01-10

Examination time: 08:15 - 11:30

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. What challenges are commonly encountered in the field of data mining?
b. Outline the six key steps involved in executing the data mining process
c. How do data quality issues impact the effectiveness of data mining algorithms?
d. Provide an explanation of text mining and distinguish it from text analytics

e. In your opinion, what are some practical applications where data mining has proven to be highly beneficial?

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. Describe a specific data mining classification algorithm, highlighting its structural components and real-world applications.
b. Explore the role of ensemble methods in improving the performance of classification algorithms
c. In association analysis, what are support count, support, frequent itemset and confidence?
d. What is the difference between linear regression and logistic regression?
e. Discuss the structural differences between decision trees and support vector machines in classification algorithms

Question 3

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

a. Explain the mechanisms behind K-means and K-means+ clustering algorithms. How do they differ?
b. How does Apriori algorithm work in association rules analysis
c. Define what a recommender system is and distinguish between collaborative filtering and content-based approaches. How do these methods leverage user preferences to provide personalized recommendations?
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. Predicting the sales price of houses based on features (e.g., size, location)
  b. Grouping news articles without predefined categories
  c. Identifying Spam emails in a dataset of email communications
  d. Segmenting internet users into behavior-based groups
  e. Predicting student performance based on academic history and demographics

Question 5

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

a. How can the choice of a classification algorithm impact the accuracy of predicting customer churn in a telecommunications dataset?
b. Select an appropriate regression algorithm for predicting housing prices based on real estate features. How does the model's interpretability and predictive accuracy vary across different regression techniques?
c. Examine the impact of feature selection on the performance of clustering algorithms in high-dimensional biological data. How does the choice of feature selection method influence the quality of clustering results?
d. Assess the ethical implications of using data mining algorithms for decision-making in sensitive domains such as criminal justice. How can bias and fairness be addressed when selecting and deploying these algorithms?
e. What is the difference between Bag of Words (BOW) and Word Embedding, motivate your answer by providing examples.