UNIVERSITY OF OULU

YLEISEN TENTIN TENTTILOMAKE - GENERAL EXAM FORM

Opiskelija täyttää / Student fills in

Opiskelijan nimi / Student name:	**Opiskelijanumero / Student number:**

Opettaja täyttää / Lecturer fills in

Opintojakson koodi / The code of the course: 521156S	
Opintojakson (tentin) nimi / The name of the course or exam: Towards Data Mining	
Opintopistemäärä / Credit units: 5 op Mikäli kyseessä on välikoe, opintopistemääräksi täytetään 0 op. 0 ECTS Credits is used for mid-term exams.	
Tiedekunta / Faculty: Tieto- ja sähkötekniikan tiedekunta / Faculty of Information Technology and Electrical Engineering	
Tentin pvm / Date of exam: 2019-11-07	**Tentin kesto tunteina / Exam in hours:** 3 h
Tentaattori(t) / Examiner(s): Satu Tamminen	**Sisäinen postiosoite / Internal address:** 9BISG
Tentissä sallitut apuvälineet / The devices allowed in the exam:	
Funktiolaskin / Scientific calculator, Ohjelmoitava laskin / Programmable calculator	
Muut tenttiä koskevat ohjeet opiskelijalle (esimerkiksi kuinka moneen kysymyksen opiskelijan tulee vastata) / Other instructions for students e.g. how many questions he/she should answer: Choose 4 out of the 5 questions to answer. Each question is worth 6 points. If you answer all the questions, only the first 4 of your answers will be graded. Voit vastata myös suomeksi.	

UNIVERSITY OF OULU

YLIOPISTOTENTTI - UNIVERSITY EXAM

521156S Towards Data Mining

Choose 4 out of the 5 questions to answer. Each question is worth 6 points. If you answer all the questions, only the first 4 of your answers will be graded. Voit vastata myös suomeksi.

1. Explain terms briefly.
    a. Artificial data (1p)
    b. Randomized controlled trial (RCT) (1p)
    c. Data management (1p)
    d. Validation set (1p)
    e. Data normalization (1p)
    f. Predictive model (1p)

2. Relational databases
    a. Explain the concept of data modeling and give two examples of reasons why it is useful to create a data model when designing a database. (2p)
    b. Examine the database below, representing the people and animals of a livestock farm, and then the following queries on the database. Explain the **intent** of each query (e.g "return the age of the oldest caretaker") and write down what it **returns** (e.g. "79"). (4p)

i. `SELECT firstName, lastName FROM Caretaker ORDER BY age DESC;`
ii. `SELECT * FROM Animal WHERE pricePerHead BETWEEN 300 AND 600;`
iii. `SELECT animalKind FROM Animal A INNER JOIN Caretaker C ON A.caretakerId = C.caretakerId WHERE firstName = 'Betty';`
iv. `SELECT MIN(pricePerHead) FROM Animal WHERE headCount <= 50;`

Old McDonald's Farm

Caretaker

1	Amos	McDonald	79
2	Betty	McDonald	76

Animal

1	1	cow	40	1000
2	1	pig	60	400
3	2	sheep	30	550
4	2	chicken	500	10

3. Choose three missing value imputation methods from the following list: mean imputation, regression imputation, stochastic regression imputation, multiple imputation (2p each = 6p total)
    a. Explain how each of the selected methods generates the imputed values.
    b. Explain the pros and cons of each of the selected methods.

4. Data collection
    a. Explain how or why data collection can go wrong. (3p)
    b. You have collected data from 300 persons: 25 of them have cancer and 275 do not have cancer. Therefore, your data set is imbalanced. Explain one sampling method that can be used to balance the data set. (3p)

5. Data normalization
    a. Explain why data normalization is needed? (1p)
    b. Describe the procedure of these two statistical normalization methods. (1p)

$$v' = \frac{v - v_{min}}{v_{max} - v_{min}} \qquad v' = \frac{v - \mathbf{mean}(v)}{\mathbf{sd}(v)}$$

    c. What can you tell about the statistical properties of the data after normalization with these methods (mean, deviation, extreme values, and distribution shape)? (4p)