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. Open data (1p)
    b. Contextual outlier(1p)
    c. Data model (1p)
    d. Accuracy paradox (1p)
    e. Variable construction (1p)
    f. Cross-validation (1p)

2. Missing data
    a. Explain the terms list wise deletion (also called complete case analysis) and pair wise deletion (also called available case analysis). (2p)
    b. Using data from Table 1, list the rows used for the following analyses using both list wise and pair wise deletion. You do not need to calculate the analysis results. (1p)
        i. Average height of female subjects
        ii. Average weight of all subjects
    c. Explain the idea of multiple imputation (MI) briefly. What is the main advantage of MI over other imputation methods? (3p)

Table 1. Data set with missing values.

Row number	Height	Weight	Gender
1	175 cm	-	Male
2	158 cm	60 kg	Female
3	169 cm	83 kg	-
4	-	65 kg	Female
5	172 cm	-	-
6	160 cm	63 kg	Male
7	-	73 kg	Male
8	161 cm	-	Female

3. Ethics and openness in data mining
    a. Explain the concepts of anonymity and consent and their relevance to the ethics of data mining. (3p)
    b. Explain the concept of open data. What kind of constraints does the definition of openness allow? What are the main benefits of open data from the perspective of an individual data miner and from that of society as a whole? (3p)

4. Data pre-processing
    a. Explain the steps of knowledge discovery process. (2p)
    b. What are the types of data reduction? Describe with examples. Why is data reduction needed? (4p)

5. Data collection
    c. Describe the ethical aspects and risks needed to take into account when collecting data from humans. (3p)
    d. 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)