TU Delft

CYBER DATA ANALYTICS (CS4035)

Exam, July 1 2019
9:00 - 12:00

Important: This exam consists of 5 questions split into subquestions. For each subquestion you can get a maximum of 5 points. Always give full explanations of your answers and number all steps of the asked algorithms. Do not forget to put your name and student number on every sheet of paper. Answers are required to be in English.

Question 1 - class imbalance

SMOTE is a commonly used method to use machine learning algorithms when the input data are imbalanced.

(a) Give pseudocode for the SMOTE procedure. (5 pt)

(b) Provide a workflow (think of KNIME, labeled boxes and arrows) of the different steps for a cross-validation scheme that includes SMOTE. Briefly explain the position of SMOTE in the workflow. (5 pt)

(c) Several researchers advice to, after SMOTE-ing the data, perform additional transformations such as removing Tomek links. Think of one positive and one negative effect of removing such links. Briefly explain these effects. (5 pt)

Question 2 - time series

(a) Time series are often normalized before applying machine learning. Normalization can be applied (1) on the whole sequence, or (2) on sliding windows. Draw two time series showing similar behavior under option (1) but not under (2). Draw two time series showing similar behavior under option (2) but not under (1). (5 pt)

(b) Dynamic Time Warping is a technique for computing a distance between time series. Give two characteristics of software and malware data that highlight the usefulness of Dynamic Time Warping. What is an important shortcoming of using Dynamic Time Warping when analyzing such data? (5 pt)

(c) Given the following time series:

- 1, 5, 8, 4, 8, 2, 10
- 1, 8, 4, 4, 10, 4

Compute the time-warping distance, and show how to compute it. (10 pt)

Question 3 - hashing and streaming

The FREQUENT algorithm (Misra-Gries) and the count-min sketch datastructure provide ways to estimate the frequency of identifiers such as IP-addresses. Suppose we observe the following sequence of identifiers:

$$1, 1, 2, 2, 2, 2, 2, 3, 3, 1, 1, 3, 3, 4, 1, 3, 3$$

(a) Wat guarantee does the FREQUENT algorithm provide with respect to these estimates? (5 pt)

(b) Wat guarantee does the count-min sketch provide with respect to these estimates? (5 pt)

(c) Give the memory content after every iteration of the FREQUENT algorithm with memory size 2 (k-1 in the lecture slides) when providing the above sequence as input. (5 pt)

(d) Give the content content of a count-min sketch after providing the same sequence using the following two hash functions: (5 pt)

$$h_1(n) = (4n + 2) \mod 3$$
$$h_2(n) = (2n - 1) \mod 3$$

Question 4 - sequential data mining

You are given the following discrete sequence data, collected by observing a network of different hosts:

host 1 B,A,A,A,A,B,A,A,B,B
host 2 A,B,A,B,A,B,A,B,A,B
host 3 B,B,B,A,A,B,B,B,B,B
host 4 A,B,A,B,A,A,B,B,B,B

(a) Provide the 3-gram probabilities for each of these hosts (do not apply smoothing). (5 pt)

(b) The sequence A,A,B,B,A,A is part of a test set generated by one of these hosts. Which host would you estimate has generate the sequence when profiling using fingerprints? Which host when profiling using probabilistic profiles? Explain why. Both use the 3-grams you obtained under (a). (10 pt)

(c) When the 3-gram probabilities obtained from two hosts are identical, does this impy the data sequences are identical? Show why (not). (5pt)

Question 5 - privacy & adversarial

(a) Classifier hardening is the process of updating a classification model by retraining using adversarial examples. Brielfy describe how adversarial examples are generated. Give one advantage and one disadvantage of hardening a classifier. (5pt)

Suppose you are asked to publish the following data table, in which Gender, Age, and Zipcode are assumed to be quasi-identifiers.

Name	Gender	Age	Zipcode	Salary
Alex	Male	35	27101	$54,000
Bob	Male	38	27120	$55,000
Cedric	Male	38	27130	$56,000
Diana	Female	38	27229	$65,000
Ellen	Female	43	27269	$75,000
Frederic	Male	47	27243	$70,000
Gennaro	Male	52	27656	$80,000
Hellan	Female	52	27686	$75,000
Indy	Male	58	27635	$85,000

(b) Describe record linkage and explain it could be used to discover the salary of Cedric if this data were published. (5pt)

(c) Make the data 2-anonymous using generalization and suppression, show the resulting table. (5pt)