NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Data Science and Machine Learning Master's Programme

Εξόρυξη Γνώσης από Δεδομένα (Data-Driven Knowledge Extraction)

Fall 2023

Instructors: D. Tsoumakos, G. Alexandridis

Assignment #2

Question 1 – Data Preprocessing:

Suppose that we are given a dataset for analysis, which includes an integer attribute, whose values are as follows:

87, 96, 80, 23, 36, 27, 64, 55, 75, 62, 93, 84, 65, 60, 78, 13, 61, 18, 99, 29, 14, 41, 32, 46, 44, 73, A, B

where A, B are the last two and the last-but-two digits of your student ID number, respectively (i.e. if your student ID ends in 1234, then A=12 and B=34). If the last two and/or the last-but-two digits of your student ID already appear in the values then use A=91 and/or B=26 (use the same assignments if the last two and the last-but-two of your student ID are the same)

Use smoothing by (a) bin means (10%) and (b) bin boundaries (10%) to smooth the above data, using a bin depth of 4. Illustrate your steps.

Question 2 – Mining Data Streams:

Suppose we are monitoring a data stream of integers and we want to monitor the values in the stream using a bloom filter of 20 bits and the following two hash functions:

1. $h_1(x) = (Ax + 11) \, mod \, 20$
2. $h_2(x) = (Bx + 2) \, mod \, 20$

where A, B are the last and the last-but-one digits of your student ID, respectively (ie if your student ID ends in 12, then A=1 and B=2). If the last and/or last-but-one digit is zero then use A=5 and B=7, respectively. Finally, if the last two digits of your ID are the same then user either A=5 or B=7.

At timestamp t, the bloom filter state is as follows:

$$[1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0]$$

Answer the following questions, justifying your answers:

1. Has the element $y = 8$ passed through the stream? (5%)
2. How certain are you about your answer? Give a probabilistic estimate. (15%)
3. What will be the bloom filter state after $x = 13$ passes through the stream at timestamp $t + 1$? (10%)
4. Can you give an estimate on the number of elements passed through the stream until $t + 2$? (10%)

Question 3 – Mining Discrete Sequences:

Suppose we are given a Hidden Markov Model with a set of states $S = \{s_1, s_1\}$, a set of symbols $\Sigma = \{a, b, c\}$, initial state probability $\Pi = \{0,1\}$, state transition probability $P = \begin{pmatrix} 0.x & 1 - 0.x \\ 0.x & 1 - 0.x \end{pmatrix}$ and probability of creating symbol σ_i at state s_j $\theta^j(s_i) = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$, where x is the last digit of your student ID number (or $x = 3$ if the last digit is zero). What is the probability of the aforementioned model producing the sequence $V = \{cba\}$? Justify your answer (40%)

Deliverable:

- This is an individual assignment.
- Your solutions must be uploaded to the helios class page in pdf format by the deadline. No late submissions, different file formats or scanned/photographed answer sheets will be accepted.