University of Wrocław: Data Science June, 2024

Math and coding problems

NOTE:

- For Task 1 provide numerical answers in a report and attach your programs you used (you may use any programming language).
- For Task 2 provide answers/discussion in a report (it may contain some code or mathematical formulas)
- For Tasks 3 and 4 provide answers in a report file (numerical answers / formulas and solutions). These can be scans of your hand-written solutions.
- For Task 5 provide (up to) three text files containing the solution.

Task 1. Write a program that prints the largest N-digit prime number n (in the decimal system) whose binary representation is $1^a0^b1^c$, where a, b, c are positive. So the binary representation of n is a sequence of ones , followed by a sequence of zeros, followed by a sequence of ones again.

Present the answers for $N \in \{6, 8, 10\}$. Your program should finish its work in several seconds.

Task 2. Our dataset consists of user's sessions in a e-commerce store. Let $I = I_1, I_2, \ldots, I_N$ be a set of N available products. A session is a sequence $(s_1, s_2, \ldots, s_n)$ of products browsed successively by the user, where n is the length of the session and $s_1, s_2, ..., s_n \in I$ are the browsed products. In the training dataset, for each session $(s_1, s_2, \ldots, s_n)$ we have a target product t being the next product browsed by the user directly after the session. A session-based recommender system aims at predicting the target product on the basis of the session. It returns a probability vector $(p_1, p_2, \ldots, p_N)$ where p_i denotes the probability that the target vector is I_i, for $i = 1, 2, \ldots, N$. In practical applications, such a probability vector enables to determine the K most probable products and display them to the user.

Example:

- Set of available products: I = {101, 102, 103, 104, 105}.
- Sessions and target products:
  (101, 102, 103), 104
  (101, 102, 103), 105
  (103, 102, 101), 102
  (105, 102, 104), 105
  (103, 102, 103), 101

- Recommendations:

  (0.10, 0.20, 0.20, 0.20, 0.30) → recommendation: 105
  (0.30, 0.10, 0.10, 0.40, 0.10) → recommendation: 104
  (0.20, 0.20, 0.20, 0.20, 0.20) → recommendation: 101 (or any other product, because of equal probabilities)
  (0.10, 0.40, 0.30, 0.10, 0.10) → recommendation: 102
  (0.00, 0.50, 0.20, 0.00, 0.30) → recommendation: 102

How to evaluate such a recommender system? Please propose and discuss possible evaluation metrics. Can we use accuracy for evaluating such a recommender system?

Task 3. Random variables X_1, X_2, X_3 are independent with exponential distrubtion with expectations $\mathbb{E}X_i = \dfrac{1}{(i+1)^2}, i = 1, 2, 3.$. Compute $P(X_2 = \min(X_1, X_2, X_3))$.

Task 4. Let $X_1, \ldots, X_n$ be independent identically distributed random variable coming from the population with $N(0, \theta)$ distribution, where $\theta = EX_1^2$. Find the maximum likelihood estimator of the parameter θ. Is the obtained estimate the minimum variance unbiased estimator? Justify the answer.

Task 5. In this task we consider a variant of the Sudoku puzzle (for the original puzzle see: Sudoku).
Our variant will be played on a bigger board (16×16, divided into 4×4 squares), and we will use haxadecimal digits (i.e. 0-9, A-F). Moreover, we define the score for every solved board to be the number of English words occuring horizontally on the board. We accept only words of length 3 and longer. We will count every occurence separately. Moreover if a word is a substring of another, then both will be counted (like ACE and ACED).
Task definition is a 16-line string where some digits are shown, and there is a placeholder (`'.'`) fot the rest. Here is the example task:

```
.........7E5AC49
A9C....7G4....32
............5BED
.83......2BD...G
9C............2A
8..............B
B5....CAE....D8F
........F5......
......D...6.G..8
3........D1...B5
..81....C...E.F7
CA......5B2....6
....F1E....2...C
.....A...8.F....
E1...78B6...FAD3
........BED.....
```

One possible solution of this task is:

```
FGD63B1287E5AC49
A9CBED57G4F18632
1472AFG83C965BED
583E964CA2BD7F1G
9CG453FED18B672A
8EFD7G9126A435CB
B56342CAEG791D8F
271AD8B6F5C39EG4
4B572CD39F6EG1A8
369G8EAF4D17C2B5
D281B965CA3GE4F7
CAEF147G5B28D396
GDA9F1E47352B86C
73BC6A2D18GF495E
E125G78B694CFAD3
6F48C539BEDA2G71
```

It is worth 5 points, since we have three occurences of the word `BED`, one occurence of the word `FED`, and one occurence of the word `FAD`.
Write a program which reads a single task description from the file `input.txt`, and outputs the solution followed by the info about the score into the file `outputs.txt`.
Use the attached `english_words.txt` file. Put the results of your program for files `sudoku1.txt`, `sudoku1.txt`, `sudoku1.txt` to your report. Note that there can sometimes be more than one correct answer. In this case, the results with higher scores are preferred.