University of Wrocław: Data Science June, 2024

Math and coding problems

Solutions

Applicant Name: Mohammad Faisal Ahmed

Country: Bangladesh

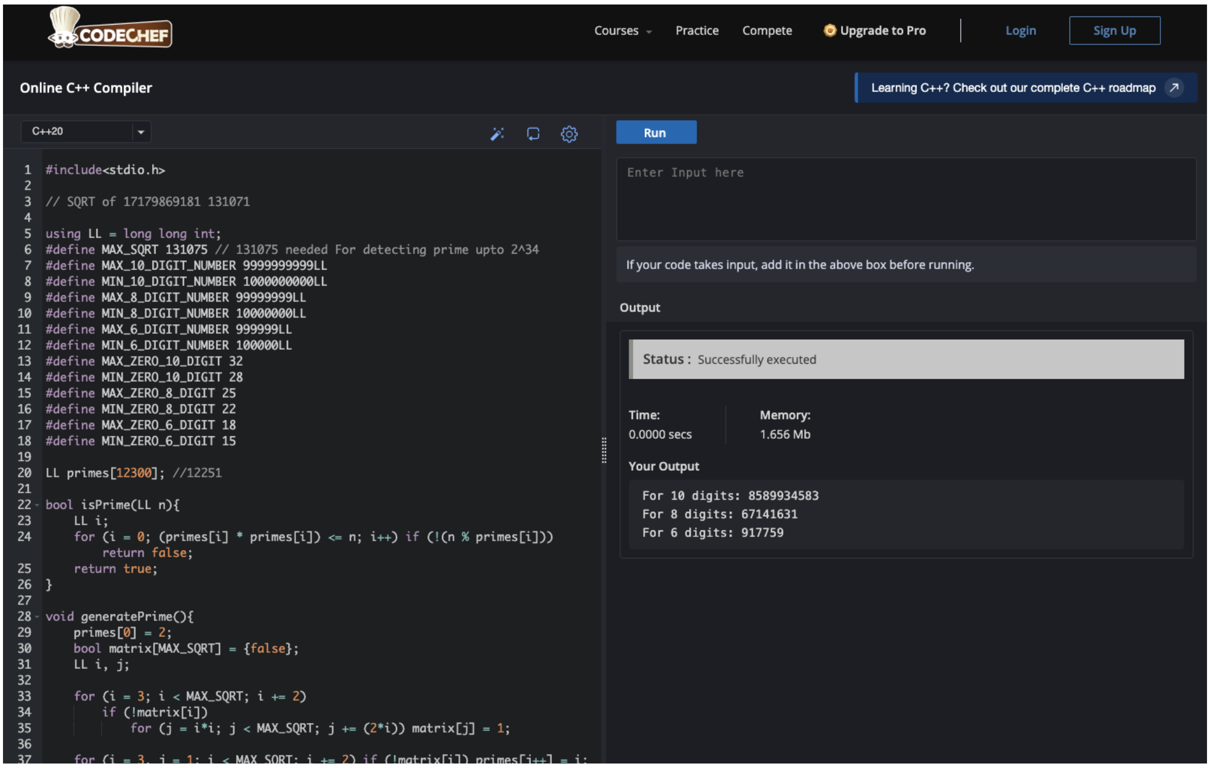
Applied Program: MS in Data Science

Submitted On: 5th of July, 2024.

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 ∈ {6, 8, 10}. Your program should finish its work in several seconds.

Solution: I have written a program in C (along with some C++ STL) that runs in 0.000 seconds in a popular online compiler named CODECHEF. Please refer to the screenshot:



Online C++ Compiler

Learning C++? Check out our complete C++ roadmap

```
#include<stdio.h>

// SQRT of 17179869181 131071

using LL = long long int;
#define MAX_SQRT 131075 // 131075 needed For detecting prime upto 2^34
#define MAX_10_DIGIT_NUMBER 9999999999LL
#define MIN_10_DIGIT_NUMBER 1000000000LL
#define MAX_8_DIGIT_NUMBER 99999999LL
#define MIN_8_DIGIT_NUMBER 10000000LL
#define MAX_6_DIGIT_NUMBER 999999LL
#define MIN_6_DIGIT_NUMBER 100000LL
#define MAX_ZERO_10_DIGIT 32
#define MIN_ZERO_10_DIGIT 28
#define MAX_ZERO_8_DIGIT 25
#define MIN_ZERO_8_DIGIT 22
#define MAX_ZERO_6_DIGIT 18
#define MIN_ZERO_6_DIGIT 15

LL primes[12300]; //12251

bool isPrime(LL n){
    LL i;
    for (i = 0; (primes[i] * primes[i]) <= n; i++) if (!(n % primes[i]))
        return false;
    return true;
}

void generatePrime(){
    primes[0] = 2;
    bool matrix[MAX_SQRT] = {false};
    LL i, j;

    for (i = 3; i < MAX_SQRT; i += 2)
        if (!matrix[i])
            for (j = i*i; j < MAX_SQRT; j += (2*i)) matrix[j] = 1;

    for (i = 3, j = 1; i < MAX_SQRT; i += 2) if (!matrix[i]) primes[j++] = i;
```

Enter Input here

If your code takes input, add it in the above box before running.

Status : Successfully executed

Time: 0.0000 secs

Memory: 1.656 Mb

Your Output

```
For 10 digits: 8589934583
For 8 digits: 67141631
For 6 digits: 917759
```

The answer

For 10 digits: 8589934583

For 8 digits: 67141631

For 6 digits: 917759

***For coding script please check the attachment.**

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 $s_1, s_2, \ldots, s_n \in I$ 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 prob- ability that the target vector is I_i, for i = 1,2,...,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, be- cause 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?

Solution:

Evaluating a recommender system is highly important to measure its performance and identify areas for improvement. Predicting a product a user may choose, we need different types of metrics that assess the accuracy of the prediction. My proposal for the possible evaluation metrics would be customize and combine some of the following tactics:

- **Top-K Accuracy:** Measure the percentage of times the recommended products (Top- K) are correct using the provided data. For example, if the system recommends the top 3 products and 2 out of 3 are correct then the Top-3 accuracy would be 66.67%.
- **Hit Rate:** We can calculate the percentage of sessions where at least one of the recommended products matches the target product (e.g.: either clicked or purchased by the user). This metric is useful when there are multiple relevant products.

- **Monitor and analyze user feedback:** We may even ask the user a reward-based survey at the end of each session. Using this metric we can easily calculate the accuracy.
- **Mean Reciprocal Rank (MRR)**: We can measure the average reciprocal of the rank of the correct product in the recommended list. For example, if the correct product is ranked 2nd out of 5, the MRR would be 1/2 = 0.5.
- **Recall at K:** Calculate the proportion of all relevant products that are among the top K recommended products.
- **Consider the type of user interactions:** If the users tend to interact with products in a specific way (e.g., buying, adding to cart, or simply browsing), we may want to use metrics that account for these interactions. For example, if users often add products to their cart, we could use metrics like "Cart Hit Rate" or "Purchase Conversion Rate".

A balanced evaluation approach should include a combination of multiple metrics to provide a comprehensive understanding of the recommender system's performance, ensuring it effectively meets user needs by recommending relevant and engaging products based on browsing sessions.

Can we use accuracy for evaluating such a recommender system?

Accuracy may not be an ideal metric for evaluating a recommender system for several reasons:

- **Accuracy is sensitive:** If there are many more sessions with a particular product being browsed frequently, accuracy might be skewed towards that product. For example: Umbrella might be browsed in the rainy season along with raincoat. But user might also use it in the scorching heat of sunlight without the raincoat.
- **Accuracy doesn't consider ranking:** A recommender system can have high accuracy but provide poor rankings. For example: Different types of chairs might be suggested with a table, but a particular chair may be at the pick of the choice in a particular area such as home or office. Therefore, their rankings might not be accurate despite of having high accuracy.
- **Accuracy doesn't account for diversity:** A system might always recommend the same product, even if it's not very accurate. For example: Combination of Tea, Sugar, Coffee and Milk.

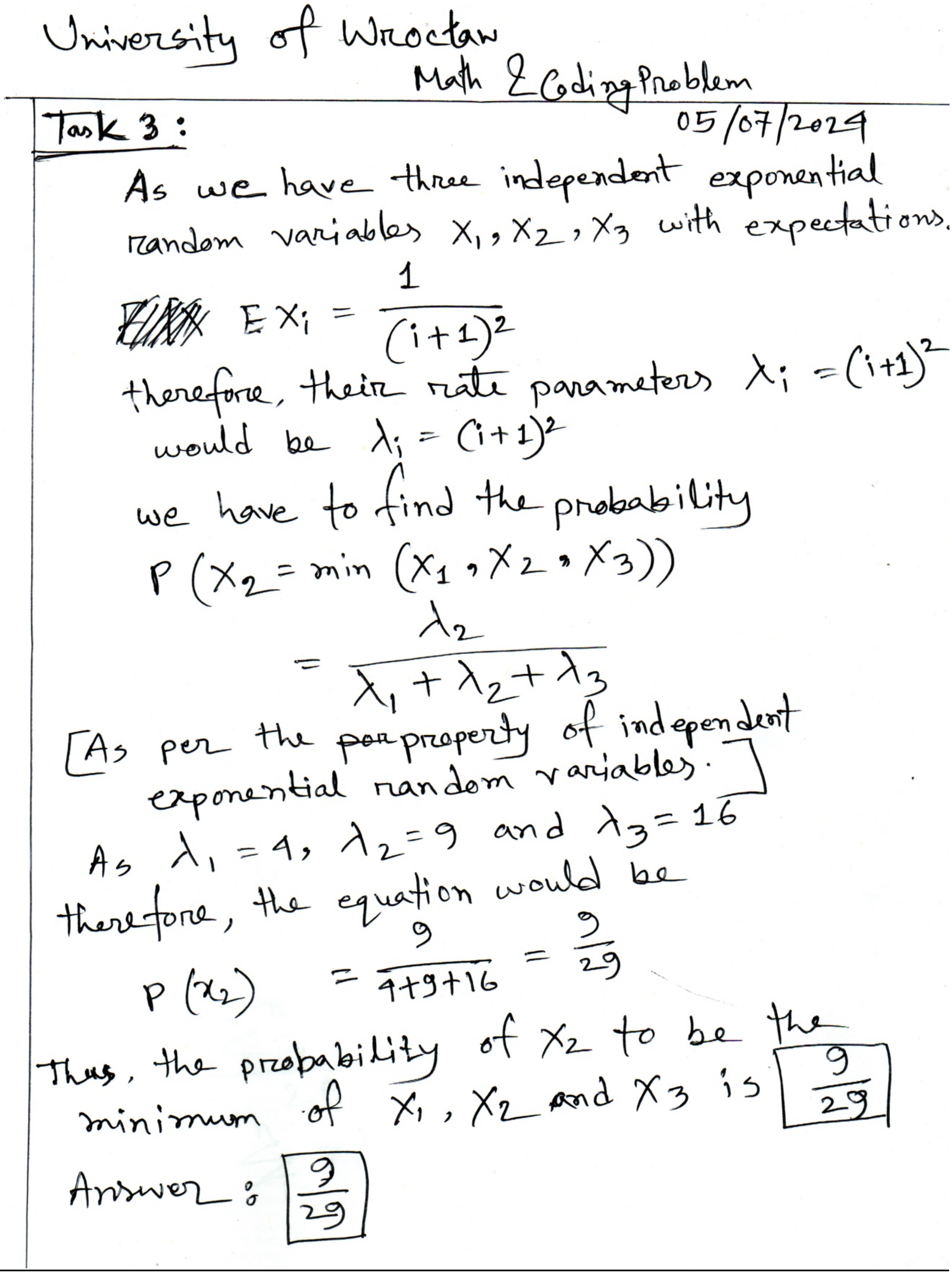
In the example, using accuracy as a sole metric might not be sufficient, as it would penalize systems that recommend multiple products with similar probabilities, even if they're all correct.

Conclusion: While accuracy metrics such as Top-K Accuracy are useful and commonly used for evaluating session-based recommender systems, they should be complemented with other metrics like MRR, Hit Rate, Recall at K, and so on to provide a comprehensive evaluation of the system's performance. This approach ensures that the recommender system is not only accurate but also effective in recommending relevant and useful products to users based on their browsing sessions. Metrics should consider user satisfaction beyond simple correctness, including novelty and diversity of recommendations.

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

Compute P ($X_2 = \min(X_1, X_2, X_3)$).

Solution: Please check the scan of the hand-written solution.



University of Wroclaw

Math & Coding Problem

Task 3 : 05/07/2024

As we have three independent exponential random variables X_1, X_2, X_3 with expectations.

$$E X_i = \frac{1}{(i+1)^2}$$

therefore, their rate parameters $\lambda_i = (i+1)^2$ would be $\lambda_i = (i+1)^2$

we have to find the probability

$$P(X_2 = \min(X_1, X_2, X_3))$$

$$= \frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3}$$

[As per the ~~por~~ property of independent exponential random variables.]

As $\lambda_1 = 4$, $\lambda_2 = 9$ and $\lambda_3 = 16$

therefore, the equation would be

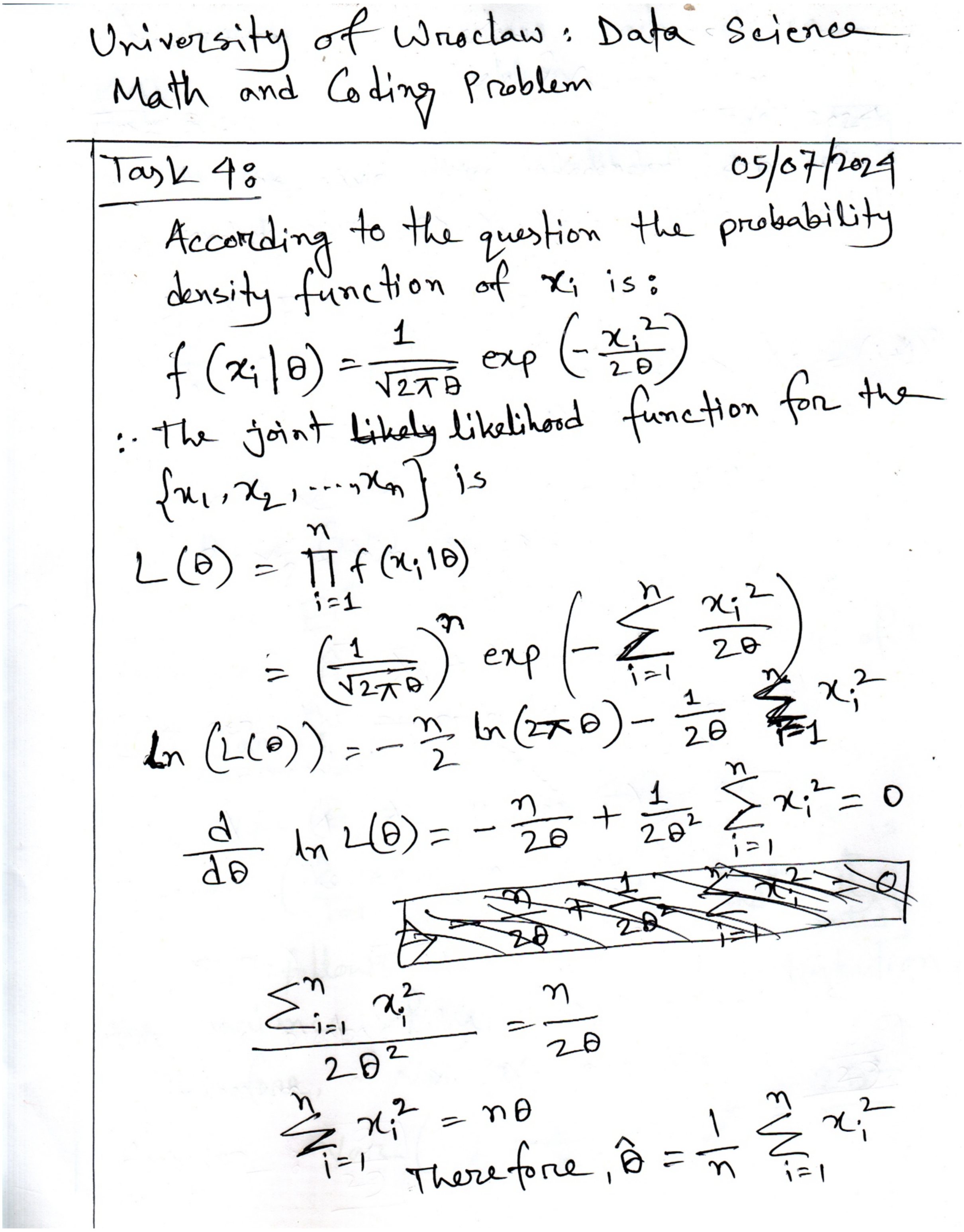
$$P(x_2) = \frac{9}{4+9+16} = \frac{9}{29}$$

Thus, the probability of X_2 to be the minimum of X_1, X_2 and X_3 is $\boxed{\frac{9}{29}}$

Answer : $\boxed{\frac{9}{29}}$

Task 4. Let X1, . . . , Xn be independent identically distributed random variable coming from the population with N(0,θ) 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.

Solution: Please check the scan of the hand-written solution.



University of Wroclaw : Data Science
Math and Coding Problem

Task 4: 05/07/2024

According to the question the probability density function of x_i is:

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x_i^2}{2\theta}\right)$$

∴ The joint ~~likely~~ likelihood function for the $\{x_1, x_2, \dots, x_n\}$ is

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

$$= \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{x_i^2}{2\theta}\right)$$

$$\ln(L(\theta)) = -\frac{n}{2}\ln(2\pi\theta) - \frac{1}{2\theta}\sum_{i=1}^{n} x_i^2$$

$$\frac{d}{d\theta}\ln L(\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 = 0$$

$$\cancel{-\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 = 0}$$

$$\frac{\sum_{i=1}^{n} x_i^2}{2\theta^2} = \frac{n}{2\theta}$$

$$\sum_{i=1}^{n} x_i^2 = n\theta$$

Therefore, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

05/07/2024

Therefore, Maximum likelihood estimator of

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

□ Unbiasedness:

the expectation of θ is

$$E(\hat{\theta}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) = \frac{1}{\cancel{2}n}\sum_{i=1}^{n} E(x_i^2)$$

As $\theta = E(x_i^2)$, therefore

$$E(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta$$

Therefore, $\hat{\theta}$ is an unbiased estimator of θ.

□ Variance:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x^2\right)$$

$$\mathrm{var}\left(\sum_{i=1}^{n} x_i^2\right) = n\,\mathrm{Var}(x_i^2)$$ [As x_i^2 is independent indentically distributed]

As x_i^2 follows a chi-squared distribution, the variance of x_i^2 is $2\theta^2$.

Therefore, $\mathrm{Var}(x_i^2) = 2\theta^2$

$$\therefore \quad \mathrm{Var}(\hat{\theta}) = \frac{1}{\cancel{2}n^2}\, n\, 2\theta^2 = \frac{2\theta^2}{n}$$

University of Wroclaw : Data Science

Math & Coding Problem 05/07/2024

According to the Cramer-Rao lower bound of unbiased estimators, the ~~mini~~ minimum variance unbiased ~~ex~~ estimator for θ is $\frac{2\theta^2}{n}$.

As, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ achieves this bound, therefore $\hat{\theta}$ is the minimum variance unbiased estimator for θ.

Answer :

The maximum likelihood estimator of θ is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

This estimator is also the minimum variance unbiased estimator for θ

Submitted by : Mohammad Faisal Ahmed

<u>Task 5.</u> 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 some- times be more than one correct answer. In this case, the results with higher scores are preferred.

Solution: I have written several solutions for this problem. Please check the attached files which corresponds:

- **Solution_For_Input_1.txt [Contains the result for sudoku1.txt]**
- **Solution_For_Input_2.txt [Contains the result for sudoku2.txt]**
- **Solution_For_Input_3.txt [Contains the result for sudoku3.txt]**
- **Sudoku_Solution_Code.py [It requires input.txt, output.txt and english_words.txt to run]**

My solution can produce thousands of simulations if you have required computational power. For this problem, I have just generated as much as 30 solutions and ranked them based on the scores among these solutions only. The solution files contains the maximum scored sudoku solution and also all other generated simulations (solved and unsolved state).

Solutions that generated the highest scores among 30 simulations are:

Sudoku 1:

```
E867......5..4AC
.3DF4.5E..C7B.9G
5.GA..6C49.B1E8.
9C.48.1.EFA23576
AFED1C.9B.G35..4
8.9..B.514F.C3.7
.136..87DC25GAB9
75CB..D.6A982F..
.D4G59...21FE..8
....C..6.8B...2A
.B5....1CG6A493.
F..C2.B83ED..65.
D62..87.9..C.1.3
...3.542F17.6D.B
G.15.6C3AB..97F.
B.F.A1G.2.E6.C4.
```

Solution to Sudoku 1:

```
E867329BGD51F4AC
13DF4A5E86C7B29G
52GA7F6C493B1E8D
9CB48D1GEFA23576
AFED1C29B7G35864
8G926BA514FEC3D7
4136FE87DC25GAB9
75CBG3D46A982F1E
6D4G593A721FEBC8
3E71C4F658B9DG2A
2B58D7E1CG6A493F
F9AC2GB83ED47651
D62EB87F954CA1G3
CA839542F17G6DEB
G415E6C3AB8D97F2
B7F9A1GD23E68C45
Maximum Score: 3
```

Sudoku 2:	Solution to Sudoku 2:
..AF4..E..29B7..	G1AF465E3829B7CD
..6....2....A...	E465FG127DBCA938
......37.....2..	8BCDA937G415E2F6
93......FE......	93278BCDFEA6G145
.........G......	1E423F685G97CADB
..........3.....	5G78E1DCAF3B2694
3.....B.D.6.8..F	3C9AG7B4D26E851F
F......5.C....G.	F6DB92A51C483EG7
.....C...B84...2	6F3E1C7G9B845DA2
....D.2..1....8.	A7G9DE2F61534B8C
.....3....CG9...	BD1453862ACG9F7E
.....A......1...	285CBA49E7DF1G63
....2......A....	D5BG249386FA7CE1
.A.3......7....9	CAE368G1B57DF429
7....DE......8..	79F1CDEB43G2685A
4.8675.A.....3B.	428675FAC9E1D3BG
	Maximum Score: 2

Sudoku 3:	Solution to Sudoku 3:
....	BCDEFG123456789A
....	FG12BCDE789A3456
....	3456789ABCDEFG12
....	789A3456FG12BCDE
....	CBEDGF21436587A9
....	GF21CBED87A94365
....	436587A9CBEDGF21
....	87A94365GF21CBED
....	DEBC12FG56349A78
....	12FGDEBC9A785634
....	56349A78DEBC12FG
....	9A78563412FGDEBC
....	EDCB21GF6543A987
....	21GFEDCBA9876543
....	6543A987EDCB21GF
....	A987654321GFEDCB
	Maximum Score: 12