CS215 Assignment 1 Report

Atharva Bendale (22B0901)
Vishal Bysani (22B1061)

August 2023

Contents

Introduction

In this report, we have written down our solutions for the problems in the Assigment 1 of CS 215 course on Data Analysis and Interpretation.

1 Problem 1

Consider n people each of whom owns a book. The book belonging to each of n persons is put into a basket. The people then pick up a book at random, due to which it is equally likely that a given person could pick any one of the n books from the basket. What is the probability that

1. every person picks up his or her book back?
2. the first $m < n$ persons who picked up a book receive their own book back again?
3. each person among the first m persons to pick up the book gets back a book belonging to one of the last m persons to pick up the books?
4. Now suppose that every book put into the box has an independent probability p of getting unclean, i.e. this is independent of who picked up which book and independent of whether other books became unclean. What is the probability that the first m persons will pick up clean books?
5. Continuing from the previous point, what was the probability that exactly m persons will pick up clean books? [3 × 5 = 15 points]

Solution :

Let A_i be the event of i^{th} person picking up his/her own book.

a) The event when everyone picks up his/her own book is $\bigcap_{k=1}^{n} A_i$. We know that :

$$P(\bigcap_{k=1}^{n} A_i) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot P(A_4|A_1 \cap A_2 \cap A_3) \dots P(A_n|\bigcap_{k=1}^{n-1} A_i)$$

$$P(\bigcap_{k=1}^{n} A_i) = \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \cdot \dots \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{1}.$$

$$P(\bigcap_{k=1}^{n} A_i) = \frac{1}{n!}$$

b) The event when first $m(\leq n)$ persons pick up their own books is $\bigcap_{k=1}^{m} A_i$. We know that :

$$P(\bigcap_{k=1}^{m} A_i) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot P(A_4|A_1 \cap A_2 \cap A_3) \dots P(A_m|\bigcap_{k=1}^{m-1} A_i)$$

$$P(\bigcap_{k=1}^{m} A_i) = \frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \cdot \dots \frac{1}{n-m+3} \cdot \frac{1}{n-m+2} \cdot \frac{1}{n-m+1}.$$

$$P(\bigcap_{k=1}^{m} A_i) = \frac{(n-m)!}{n!}$$

c) Let B_i be the event of the i^{th} person picking up one of the books of the last m persons. Then the event of the first m persons picking up the books of last m persons is $\bigcap_{k=1}^{m} B_i$

$$P(\bigcap_{k=1}^{m} B_i) = P(B_1) \cdot P(B_2|B_1) \cdot P(B_3|B_1 \cap B_2) \cdot P(B_4|B_1 \cap B_2 \cap B_3) \dots P(B_m|\bigcap_{k=1}^{m-1} B_i)$$

$$P(\bigcap_{k=1}^{m} B_i) = \frac{m}{n} \cdot \frac{m-1}{n-1} \cdot \frac{m-2}{n-2} \cdot \dots \frac{3}{n-m+3} \cdot \frac{2}{n-m+2} \cdot \frac{1}{n-m+1} \cdot$$

$$P(\bigcap_{k=1}^{m} B_i) = \frac{m! \cdot (n-m)!}{n!}$$

d) The events of picking any book and any book being unclean are independent, therefore, probability of first m persons picking up clean books is same as probability of some m books being clean. Let the event be X then:

$$P(X) = (1-p)^m$$

e) Similar to the above d problem this problem is also reduced to calculating probability of chossing m persons and the m books with them are clean and remaining $n-m$ books being unclean. Let the event be X:

$$P(X) = \binom{n}{m} \cdot p^{n-m} \cdot (1-p)^m$$

2 Problem 2

Given n distinct values $\{x_i\}_{i=1}^n$ with mean μ and standard deviation σ, prove that for all i, we have $|x_i - \mu| \leq \sigma\sqrt{n-1}$. How does this inequality compare with Chebyshev's inequality as n increases? (give an informal answer) [7+3=10 points]

Solution :

We know that square of standard deviation (variance) is given by,

$$\sigma^2 = \frac{\Sigma_{i=1}^n (x_i - \mu)^2}{n-1}$$

$$\implies (n-1)\sigma^2 = \Sigma_{i=1}^n (x_i - \mu)^2$$

$$\implies (n-1)\sigma^2 \geq (x_i - \mu)^2 \quad \forall i \quad \text{from} \quad 1 \quad \text{to} \quad n$$

$$\implies \sqrt{n-1} \cdot \sigma \geq |x_i - \sigma|$$

Hence Proved.

Now since we have proved that the inequality $|x_i - \mu| \leq \sigma\sqrt{n-1}$ holds for all i, we can say that number of sample points which deviate from the mean μ by $\sqrt{n-1}$ times standard deviation σ is n. So the proportion of such samples points is 1.
According to Chebyshev's inequality, the proportion of samples points which deviate from the mean by k or more than k times the standard deviation is lesser than or equal to $\frac{1}{k^2}$

$$S_k = x_i : |x_i - \mu| \leq k\sigma$$

$$\frac{|S_k|}{N} \leq \frac{1}{k^2}$$

So according to the problem , the value of k is $\sqrt{n-1}$ and the required proportion of samples points satisfying the given condition is given by,

$$P = 1 - \frac{|S_k|}{N} \geq 1 - \frac{1}{k^2}$$

$$\implies P = 1 - \frac{|S_k|}{N} \geq 1 - \frac{1}{n-1}$$

$$\implies P = 1 - \frac{|S_k|}{N} \geq \frac{n-2}{n-1}$$

$$P \to 1 \text{ as } x \to \infty$$

As n becomes very large, we can see that the right side of the inequality tends to 1. Hence the lower bound of Chebyshev's inequality becomes stricter and approaches the conclusion that we have reached from our proof (that the proportion of such samples points is 1).

3 Problem 3

Consider $\epsilon > 0$ and any two values Q_1, Q_2. Let F be the event that $\{|Q_1 + Q_2| > \epsilon\}$ and let E be the event that $\{|Q_1| + |Q_2| > \epsilon\}$. Also, let E_1 and E_2 be the events $\{|Q_1| > \epsilon/2\}$ and $\{|Q_2| > \epsilon/2\}$ respectively. Then show that $P(F) \leq P(E_1) + P(E_2)$. [10 points]

Solution :

We know that for any two numbers a & b :

$$|a| + |b| \geq |a + b| \qquad \textit{(By triangle inequality)}$$

If event F occurs then:

$$|Q_1 + Q_2| \geq \epsilon \quad \Rightarrow \quad |Q_1| + |Q_2| \geq \epsilon$$

$$\therefore F \subseteq E$$

$$\therefore P(F) \leq P(E) \tag{1}$$

We can also observe that:

$$E_1^C \cap E_2^C \subseteq E^C$$

$$P(E_1^C \cap E_2^C) \leq P(E^C) \tag{2}$$

As we know E_1 and E_2 are independent events we can say that:

$$P(E_1^C \cap E_2^C) = P(E_1^C) \cdot P(E_2^C)$$

$$P(E_1^C) \cdot P(E_2^C) \leq P(E^C) \qquad : from \quad (2)$$

$$(1 - P(E_1)) \cdot (1 - P(E_2)) \leq 1 - P(E)$$

$$P(E) \leq P(E_1) + P(E_2) - P(E_1) \cdot P(E_2)$$

$$P(E) \leq P(E_1) + P(E_2)$$

$$P(F) \leq P(E_1) + P(E_2) \qquad : from \quad (1)$$

$\therefore$ Hence proved.

4 Problem 4

Let Q_1, Q_2 be non-negative random variables. Let $P(Q_1 < q_1) \geq 1 - p_1$ and $P(Q_2 < q_2) \geq 1 - p_2$, where q_1, q_2 are non-negative. Then, show that $P(Q_1 Q_2 < q_1 q_2) \geq 1 - (p_1 + p_2)$. [10 points]

Solution :

We can observe that:

$$(Q_1 < q_1) \cap (Q_2 < q_2) \subseteq (Q_1 \cdot Q_2 < q_1 \cdot q_2)$$

$$P((Q_1 \cdot Q_2 < q_1 \cdot q_2)) \geq P((Q_1 < q_1) \cap (Q_2 < q_2))$$

As both events $(Q_1 < q_1)\&(Q_2 < q_2)$ are independent:

$$P((Q_1 \cdot Q_2 < q_1 \cdot q_2)) \geq P((Q_1 < q_1)) \cdot P((Q_2 < q_2))$$

$$P((Q_1 \cdot Q_2 < q_1 \cdot q_2)) \geq (1 - p_1) \cdot (1 - p_2)$$

$$P((Q_1 \cdot Q_2 < q_1 \cdot q_2)) \geq 1 - (p_1 + p_2) + p_1 \cdot p_2$$

$$P((Q_1 \cdot Q_2 < q_1 \cdot q_2)) \geq 1 - (p_1 + p_2)$$

$\therefore$ Hence proved

5 Problem 5

A contestant is on a game show and is allowed to choose between three doors. Behind one of them lies a car, behind the other two there lies a stone. The contestant will be given whatever is behind the door that (s)he picked, and quite naturally (s)he wants the car. Suppose (s)he chooses the first door, and the host of the show who knows what is behind every door, opens (say) the third door, behind which there lies a stone (without opening the first door). The host now asks the contestant whether (s)he wishes to choose the second door instead of the first one. Your task is to determine whether switching the contestant's choice is going to increase his/her chance of winning the car. Remember that the host is intelligent: (s)he is always going to open a door not chosen by the contestant, *and* is also going to open a door behind which there is a stone. You should approach this problem only from the point of view of conditional probability as follows. To this end, let C_1, C_2, C_3 be events that the car is behind doors 1,2,3 respectively. Assume $P(C_i) = 1/3, i \in \{1, 2, 3\}$.

1. Let Z_1 be the event that the contestant chose door 1. Write down the value of $P(C_i|Z_1)$ for all $i \in \{1, 2, 3\}$.
2. Let H_3 be the event that the host opened door 3. Write down the value of $P(H_3|C_i, Z_1)$ for all $i \in \{1, 2, 3\}$.
3. Clearly the conditional probability of winning by switching is $P(C_2|H3, Z1)$. This is equal to $\frac{P(H_3|C_2, Z_1)P(C2, Z1)}{P(H_3, Z_1)}$. Evaluate this probability. Note that $P(A_1, A_2)$ denotes the joint probability of events A_1, A_2.
4. Likewise evaluate $P(C_1|H3, Z1)$.
5. Conclude whether switching is indeed beneficial.
6. Now let us suppose that the host were quite whimsical and decided to open one of the two doors not chosen by the contestant, with equal probability, not caring whether there was a car behind the door. In this case, repeat your calculations and determine whether or not it is beneficial for the contestant to switch choices. [2+2+5+5+1+5=20 points]

Solution:

1. The probability of the car occurring behind any door is independent of the contestant choosing a door. So $P(C_i|Z1) = P(C_i) = \frac{1}{3} \quad i \in \{1, 2, 3\}$.

2. $P(H_3|C_1, Z_1) = \frac{1}{2}$ (If the car is behind door 1, and the contestant also chooses door 1, then the host can choose either door 2 or 3)

   $P(H_3|C_2, Z_1) = 1$ (If the car is behind door 2, and the contestant chooses door 1, the the host can only choose door 3, since he can neither choose the door behind which car is there nor the one which contestant chooses)

   $P(H_3|C_3, Z_1) = 0$ (If the car is behind door 3, then the host can't choose it)

3. $$P(C_2|H_3, Z_1) = \frac{P(C_2 \cap H_3, Z_1)}{P(H_3, Z_1)} = \frac{P(H_3 \cap C_2, Z_1)}{P(H_3, Z_1)} = \frac{P(H_3|C_2, Z_1)P(C2, Z1)}{P(H_3, Z_1)} = \frac{1 \cdot P(Z_1) \cdot P(C_2)}{P(Z_1) \cdot P(H_3|Z_1)}$$

   $$P(H_3|Z_1) = P(C_1) \cdot P(H_3|C_1, Z_1) + P(C_2) \cdot P(H_3|C_2, Z_1) + P(C_3) \cdot P(H_3|C_3Z_1) = \frac{1}{3} \cdot (\frac{1}{2} + 1 + 0) = \frac{1}{2}$$

   $$\therefore P(C_2|H_3, Z_1) = \frac{2}{3}$$

4. $$P(C_1|H_3, Z_1) = \frac{P(C_1 \cap H_3, Z_1)}{P(H_3, Z_1)} = \frac{P(H_3 \cap C_1, Z_1)}{P(H_3, Z_1)} = \frac{P(H_3|C_1, Z_1)P(C1, Z1)}{P(H_3, Z_1)} = \frac{\frac{1}{2} \cdot P(Z_1) \cdot P(C_1)}{P(Z_1) \cdot P(H_3|Z_1)} = \frac{1}{3}$$

5. From 3. and 4. we can conclude that switching is indeed beneficial.

6. If the host decides to open one of the two doors not chosen by the contestant, with equal probability, not caring whether there was a car behind the door,

   $$P(H_3|C_1, Z_1) = P(H_3|C_2, Z_1) = P(H_3|C_3, Z_1) = \frac{1}{2}$$

   $$P(C_2|H_3, Z_1) = \frac{P(C_2 \cap H_3Z_1)}{P(H_3, Z_1)} = \frac{P(H_3 \cap C_2Z_1)}{P(H_3, Z_1)} = \frac{P(H_3|C_2, Z_1)P(C2, Z1)}{P(H_3, Z_1)} = \frac{\frac{1}{2} \cdot P(Z_1) \cdot P(C_1)}{P(Z_1) \cdot P(H_3|Z_1)}$$

   $$P(H_3|Z_1) = P(C_1) \cdot P(H_3|C_1Z_1) + P(C_2) \cdot P(H_3|C_2Z_1) + P(C_3) \cdot P(H_3|C_3Z_1) = \frac{1}{3} \cdot (\frac{1}{2} + \frac{1}{2} + \frac{1}{2}) = \frac{1}{2}$$

   $$\therefore P(C_2|H_3, Z_1) = \frac{1}{3}$$

   Similarly, $P(C_1|H_3, Z_1) = \frac{1}{3}$

In this case, since $P(C_1|H_3, Z_1) = P(C_2|H_3, Z_1) = \frac{1}{3}$, it would not be beneficial for the contestant to switch choices. If the host opens the door behind which the car is there, then the contestant would win the car irrespective of switching.

6 Problem 6

Generate a sine wave in MATLAB of the form $y = 6.5\sin(2.1x + \pi/3)$ where x ranges from -3 to 3 in steps of 0.02. Now randomly select a fraction $f = 30\%$ of the values in the array y (using MATLAB function 'randperm') and corrupt them by adding random values from 100 to 120 using the MATLAB function 'rand'. This will generate a corrupted sine wave which we will denote as z. Now your job is to filter z using the following steps.

- Create a new array y_{median} to store the filtered sine wave.
- For a value at index i in z, consider a neighborhood $N(i)$ consisting of $z(i)$, 8 values to its right and 8 values to its left. For indices near the left or right end of the array, you may not have 8 neighbors in one of the directions. In such a case, the neighborhood will contain fewer values.
- Set $y_{median}(i)$ to the median of all the values in $N(i)$. Repeat this for every i.

This process is called as 'moving median filtering', and will produce a filtered signal in the end. Repeat the entire procedure described here using the arithmetic mean instead of the median. This is called as 'moving average filtering'. Repeat the entire procedure described here using the first quaritle (25 percentile) instead of the median. This is called as 'moving quartile filtering'. Plot the original (i.e. clean) sine wave y, the corrupted sine wave z and the filtered sine wave using each of the three methods on the same figure in different colors. Introduce a legend on the plot (find out how to do this in MATLAB). Include an image of the plot in your report. Now compute and print the relative mean squared error between each result and the original clean sine wave. The relative mean squared error between y and its estimate $\hat{y}$ (i.e. the filtered signal - by any one of the different methods) is defined as $\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$.
Now repeat all the steps above using $f = 60\%$, and include the plot of the sine waves in your report, and write down the relative mean square error values.
Which of these methods (median/quartile/arithmetic mean) produced better relative mean squared error? Why? Explain in your report. [5+5+4+3+3=20 points]

Solution :

Solution : This problem is solved in matlab, all the relevant files are uploaded along with the other files in the folder **Problem6**. Directions for code usage :

1. You can load all the variables we used with `load("matlab.mat")` or you can run the code afresh using `run("main.m")`.
2. You can see the flow of code in the `main.m` file and can run files manually also.
3. Running "main.m" will display errors of various filtering methods.
4. Finally you can use command `run("plot_all_graphs30.m")` and `run("plot_all_graphs60.m")` for plotting graphs for $f = 30\%$ and $f = 60\%$ respectively.

For fraction f = 30%,

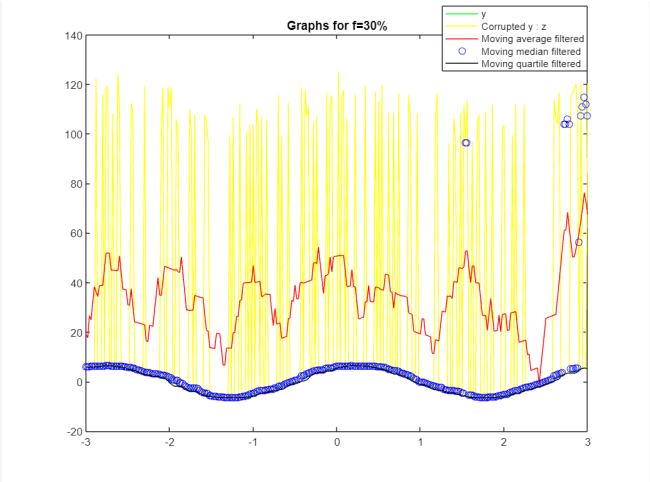

Figure 1: X, Y, corrupted wave and all filtered waves.

The RMSE (Relative Mean squared error) values of all filtering methods:

Moving Average Filtering = 58.9010
Moving Median Filtering = 18.8291
Moving Quartile Filtering = 0.0147

For fraction f = 60%,

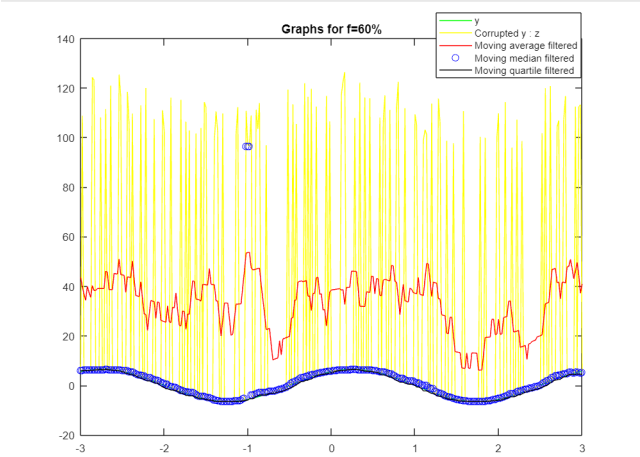

Figure 2: X, Y, corrupted wave and all filtered waves.

The RMSE (Relative Mean squared error) values of all filtering methods:

Moving Average Filtering = 55.1725
Moving Median Filtering = 4.9410
Moving Quartile Filtering = 0.0131

From the relative mean squared errors, it is evident that the moving quartile method produced better relative mean squared error. The reason is that the quartiles are less sensitive to outliers as compared to mean and median. If the data consists of outliers then the moving quartile method provided robust estimates of central tendency.

7 Problem 7

Suppose that you have computed the mean, median and standard deviation of a set of n numbers stored in array A where n is very large. Now, you decide to add another number to A. Write a MATLAB function to update the previously computed mean, another MATLAB function to update the previously computed median, and yet another MATLAB function to update the previously computed standard deviation. Note that you are *not* allowed to simply recompute the mean, median or standard deviation by looping through all the data. You may need to derive formulae for this. Include the formulae and their derivation in your report. Note that your MATLAB functions should be of the following form

```
function newMean = UpdateMean (OldMean, NewDataValue, n),
function newMedian = UpdateMedian (oldMedian, NewDataValue, A, n),
function newStd = UpdateStd (OldMean, OldStd, NewMean, NewDataValue, n).
```

Also explain, how would you update the histogram of A, if you received a new value to be added to A? (Only explain, no need to write code.) **Note:** For updating the median, you may assume that the array A is sorted in ascending order, that the numbers are all unique. For sorted arrays with a even number of elements, MATLAB returns the answer as $(A(N/2) + A(N/2+1))/2$. You may use MATLAB's convention though it is not strictly required. Recall that the standard deviation with n values $A_1, ..., A_n$ is given as $s_n = \sqrt{\sum_{i=1}^{n}(A_i - \bar{A}_n)^2/(n-1)}$ and $\bar{A}_n = \sum_{i=1}^{n} A_i/n$. [4+5+5+1 = 15 points]

Solution:

For finding the newMean:

$$\text{oldMean} = \Sigma_{i=1}^{n} A_i/n$$

$$\implies \Sigma_{i=1}^{n} A_i = \text{oldMean} \cdot n$$

Let the added element be A_{n+1}. So the new sum of all elements is

$$\Sigma_{i=1}^{n+1} A_i = \Sigma_{i=1}^{n} A_i + A_{n+1} = \text{oldMean} \cdot n + A_{n+1}$$

$$\therefore \text{The new mean is :} \frac{\text{oldMean} \cdot n + A_{n+1}}{n+1}$$

The function to implement the above updation of mean is present in the **UpdateMean.m** file under **Problem7 folder**

For finding the newMedian:

Two cases are made depending on whether the initial number of elements in the array is even or odd. If it is even, then the newly added value is compared with the 2 middle elements of the original array and the median is updated accordingly. If the initial number of elements are odd, then the newly added value is compared with the 3 elements present in the center of the original array and median is updated accordingly.

The function to implement the updation of median is present in the **UpdateMedian.m** file under **Problem7 folder**.

For finding the new standard deviation:
The old standard deviation is given by

$$\text{OldStd} = \sqrt{\sum_{i=1}^{n}(A_i - \text{OldMean})^2/(n-1)}$$

$\implies \text{OldStd}^2 \cdot (n-1) = \sum_{i=1}^{n} A_i^2 - n \cdot \text{OldMean}^2$

$\implies \sum_{i=1}^{n} A_i^2 = \text{OldStd}^2 \cdot (n-1) + n \cdot \text{OldMean}^2$

Now the new summation of squares of the elements after adding new element is

$\sum_{i=1}^{n+1} A_i^2 = \text{OldStd}^2 \cdot (n-1) + n \cdot \text{OldMean}^2 + A_{n+1}^2$

The new standard deviation is given by:

$$\text{NewStd} = \frac{\sum_{i=1}^{n+1} A_i^2 - (n+1) \cdot \text{NewMean}^2}{n}$$

$$\implies \text{NewStd} = \frac{\text{OldStd}^2 \cdot (n-1) + n \cdot \text{OldMean}^2 + A_{n+1}^2 - (n+1) \cdot \text{NewMean}^2}{n}$$

The function to implement the above updation of standard deviation is present in the **UpdateStd.m** file of **Problem7** folder.

On adding a new value to A, first we would find the particular interval (bin) in which the value lies and increase the frequency of that bin by one to update the histogram.