CS360/SE360
Artificial Intelligence
Group project for Optimal Samples Selection System

It is known that the amount of data generated has been increasingly tremendous in the past few years due to Covid-19 pandemic, the ease of accessing the internet and cheap or inexpensive mass storage devices. The ease of transferring data through internet, communication lines and digital data are used in all walks of life. Nowadays, these big data have been used for data mining, knowledge discovery, machine learning, statistical learning, statistical analysis and experiments. In order to extract or discover useful data, information or knowledge from these big data, one of the methods we usually adopted is the optimal samples selection.

In this group project, you are expected to extract a subset of samples from these big data. In order to extract this subset of data (samples), we have to make sure that the subset of samples extracted or selected should be as fair and unbiased as possible and also as optimal as possible. In the following we propose a useful method.

Assuming that there are m samples ($45 \le m \le 54$), any group of n ($7 \le n \le 25$) samples out of these m samples are randomly selected. There are ${}_mC_n$ groups of n samples. From one of these groups of n samples, we randomly select e.g., $k=6$ ($4 \le k \le 7$) samples to form some groups. So there will be ${}_nC_k$ groups of $k=6$ samples selected. Among these groups of $k=6$ samples, we would like to optimize them by selecting ONLY some of them. The conditions that need to be fulfilled are listed as follows:

1. There are at least ONE group of k samples, in which s ($3 \le s \le 7$) samples have been selected from the j (where $s \le j \le k$) samples, i.e., when $j=4$, we have $s=3$ or 4; when $j=5$, we have $s=3$, 4 or 5; when $j=6$, we have $s=3, 4, 5$ or 6; and when $j=7$, we have $s=3, 4, 5, 6$ or 7.

   E.g. 1, when $m=45$, $n=7$ (we randomly choose 7 samples, A,B,C,D,E,F,G and $k=6$, $j=5$, $s=5$, we obtain the following minimum 6 groups of $k=6$ samples, which guarantee at least ONE group of $k=6$ samples has ALL $s=5$ samples groups from ALL $j=5$ samples groups of $n=7$ samples,(i.e.,${}_nC_j={}_7C_5$ and ${}_jC_s={}_5C_5$).

   1. A,B,C,D,E,G
   2. A,B,C,D,F,G
   3. A,B,C,E,F,G
   4. A,B,D,E,F,G,
   5. A,C,D,E,F,G
   6. B,C,D,E,F,G

   E.g. 2, when $m=45$, $n=8$ (we randomly choose 8 samples, A,B,C,D,E,F,G,H and $k=6$, $j=4$, $s=4$, we obtain the following minimum 7 groups of $k=6$ samples, which guarantees at least ONE group of $k=6$ samples has ALL $s=4$ samples groups from ALL $j=4$ samples groups of $n=8$ samples, (i.e., ${}_nC_j={}_8C_4$ and ${}_jC_s={}_4C_4$).

   1. A,B,C,D,G,H
   2. A,B,C,E,G,H
   3. A,B,C,F,G,H
   4. A,B,D,E,F,G
   5. A,C,D,E,F,H
   6. B,C,D,E,F,H
   7. C,D,E,F,G,H

   E.g. 3, when $m=45$, $n=9$ (we randomly choose 9 samples, A,B,C,D,E,F,G,H,I and $k=6$, $j=4$, $s=4$, we obtain the following minimum 12 groups of $k=6$ samples, which guarantees at least ONE group of $k=6$ samples has ALL $s=4$ samples groups from ALL $j=4$ samples groups of $n=9$ samples, (i.e., ${}_nC_j={}_9C_4$ and ${}_4C_4$).

   1. A,B,C,D,E,I
   2. A,B,C,E,G,H
   3. A,B,C,F,H,I
   4. A,B,D,E,F,G
   5. $A,B,D,G,H,I.$
   6. A,C,D,E,F,H
   7. A,C,D,F,G,I
   8. A,E,F,G,H,I
   9. B,C,D,F,G,H
   10. B,C,E,F,G,I
   11. B,D,E,F,H,I
   12. C,D,E,G,H,I

   E.g.4, when $m=45$, $n=8$ (we randomly choose 8 samples, A,B,C,D,E,F,G,H and $k=6$, $j=6$, $s=5$, we obtain the following minimum 4 groups of $k=6$ samples, which guarantees at least ONE group of $k=6$ samples has at least ONE $s=5$ samples group from ALL $j=6$ samples groups of $n=8$ samples, (i.e., ${}_nC_j={}_8C_6$ and ${}_6C_5$).

   1. A,B,C,E,G,H
   2. A,B,D,F,G,H
   3. A,C,D,E,F,H
   4. B,C,D,E,F,G

E.g. 5, when *m=45, n=8* (we randomly choose *8* samples, *A,B,C,D,E,F,G,H* and *k=6, j=6, s=5*, we obtain the following minimum *10* groups of *k=6* samples, which guarantees at least ONE group of *k=6* samples has at least FOUR *s=5* samples groups from ALL *j=6* samples groups of *n=9* samples, (i.e., ${}_nC_j={}_8C_6$ and ${}_6C_5$).

1. *A,B,C,D,E,H*
2. *A,B,C,E,F,H*
3. *A,B,C,E,G,H*
4. *A,B,D,E,F,G*
5. *A,B,D,F,G,H.*
6. *A,C,D,E,F,G*
7. *A,D,E,F,G,H*
8. *B,C,D,E,G,H*
9. *B,C,D,F,G,H*
10. *B,D,E,F,G,H*

E.g. 6, when *m=45, n=9* (we randomly choose *9* samples, *A,B,C,D,E,F,G,H,I* and *k=6, j=5, s=4*, we obtain the following minimum *3* groups of *k=6* samples, which guarantees at least ONE group of *k=6* samples has at least ONE *s=4* samples group from ALL *j=5* samples groups of *n=9* samples, (i.e., ${}_nC_j={}_9C_5$ and ${}_5C_4$).

1. *A,B,D,F,G,H*
2. *A,C,E,G,H,I*
3. *B,C,D,E,F,I*

E.g. 7, when *m=45, n=10* (we randomly choose *10* samples, *A,B,C,D,E,F,G,H,I,J* and *k=6, j=6, s=4*, we obtain the following minimum *3* groups of *k=6* samples, which guarantees at least ONE group of *k=6* samples has at least ONE *s=4* samples group from ALL *j=6* samples groups of *n=10* samples, (i.e., ${}_{10}C_6$ and ${}_6C_4$).

1. *A,B,E,G,I,J*
2. *A,C,E,G,H,J*
3. *B,C,D,F,H,I*

E.g. 8, when *m=45, n=12* (we randomly choose *12* samples, *A,B,C,D,E,F,G,H,I,J,K,L* and *k=6, j=6, s=4*, we obtain the following minimum *6* groups of *k=6* samples, which guarantees at least ONE group of *k=6* samples has at least ONE *s=4* samples group from ALL *j=6* samples groups of *n=12* samples. (i.e., ${}_nC_j={}_{12}C_6$ and ${}_jC_s={}_6C_4$).

1. *A,B,D,G,K,L*
2. *A,C,D,H,J,L*
3. *A,D,E,F,I,L*
4. *B,C,G,H,J,K.*
5. *B,E,F,G,I,K*
6. *C,E,F,H,I,J*

2. A user friendly interface should be provided. A system title is given as, e.g., "An Optimal Samples Selection System".
3. The user needs to input the values for parameters *m, n, k, j* and *s*. They are all positive integers. See Screens below.
4. The system can either randomly select *n* out of *m* numbers or a user can input *n* out of *m* numbers, and displays these *n* numbers on screen.
5. Output groups of *k=6* samples (results) to a DB file, e.g., *40-9-6-4-4-x-y* for *m=40, n=9, k=6, j=s=4* for the x^{th} run and *y* is the number of results obtained.
6. Provide a way to **EXECUTE and DELETE, etc.** the obtained groups of *k* samples (results) onto the screen from a DB file, e.g., *45-9-6-4-4-x-y.*These groups of *k=6* samples are selected from the list. Please see the screens below.
7. Students are required to form groups yourselves. Each group should/must have 4 students. No group of less than 4 students is allowed unless under some specific condition, i.e., the total number of students not divisible by 4. You are advised to include in your group at least ONE student who knows how to do programming in any programming languages.
8. Use numeral values, e.g., positive INTEGERS, 01,02,03,…..,54 instead of big capital letters A,B,C,D,E,F….,Z for the *m, n, k, j* and *s* numbers.
9. Submit to me **names and student IDs** of your group members using half of an A4 paper next week. Group numbers will be provided to you later for presentation in week 13.
10. A presentation and demonstration of your project is a MUST in weeks 14 and/or **15**.
11. Each group is required to have a **15 minutes** presentation which includes the brief introduction, detailed description of method(s) adopted/used, what you have been achieved in this project, and a demonstration of your system is a MUST in the presentation.

12. A clear, succinct, easy to understand **detailed** *user manual/guide* on how to **INSTALL and EXECUTE** your DEVELOPED system **step by step** and a *project report* must be submitted in hardcopy. The *PROJECT REPORT* must include **method(s)/methodology (supported by diagram(s), etc.), features you have developed/used, contributions such as good running time, optimal/near optimal results, etc., and problems such as long time to get results, results not good enough, etc. of your system, results of sample runs,** etc. should be submitted also in **hardcopy**.
13. You are required to submit a **USB** which contains your developed system, all your source files (codes), free/share ware, database files, DB files of $k=6$ samples (sample runs outputs/results), and the **REPORT** mentioned in point 12.
14. Bonuses will be given to group(s) that allow users to select as many different parameters as possible for *m, n, k, j* and *s*, good method(s) adopted, could generate optimal/ **near optimal** *solutions*. Furthermore, bonuses will be given to the developed system(s) that could be executed in a short time and run on a mobile phone.
15. All teams must submit their projects in a **USB** and **hardcopy** of the **REPORTs (User Manual and Project Report)** in **Week 15**. Group number, names, student numbers of your group members should be listed in your **User Manual** and **REPORT**.

S1 An Optimal Samples Selection System

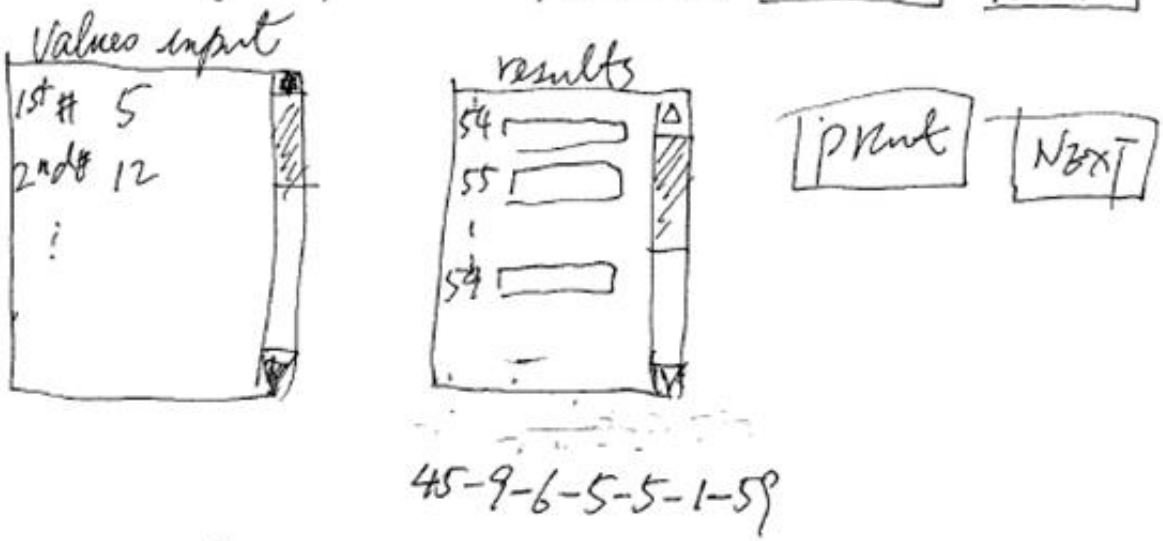
Please input the following parameters

m ☐ $45 \le m \le 54$ n ☐ $7 \le n \le 25$

K ☐ $4 \le K \le 7$ j ☐ $S \le j \le K$ S ☐ $3 \le S \le 7$

at least ☐ S samples

○ Random n ○ Input n [execute] [store] [clear]



User input

S2 An Optimal Samples Selection System

Data Base Resources

Display Delete

- 45-8-6-5-5-1-59
- 45-10-6-5-4-1-20
- 45-10-6-4-4-1-30
- ...