CS203A: Assignment 2 (Spring 2022)

Submission Deadline: 22 April 2022 23:59 IST. Total Marks:100

Author:

Name: **Harshit Raj** Roll: **200433** Email: **harshitr20@iitk.ac.in**

---

1. (15+15 marks) **Endsem allocation**

   You are allocated as the Tutor of CS203, with n students. Rajat has created 2 sets of Endsem papers to decrease cheating. He has asked you to help decide which paper should be given to whom. You scraped through the data on Hello, and found out who have been project partners in previous courses, as they will be friends now. Thus, you have found out m friendship connections among the students. You reported this to Rajat, and he said he is fine with any allocation that disrupts atleast half of the friendship connections. A friendship connection is disrupted if the students get different sets of papers.

   (a) You are really busy, and just randomly allocated each student to set 1 or set 2. Show that the expected value of disrupted friendship connections is $\frac{m}{2}$.

   (b) Getting expected value is not enough, you need to find a proper allocation. But you cannot go over all the 2^n allocations as $n \approx 150$. Using the construction for pairwise independence given in class, show that you can find an allocation with at least half of the friendship connections disrupted in poly(n)-time.

   Solution:

   (a) We know that students have been allocated randomly.
   Say, $P_i(x)$ is probability of i^{th} student getting set x. Where x is either 1 or 2.
   Clearly,

   $$P_i(1) = \frac{1}{2} \text{ and } P_i(2) = \frac{1}{2} \ \forall i \in \{1, ..., n\}$$

   For each connection C_t of Student S_i and Student S_j. Say probability of Connection t being disrupted is $PC(t)$.
   For any general C_t,

   $$PC(t) = P(\text{Both student have set 1}) + P(\text{Both student have set 2})$$

   $$\implies PC(t) = P_i(1) \times P_j(1) + P_i(2) \times P_j(2)$$

   $$PC(t) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}$$

   $$PC(t) = \frac{1}{2}$$

   Any general connection C_t is disrupted with a probability $\frac{1}{2}$.
   Say $E[A]$ is expected value of disrupted friendship connections,

   $$E[A] = \sum_{i=1}^{i \leq m} PC(i) \cdot 1$$

   $$\implies E[A] = \sum_{i=1}^{i \leq m} \frac{1}{2}$$

   $$\implies E[A] = \frac{m}{2}$$

(b) Since the expected value is $m/2$ there will surely be some combinations greater who have more connections disrupts than expectation. We can put first try to make pools of students and make pools like no inter-pool connections exist to ease our computation.

Every discussion and computation will happen for individual pools. Randomly assign opposite set to a pair. Now, either all students have set or students there's some connection to some student who already assign set opposite to the connection having set.
eg. $S1$ and $S2$ have a connection, both have set 1 and 2 respectively. Now there must exist students connected with these two and assign them set opposite to their connection. Hit and trial with this method.

□

2. (5+10+10+15 marks) **Estimating the number of tickets**

You are given a bag full of N tickets numbered $1, \ldots, N$(N is unknown to you). You can take out tickets one at a time, note their label, and put them back in the bag. Your task is to estimate N. We will do this in the same way as we estimated π in lecture:

(a) Assume you drew out k tickets. What will be the expected value of the mean of these tickets ? Calculate N in terms of this mean, call this $\tilde{N}$.

(b) Chernoff bound can be extended to work on the case when the Random Variables take values other than $\{0, 1\}$. This is known as Hoeffding's inequality. Use it to find a lower bound on the probability that the error in N, using the above calculation, will be less than $\delta N(\delta < 1/2)$. (in terms of N, δ, k)

(c) Assume k, N are odd. In calculation of part (a), instead of using the value of mean, we use the median of the labels of tickets drawn. Prove a lower bound of $1 - 2e^{-\frac{k(1+2\delta)^2}{2(3-2\delta)}}$ on the probability that the error in N using the median will be less than $\delta N(\delta < 1/2)$. (in terms of N, δ, k)

(d) Start with a random hidden value of N in range $10^4 - 10^6$. Write a function that gives k values from $[N]$ when queried with equal probability. Use these values to calculate $\tilde{N}$ as in part (a) and (c), and plot them with respect to increasing $k \leq 1000$. Repeat this estimation for a total of 3 different N, and put the plots in the main answer file. Submit the code you used to generate these plots, along with a readme on how to execute the code, zipped together with the main answer file into a single .zip file.

Solution:

(a) There are N cards and probability of drawing i^{th} card is $\frac{1}{N}$.
Say, expected value when a card is drawn is $E[A]$.

$$E[A] = \sum_{i=1}^{i \leq N} P(i^{th} \text{ card is drawn}) \cdot i$$

$$\implies E[A] = \sum_{i=1}^{i \leq N} \frac{1}{N} \cdot i$$

$$\implies E[A] = \frac{1}{N} \cdot \frac{N \cdot (N+1)}{2}$$

$$\implies E[A] = \frac{N+1}{2}$$

Expected value of draw in any general draw is $\frac{N+1}{2}$ as many random draws one make more closer the mean value will be to this value.

Say the mean value obtained in k draws is m,

$$m = \frac{\tilde{N}+1}{2}$$

$$\implies \tilde{N} = 2m - 1$$

(b) Define a family of random variable,

$$X_i = 2Q_i - 1$$

where Q_i is the number observed in i^{th} draw.
Define,

$$S_k = \sum_{i=1}^{i\leq k} X_i$$

Observe,

i. Since, family of Q_i are mutually independent by nature. Hence, family of X_i are also mutually independent.
ii. $E[S_k] = kN$ (from part (a))
iii. $S_k = k\tilde{N}$ (by definition)
iv. Since, $Q_i \in \{1, ..., N\}$
Hence, $X_i \in \{1, ..., 2N-1\}$

We want to know the lower bound of probability,

$$P((\text{Error in } N) < \delta N)$$

where $\delta < \frac{1}{2}$

$$\implies P(|\tilde{N} - N| < \delta N)$$

multiplying both side of inequality by k

$$\implies P(|k\tilde{N} - kN| < \delta kN)$$

We can also calculate upper bound of the probability:

$$P(|k\tilde{N} - kN| \geq \delta kN)$$

the upper bound of above probability is same as 1 - lower bound of first probability. Above can also be written as,

$$P(|S_k - E[S_k]| \geq \delta kN)$$

Using Hoeffding's inequality,

$$P(|S_k - E[S_k]| \geq \delta kN) \leq 2exp\bigg(- \frac{2(\delta kN)^2}{\sum_{i=1}^{i\leq k}\big((2N-1)-1\big)^2}\bigg)$$

$$\implies P(|S_k - E[S_k]| \geq \delta kN) \leq 2exp\bigg(- \frac{2(\delta kN)^2}{k\cdot(2N-2)^2}\bigg)$$

$$\implies P(|S_k - E[S_k]| \geq \delta kN) \leq 2exp\bigg(- \frac{k}{2}\Big(\frac{\delta N}{N-1}\Big)^2\bigg)$$

Lower bound on the probability that the error in N will be less than $\delta N (\delta < 1/2)$ is

$$1 - 2exp\bigg(- \frac{k}{2}\Big(\frac{\delta N}{N-1}\Big)^2\bigg)$$

(c) Expected value of median draw is $\frac{N+1}{2}$ by uniformity of random variable. Say the observed median is e.

$$\tilde{N} = 2e - 1$$

(d) images and code attached

□

3. (15+15 marks) **Markov Chain**

Consider a homogeneous regular Markov chain with state space S of size $|S|$, and transition matrix M. Suppose that M is symmetric and entry-wise positive.

(a) Show that all the eigenvalues of M are bounded by 1 and that the uniform distribution is the unique stationary probability distribution for M.

(b) Starting from the stationary distribution, express the probability of returning to the same state as the state at $t = 0$ after $n \in \mathbb{N}$ steps in terms of the eigenvalues of M. Compute the limit of the above probability as $n \to \infty$.

You might find the second part to be easier than the first. Feel free to assume the first part and finish the second part (even when you can't prove the first part).

Solution:

(a) We know(from lecture notes) that M can be represented in form of $\sum \lambda_i v_i v_i^T$ where λ_i is a eigenvalue and v_i is its respective eigenvector.

We know that for λ to be a eigenvalue of matrix M, $det(M - \lambda I) = 0$
Put $\lambda = 1$. Consider, $det(M - I)$
We also know that all columns of M add up to 1. Hence, all columns of $M - I$ will add up to 0.
This implies sum of all row vectors add to 0 vector. Hence we can say that all rows are linearly dependent. Which implies $det(M - I) = 0$.
Hence, 1 is an eigenvalue of matrix M.

Assume we have an eigenvalue greater than 1. Also, $M^p = \sum \lambda_i^p v_i v_i^T$
we know that M^p as $p \to \infty$ converges. But since, $\exists i$ such that $\lambda_i > 1$ the term $\lambda_i^p v_i v_i^T$ will diverge. Hence we have a contradiction and we cannot have any $\lambda_i > 1$.

Hence we conclude, $\forall i \ \lambda_i \leq 1$ i.e. bounded by 1.

Since, Eigenvalues are unique in nature. Stationary distribution is eigenvector related to eigenvalue 1.

(b) The m_{ij} of matrix M denotes probability of going from state i to state j in one step.
From Lecture notes,
$(M^n)_{ij}$ denotes probability of going from state i to state j in n step. Also, the stationary distribution for this matrix is going to be unique.
Since, it is probability of returning to the same state as the initial state, change under observation is from i to i.
Now, the probability that state didn't change will be,

$$\sum P(X_n = i | X_0 = i)$$

$$\implies \sum (M^n)_{ii}$$

the above sum is just sum of diagonals of matrix M^n which is $trace(M^n) = \sum eigenvalue(M^n)$
Eigenvalues of M^n is $eigenvalue(M)^n$. $\implies eigenvalues(M^n) = \{\lambda_1^n, ..., \lambda_k^n\}$

Probability of returning to the same state as the initial state, say $P_n(S)$

$$P_n(S) = \sum \lambda_i^n$$

Where λ_i is family of eigenvalues of M, as defined before.

Analysis at infinity,

$$\lim_{n\to\infty} P_n(S)$$

We know that $\forall i$, $\lambda_i \leq 1$ and $\exists i$, such that $\lambda_i = 1$

$$\lim_{n\to\infty} \lambda_i^n = \begin{cases} 0 & \lambda_i < 1 \\ 1 & \lambda_i = 1 \text{ (only one of this type)} \end{cases}$$

Now from these results,

$$\implies \lim_{n\to\infty} P_n(S) = \sum \lambda_i^n = 1$$

With almost surety after enough iteration the chain stops at stationary state.

□