CS203A: Assignment 2 (Spring 2022)

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

Instructions:

- You should write the solutions on your own. Dishonest behaviour and cheating in the assignment will be penalized with extreme measures. See: `https://www.cse.iitk.ac.in/pages/AntiCheatingPolicy.html`
- You should submit a single **.zip** file (with the name **[Your roll number].zip**) containing the following files to hello.iitk before the submission deadline:
  1. Your LaTeX-ed main answer script compiled as a **.pdf** file with the name **[Your roll number].pdf**. Using LaTeX is compulsory.
  2. Your LaTeX code for the main answer script as **.tex** file with the name **[Your roll number].tex**.
  3. The program file for part (d) of Question 3 with the name **[Your roll number]_Q3** (with extension **.py, .c** etc according to your choice of the programming language)
  4. A **readme** document explaining how the program file can be executed.

---

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) $Y_1, \ldots, Y_n$ are RVs such that $Y_i = 0$ if i^{th} student gets set 1, and $Y_i = 1$ if he gets set 2. Define $Z_1, \ldots, Z_m$ such that $Z_j = 1$ if j^{th} friendship connection is disrupted, else $Z_j = 0$. We need $E[Z]$, $Z = \sum_{i=1}^{m} Z_i$. $E[Z] = \sum_{i=1}^{m} E[Z_i].Z_i$ connects a and b, we have $E[Z_i] = Pr[Z_i = 1] = Pr[Y_a = 0 \& Y_b = 1] + Pr[Y_a = 1 \& Y_b = 0]$ Using pairwise independence of Y_i's, we have $Pr[Z_i = 1] = 1/2$.

   (b) Since we only need pairwise independence in Y_i, we can have that using $\log(n)$ random bits, as done in class. Going over all $2^{\log(n)}$ possibilities gives the poly(n) time algorithm. Needs to be explained and analyzed formally in the solution.

   □

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 $1 - 2e^{-\frac{(1+k\delta)^2}{2(2k+1-k\delta)}}$ 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) $E[\bar{X}] = \frac{E[\sum_{i=1}^{k} X_i]}{k} = E[X_i] = \sum_{j=1}^{N} \frac{i}{n} = \frac{N+1}{2}$. Therefore, $\tilde{N} = 2\mu - 1$

(b) δ error in N means $\delta/2$ error in $\bar{X}$. So we need $Pr[|\bar{X} - E[\bar{X}]| > \frac{\delta \bar{X}}{2}]$. $\bar{X} = \frac{\sum_{i=1}^{k} X_i}{k}$, where each X_i can have value from $[1, N]$. Use Hoeffding's inequality.

(c) δ error in N means $\delta/2$ error in $\bar{X}$. We have Error when more than $(k-1)/2$ elements on left of $(N+1)/2 - \delta N/2$ or right of $(N+1)/2 + \delta N$. Sum of k independent bernoulli variables with probability $1/2 - \delta/2$. Need probability of $> (k-1)/2$ successes. $E[X] = kE[X_i] = k(1/2 - \delta/2)$. $2 * Pr[X \geq (k+1)/2] = 2 * Pr[X - E[X] \geq 1/2 + k\delta/2] = 2 * Pr[X \geq E[X](1 + \frac{1+k\delta}{k(1-\delta)})]$. Use Chernoff. So $2 * Pr[X \geq E[X](1 + \frac{1+k\delta}{k(1-\delta)})] \leq 2e^{-\frac{(1+k\delta)^2}{2(2k+1-k\delta)}}$.

□

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) Notice that the rows of M sum to 1, so maximum eigenvalue is 1. It can be achieved by the uniform distribution. We prove uniqueness by contradiction. Let μ be a stationary distribution such that $\mu_j = \min_i \mu_i \neq \max_i \mu_i = \mu_k$. Then, using the positivity of $M_{j,k}$ the absolute value of the $(\mu^T M)_k$ is strictly smaller than μ_k, contradicting that μ is an eigenvector of M.

(b) Suppose the initial state is i. Then the probability of arriving at the i_{th} state at the n_{th} step is $(M^n)_{ii}$. Therefore using the law of total probability, the probability of arriving at the same state as the inital state is $\frac{1}{n}\sum_{i=1}^{n}(M^n)_{ii}$. Using the fact that the trace of a matrix equals the sum of its eigenvalues, this equals $\frac{1}{|S|}(\sum_{i=1}^{|S|}\lambda_i^n)$. Then, since from (a), $\lambda_1 = 1$ and $\lambda_i < 1$ for $i > 1$, the limit as $n \to \infty$ equals $\frac{1}{n}$.

□

4. (Optional) **DNF Counting**

Given a DNF formula F of n variables, how can we estimate the number of satisfying assignments without considering all cases? A natural approach is to randomly sample m assignments uniformly from the set of all possible assignments. Let X be the number of satisfying assignments among the m samples.

(a) Derive an unbiased estimate for the total number of satisfying assignments in terms of X, m, n.

(b) Construct one DNF formula each for number of variables n for $n = 6, 8, 10, 12$ having n clauses with each clause composed of subsets of $n - 4$ variables.

(c) Simulate the above algorithm on the constructed DNF formulas with $m = 4n$. Compare the estimate from the algorithm against the actual number satisfying assignments for each of the constructed formula. Plot the estimate and the number of satisfying assignments on the y-axis vs n on the x-axis..

(d) Explain qualitatively why the above scheme requires a large number of samples to produce accurate estimates.