SDSC3001 Assignment 1

2024.10.24
TA Lyuyi ZHU

Q1

1. (5 points) Prove that for any associate rule $X \to Y$ and $a \in X$, if we move the item a from X to Y, then the **confidence** of the new association rule $X - \{a\} \to Y \cup \{a\}$ is at most the confidence of $X \to Y$.

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

$$\text{Confidence}(X - \{a\} \to Y \cup \{a\}) = \frac{\text{Support}((X - \{a\}) \cup (Y \cup \{a\}))}{\text{Support}(X - \{a\})} \quad \leq \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

Q2

2. (15 points) Prove the Lower Tail of Chernoff bound (page 11 of Lec 2).

- X_i: Bernoulli random variable, $\Pr(X_i = 1) = p_i$
- $X = \sum_{i=1}^{n} X_i$, $E[X] = \mu = \sum_{i=1}^{n} p_i$

- Upper tail

$$\Pr(X \geq (1+\epsilon)\mu) \leq e^{-\frac{\epsilon^2 \mu}{2+\epsilon}}, \ \epsilon \geq 0$$

- Lower tail

$$\Pr(X \leq (1-\epsilon)\mu) \leq e^{-\frac{\epsilon^2 \mu}{2}}, \ 0 \leq \epsilon \leq 1$$

Q2

2. (15 points) Prove the Lower Tail of Chernoff bound (page 11 of Lec 2).

- Lower tail

$$\Pr(X \le (1-\epsilon)\mu) \le e^{-\frac{\epsilon^2\mu}{2}},\ 0 \le \epsilon \le 1$$

Solution:

By Markov inequality:

$$\begin{aligned}
P(X \le (1-\epsilon)\mu) &= P(-X \ge -(1-\epsilon)\mu) \\
&= P(e^{-sX} \ge e^{-s(1-\epsilon)\mu}) \\
&\le \frac{E(e^{-sX})}{e^{-s(1-\epsilon)\mu}},\ s > 0
\end{aligned}$$

Q2

Solution:

$$E(e^{-sX}) = E(e^{-s\sum X_i}) = E(\prod_{i=1}^{n} e^{-sX_i}) = \prod_{i=1}^{n} E(e^{-sX_i})$$

$$E(e^{-sX_i}) = p_i e^{-s} + (1-p_i) = p_i(e^{-s}-1)+1 \leq e^{p_i(e^{-s}-1)}$$

$$\Downarrow$$

$$E(e^{-sX}) \leq e^{np_i(e^{-s}-1)} = e^{\mu(e^{-s}-1)}$$

By Markov inequality:

$$\begin{aligned} P(X \leq (1-\epsilon)\mu) &= P(-X \geq -(1-\epsilon)\mu) \\ &= P(e^{-sX} \geq e^{-s(1-\epsilon)\mu}) \\ &\leq \frac{E(e^{-sX})}{e^{-s(1-\epsilon)\mu}}, \ s > 0 \end{aligned}$$

$$\Rightarrow \quad P(X \leq (1-\epsilon)\mu) \leq e^{\mu(e^{-s}-1+s(1-\epsilon))}, \ s > 0$$

Q2

Solution:

Now we have: $P(X \leq (1-\epsilon)\mu) \leq e^{\mu(e^{-s}-1+s(1-\epsilon))}, \ s > 0$

Denote $f(s) = e^{-s} - 1 + s(1-\epsilon)$, then we have:

$$f'(s) = -e^{-s} + 1 - \epsilon$$

Let $f'(s) = 0$. We have $s = -ln(1-\epsilon)$ and:

$$\begin{aligned}\min_s f(s) &= f(-ln(1-\epsilon)) \\ &= -\epsilon - (1-\epsilon)ln(1-\epsilon)\end{aligned}$$

Q3

3. (20 points) There is a playlist of n songs. A random music player randomly selects a song from the list to play (that is, each song is selected with probability $\frac{1}{n}$). Suppose that after listening to T songs played by the random music player, all the n songs have been played at least once. Prove that (1) $E\left[T\right] = nH_n$, where $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$ is the n-th Harmonic number; (2) $\Pr\left(\left|T - nH_n\right| \geq cn\right) \leq \frac{\pi^2}{6c^2}$. (Hint: $1 + \frac{1}{2^2} + \frac{1}{3^2} + \ldots. = \frac{\pi^2}{6}$)

Q3

Geometric distribution:

In Bernoulli trials, where the probability of event A occurring in each trial is denoted as p, the experiment stops when event A appears. The number of trials conducted until event A occurs is represented as X, and its probability mass function is given by:

$$P(X = k) = (1 - p)^{k-1} p, k = 1, 2, 3, \dots$$

The expected value and variance of X are as follows:

$$E(X) = \frac{1}{p}$$

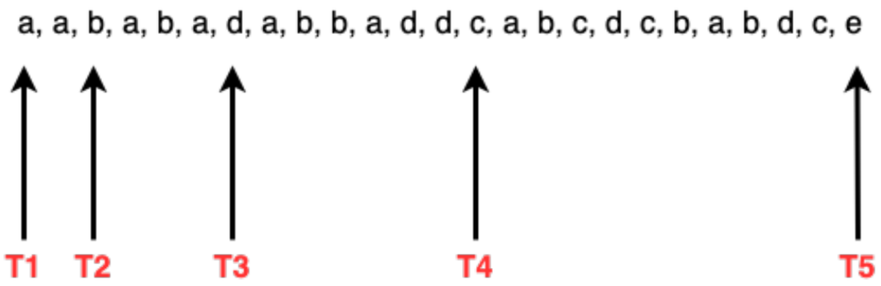
$$Var(X) = \frac{1-p}{p^2}$$

Q3

Solution:

(1) Decompose the process of listening to songs: suppose T_i is the first time to listen to i-th new songs.

An example: If we have 5 songs $\{$a, b, c, d, e$\}$.



Q3

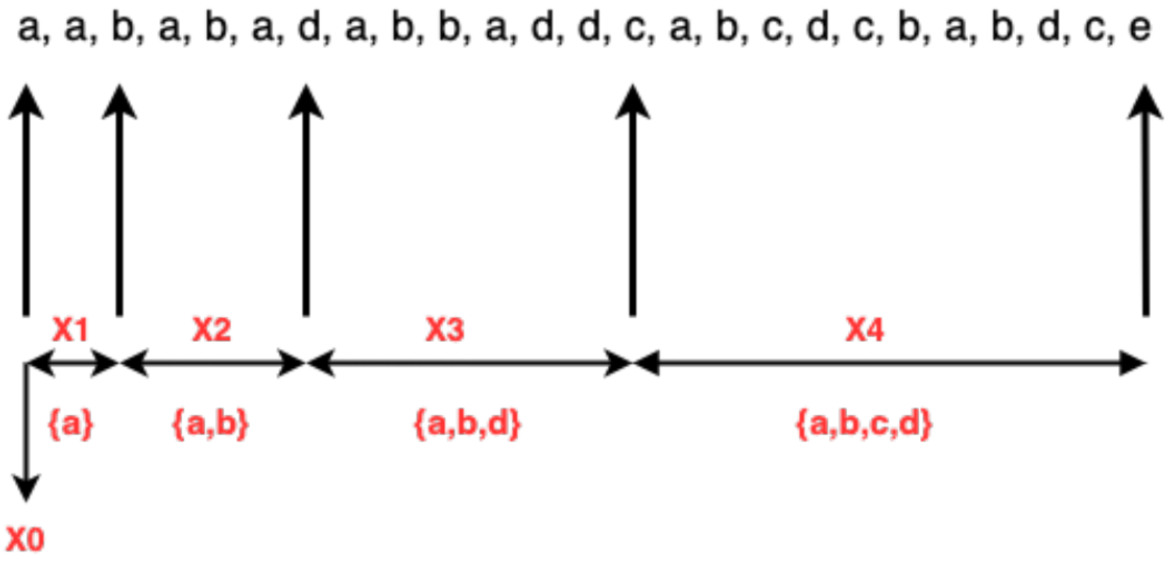
Solution:

(2) Rewrite T as follows:

$T = T_n = (T_n - T_{n-1}) + (T_{n-1} - T_{n-2}) + \ldots + (T_2 - T_1) + T_1$

After setting the $X_i = T_i - T_{i-1}$ and $X_0 = T_1$:

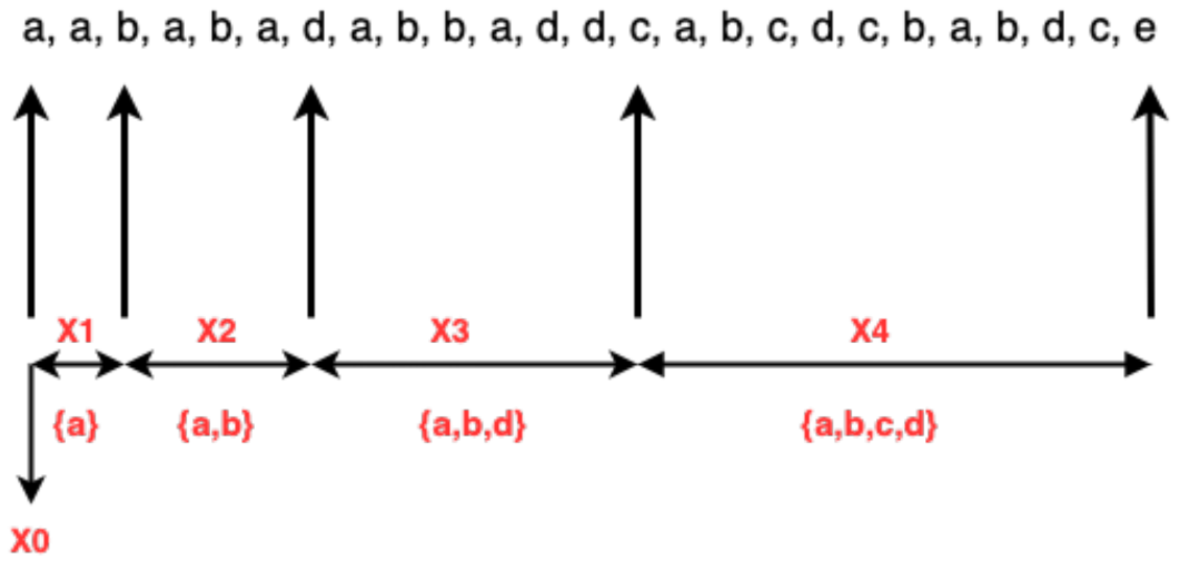
$$T = T_n = \sum_{i=0}^{n-1} X_i$$



Q3

Solution:

In time (T_i, T_{i+1}), we just repeat the previous i different old songs. And in the T_{i+1} time, we reach the $i+1$ new songs. Therefore, X_i follows a geometric distribution.



Q3

Solution:

In time (T_i, T_{i+1}), we just repeat the previous i different old songs. And in the T_{i+1} time, we reach the $i+1$ new songs. Therefore, X_i follows a geometric distribution.

$$X_i \sim Geo(1 - \tfrac{i}{n})$$

$$P(X_i = k) = (\tfrac{i}{n})^{k-1}(1 - \tfrac{i}{n})$$

$$E(X_i) = \tfrac{n}{n-i}$$

$$Var(X_i) = \tfrac{in}{(n-i)^2}$$

Q3

Solution:

$$X_i \sim Geo(1 - \frac{i}{n})$$

$$P(X_i = k) = (\frac{i}{n})^{k-1}(1 - \frac{i}{n})$$

$$E(X_i) = \frac{n}{n-i}$$

$$Var(X_i) = \frac{in}{(n-i)^2}$$

$$E(T) = \sum_{i=1}^{n-1} E(X_i) = nH_n$$

$$P(|T - nH_n| \geq cn) \leq \frac{Var(T)}{(cn)^2} = \frac{\sum_{i=1}^{n-1} Var(X_i)}{(cn)^2} \leq \frac{\pi^2}{6c^2}$$

Q4

4. (20 points) Let $\mathbf{P}$ be a $n \times n$ transition probability matrix, where $p_{ij} \geq 0$ denotes the probability of directly jumping from i to j, and $\sum_{j=1}^{n} p_{ij} = 1$ for each row i. Prove that the eigenvalues of $\mathbf{P}$ are within the range [-1,1].

Solution:

Assuming the eigenvector as $v = (v_1, \ldots, v_n)$, so $Pv = \lambda v$.
Moreover, we set $|v_k| = argmax(|v_1|, |v_2|, \ldots, |v_n|)$.

$$|\lambda v_k| = |\sum_{j=1}^{n} p_{kj} v_j| \leq \sum_{j=1}^{n} p_{kj} |v_k| = |v_k|$$

Therefore, $|\lambda| \leq 1$

Q5

5. (40 points) Download the file "com-dblp.txt" describing an undirected graph from "Files\Assignment 1" folder. Denote by d_v the degree of the node v, which is the number of neighbor nodes of v in the graph. Let $n_v = \frac{d_v}{D}$ be the normalized degree of v, where $D = \sum_v d_v$ is the sum of the degrees of all nodes. We simulate a random walk with M steps as follows: (1) the starting point is randomly selected (that is, each node is selected as the starting point with probability $\frac{1}{|V|}$ where V is the set of all nodes), and (2) at each step, randomly jump to a neighbor node of the current node. Let m_v be the number of times that v is visited in the random walk. Denote by $\mathbf{f} = \left(f_1, f_2, \ldots, f_{|V|}\right)$ the empirical frequency vector, where $f_v = \frac{m_v}{M}$. Similarly, let $\mathbf{n} = \left(n_1, n_2, \ldots, n_{|V|}\right)$ be the normalized degree vector. Write a program to simulate the above random walk and calculate the ℓ_1-distance (rounded to three decimal places) between $\mathbf{n}$ and $\mathbf{f}$ ($|\mathbf{n} - \mathbf{f}|_1 = \sum_v |n_v - f_v|$). Vary M and report the values of the ℓ_1-distance $|\mathbf{n} - \mathbf{f}|_1$ when $M = 10^7,\ 2 \times 10^7,\ 3 \times 10^7,\ 4 \times 10^7,\ 5 \times 10^7$. Briefly summarize your findings and make a guess of the relationship between $\mathbf{n}$ and $\mathbf{f}$.

Q4

- Stable distribution of random walk on graph

M	diff
10000000	0.1938
20000000	0.1367
30000000	0.1113
40000000	0.0965
50000000	0.0867