SDSC3001 - Assignment 1

Question 1

An association rule $X \rightarrow Y$ is a rule where X and Y are disjoint itemsets. The confidence of the rule $X \rightarrow Y$ is defined as:

$$\text{Conf}(X \rightarrow Y) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X)}$$

Then, moving an item $a \in X$ to the itemset Y forms a new rule $(X - \{a\}) \rightarrow (Y \cup \{a\})$.

The confidence of this new rule using the formula is:

$$\text{Conf}((X - \{a\}) \rightarrow (Y \cup \{a\})) = \frac{\text{Sup}((X - \{a\}) \cup (Y \cup \{a\}))}{\text{Sup}(X - \{a\})}$$

Considering the property of union operator, $(X - \{a\}) \cup (Y \cup \{a\}) = X \cup Y$;
therefore, the numerator in the confidence of both rules remains the same:

In the meantime, the denominator for $(X - \{a\}) \rightarrow (Y \cup \{a\})$ becomes $\text{Sup}(X - \{a\})$.

Rewriting the confidence of the new rule:

$$\text{Conf}((X - \{a\}) \rightarrow (Y \cup \{a\})) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X - \{a\})}$$

Since, $X - \{a\} \subseteq X$, every transaction that supports X also supports $X - \{a\}$; therefore, $\text{Sup}(X) \leq \text{Sup}(X - \{a\})$, which implies:

$$\frac{1}{\text{Sup}(X - \{a\})} \leq \frac{1}{\text{Sup}(X)}$$

Given:

- $\text{Conf}((X - \{a\}) \rightarrow (Y \cup \{a\})) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X - \{a\})}$
- $\text{Conf}(X \rightarrow Y) = \frac{\text{Sup}(X \cup Y)}{\text{Sup}(X)}$

This could be concluded:

$$\text{Conf}((X - \{a\}) \rightarrow (Y \cup \{a\})) \leq \text{Conf}(X \rightarrow Y)$$

Question 2

(coin tosses)

The Lower Tail of Chernoff Bound is expressed as:

$$\Pr(X \leq (1 - \epsilon)\mu) \leq \exp\left(-\frac{\epsilon^2 \mu}{2}\right), 0 < \epsilon < 1$$

- $X = \sum_{i=1}^{n} X_i$
- $\mu = \mathbb{E}[X]$

Using the fact that $ln(1 - x) \leq -x + \frac{x^2}{2}$ if $0 \leq x < 1$:

- $\Pr(X \leq (1 - \epsilon)\mu) = \Pr(e^{-sX} \geq e^{-s(1-\epsilon)\mu}) \leq \frac{\mathbb{E}[e^{-sX}]}{e^{-s(1-\epsilon)\mu}}$ from Markov's Inequality.
- $\mathbb{E}[e^{-sX}] = \mathbb{E}[e^{-s\sum_{i=1}^{n} X_i}] = \prod_{i=1}^{n} \mathbb{E}[e^{-sX_i}]$
- $\mathbb{E}[e^{-sX_i}] \leq 1 - s\mathbb{E}[X_i] + \frac{s^2}{2}\mathbb{E}[X_i^2] \leq e^{-s\mathbb{E}[X_i] + s^2/2}$

    - $e^{-sX_i} = 1 - sX_i + \frac{(sX_i)^2}{2!} + \cdots \le 1 - sX_i + \frac{s^2X_i^2}{2}$
- $\mathbb{E}[e^{-sX_i}] \le 1 - sp_i + \frac{s^2p_i}{2}$
- Using $\ln(1-x) \le -x + \frac{x^2}{2} \Rightarrow \ln(\mathbb{E}[e^{-sX_i}]) \le -sp_i + \frac{s^2p_i^2}{2}$
- Then, $\mathbb{E}[e^{-sX_i}] \le e^{-sp_i+\frac{s^2p_i}{2}}$ by taking exponentials on both sides.
- Therefore for all i, $\mathbb{E}[e^{-sX}] \le \prod_{i=1}^{n} e^{-sp_i+\frac{s^2p_i}{2}} = e^{-s\sum_i p_i+\frac{s^2\sum_i p_i}{2}} = e^{-s\mu+\frac{s^2\mu}{2}}$
- Solving it for s, $\Pr(X \le (1-\epsilon)\mu) \le \frac{e^{-s\mu+\frac{s^2\mu}{2}}}{e^{-s(1-\epsilon)\mu}} = e^{s\epsilon\mu-\frac{s^2\mu}{2}}$
    - Choosing $s=\epsilon$, $e^{\epsilon^2\mu-\frac{\epsilon^2\mu}{2}} = e^{-\frac{\epsilon^2\mu}{2}}$, minimizes the expression

$\therefore \Pr(X \le (1-\epsilon)\mu) \le \exp\left(-\frac{\epsilon^2\mu}{2}\right)$

Question 3

This problem rewrote the classic coupon collector's problem in terms of the expected number of plays to hear every song at least once and the probability bound for the deviation of the number of plays from its expectation. Listening to n distinct songs can be considered as a coupon collection process.

First, the expectation of the number of songs played T can be expressed as $E[T] = E[T_1] + E[T_2] + \cdots + E[T_n]$. Since the probability of listening to a new song at each stage is $\frac{n-i+1}{n}$, the expected number of plays until a new song is chosen is $\frac{n}{n-i+1}$, so $E[T_i] = \frac{n}{n-i+1}$. Therefore, the expected time to play every song at least once is $E[T] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i} = nH_n$.

Next, the probability bound for $|T - nH_n|$ can be shown using Chebyshev's inequality. The variance of T is $\text{Var}(T) = \sum_{i=1}^{n} \frac{n^2}{(n-i+1)^2} = n^2\sum_{k=1}^{n} \frac{1}{k^2}$ as the variance of a geometric random variable for T_i is given by $\text{Var}(T_i) = \frac{n^2}{(n-i+1)^2}$.
Since the series $\sum_{k=1}^{\infty} \frac{1}{k^2}$ converges to $\frac{\pi^2}{6}$ given as a hint, the variance of T is $\text{Var}(T) \le n^2 \cdot \frac{\pi^2}{6}$. By Chebyshev's inequality, the probability bound is:

$$\Pr\left(|T - E[T]| \ge cn\right) \le \frac{\text{Var}(T)}{(cn)^2}$$

$$\Pr\left(|T - nH_n| \ge cn\right) \le \frac{\pi^2}{6c^2}$$

where, $\frac{\text{Var}(T)}{(cn)^2} = \frac{n^2 \cdot \pi^2}{6} \cdot \frac{1}{c^2n^2}$

Question 4

To prove that the eigenvalues of the transition probability matrix $\mathbf{P}$ are within the range $[-1, 1]$, we can first show that the matrix $\mathbf{P}$ has an eigenvalue of 1.

The diagonal entries p_{ii} are non-negative and less than or equal to 1, and the sum of the off-diagonal entries in each row is $1 - p_{ii}$, since each row sums to 1 by the definition of the matrix $\mathbf{P}$. In other words, $\sum_{j=1}^{n} p_{ij} = 1, \text{ for all } i = 1, 2, \ldots, n.$

This implies that the vector $\mathbf{1} = [1, 1, \ldots, 1]^T$ (the all-ones vector) is a right eigenvector of $\mathbf{P}$ corresponding to the eigenvalue $\lambda = 1$. Thus, $\lambda = 1$ is an eigenvalue of $\mathbf{P}$, since $\lambda = 1$, $A - I$ is a singular matrix, and the $(A - I)$ is not full rank.

Also, the matrix $\mathbf{P}$, being a non-negative matrix with row sums equal to 1, cannot possess eigenvalues whose absolute value exceeds 1 by the Perron-Frobenius theorem since the sum of the absolute values of the eigenvalues is equal to the sum of the diagonal entries of the matrix.