09/20/2024

Big Data HW 1

Grokhulnath Thirumaran
675086474

1) i) Code has been uploaded on the gradescope.

2) After a spark session is initiated to read a text file, the extract_friends function structures these relationships into direct & mutual connections. Which are then combined using the create_map function. The get_sort_recommendation function aggregates and ranks the recommendation for each user.

3)

9550	9554, 9533, 9544, 9558, 153, 1220, 1421, 1436, 1951, 2413
8997	8998, 8987, 8992, 9001, 9003, 9009, 4849, 7174, 7279, 7364
4985	79, 577, 4839, 4984, 4986, 4987, 4988, 4989, 4990, 4991
4049	4871, 4875, 4889, 8492, 8685, 439, 660, 1100, 1137, 1156
2791	21185, 8783, 13280, 18359, 18363, 23667, 35740, 2204, 2786, 5996
3151	3161, 43162, 3230, 3450, 8692, 161, 2036, 3136, 3137, 3162
1724	1711, 1663, 1712, 1718, 1662, 1697, 1700, 1715, 1716, 1658
5060	5052, 5057, 5086, 14271, 98, 364, 575, 596, 611, 622
11	27552, 7785, 27573, 27574, 27589, 27590, 27600, 27617, 27620, 27667
8961	12241, 8973, 8965, 8963, 8966, 8967, 7174, 8969, 12243, 7177
739	732, 367, 381, 336, 21526, 28064, 677, 704, 728, 736

2) a)

$$Conf(A \rightarrow B) = Pr(B|A)$$

$$= \frac{Pr(A \cap B)}{Pr(A)}$$

Consider.

Basket 1 → A, B, E, F

Basket 2 → A, B, E

Basket 3 → A, B

Basket 4 → A, D

Basket 5 → B, C, E, F

Basket 6 → B, C, E

Basket 7 → B, C

Basket 8 → B, C, D

So, $Pr(A) = 4/8 = 0.5 \rightarrow (1)$

$Pr(A \cap B) = 3/8 = 0.375 \rightarrow (2)$

using (1) & (2)

$$Conf(A \rightarrow B) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{0.375}{0.5} = \underline{\underline{0.75}}$$

The confidence might misrepresent the importance of an association, which is a major drawback. It concentrates on the popularity of A rather on B. If B is as popular as A, there is a high probability of basket containing A and also B, Thus inflating confidence measure. From the data above

$$Pr(B) = 7/8 = 0.875$$

That infers that B appears in basket very often.

Lift measures how much "A & B occure together." than what if A & B are statically independent.

If lift > 1, Then B is more likely to be in the basket if A is there.

If lift < 1, Then B is unlikely to be in the basket if A is there.

$$\text{lift}(A \to B) = \frac{\text{conf}(A \to B)}{S(B)} = \frac{0.75}{0.875} < 1$$

Conviction compares the Pr of A if B is not there and are independent, with actual freq. of the appearance of A without B.

High conviction value ⇒ highly dependent on antecedent

Conviction = 1 ⇒ items are independent.

$$Conv(A \to B) = \frac{1 - S(B)}{1 - Conf(A \to B)} = \frac{1 - 0.875}{1 - 0.75}$$

$$= 0.5$$

So, B is not highly depending on antecedent.

Hence lift and conviction do not suffer from the draw back.

---

b)

Basket 1 → A B C F

Basket 2 → A B C

Basket 3 → A B

Basket 4 → A D

Basket 5 → B C E F

Basket 6 → B C E

Basket 7 → B C

Basket 8 → B C D

From the table, confidence is not symmetric

$$Conf(A \to B) = Pr(B/A)$$

$$= \frac{Pr(A \cap B)}{Pr(A)} = \frac{0.375}{0.5} = 0.75 \quad (1)$$

$$\text{conf}(B \to A) = \frac{P_r(A \cap B)}{P_r(B)}$$

$$= \frac{0.355}{0.825} = 0.428 \quad - (2)$$

We can infer from (1) & (2) that the confidence is not symmetric

To prove:

Lift is symmetric

$\Rightarrow$

$$\text{Lift}(A \to B) = \frac{\text{conf}(A \to B)}{S(B)}$$

$$= \frac{P_r(B/A)}{P_r(B)}$$

$$= \frac{\frac{P_r(A \cap B)}{P_r(A)}}{P_r(B)}$$

$$= \frac{\frac{P_r(A \cap B)}{P_r(B)}}{P_r(A)}$$

$$= \frac{P_r(A/B)}{P_r(A)} = \text{Lift}(B \to A).$$

Thus proved.

To prove: Conviction is not symmetric.

⇒

$$\text{Conv}(A \to B) = \frac{1 - S(B)}{1 - \text{conf}(A \to B)}$$

$$= \frac{1 - 0.875}{1 - 0.75} = \frac{0.125}{0.25} = 0.5 \quad -(1)$$

$$\text{Conv}(B \to A) = \frac{1 - S(A)}{1 - \text{Conf}(B \to A)}$$

$$= \frac{1 - 0.5}{1 - 0.428} = \frac{0.5}{0.571} = 0.875 \to (2)$$

$$(1) \neq (2)$$

Thus proved.

---

c)

Basket 1 A B

Basket 2 A B

Basket 3 C D

Basket 4 C F

$$\text{conf}(A \to B) = Pr(B|A)$$

$$= \frac{Pr(A \cap B)}{Pr(A)} = 1$$

$$Conv(A \to B) = \frac{1 - Pr(B)}{1 - conf(A \to B)}$$

$$= \frac{1 - 0.5}{1 - 1} = \infty$$

~~Con~~

$$Lift(A \to B) = \frac{conf(A \to B)}{Pr(B)}$$

$$= \frac{1}{0.5} = 2$$

$$Lift(C \to D) = \frac{conf(C \to D)}{Pr(D)}$$

$$= \frac{1}{0.25} = 4$$

Thus $A \to B$, $C \to D$ are 100% of the time, they have different Lift value.

d)

```
Number of frequent item sets: 553
DAI93865 -> FRO40251 = 1.0000000000
GRO85051 -> FRO40251 = 0.9991762768
DAI88079 -> FRO40251 = 0.9867256637
FRO92469 -> FRO40251 = 0.9835100118
DAI43868 -> SNA82528 = 0.9729729730
```

e)


```
DAI23334, ELE92920 -> DAI62779 = 1.0000000000
DAI55911, GRO85051 -> FRO40251 = 1.0000000000
DAI75645, GRO85051 -> FRO40251 = 1.0000000000
ELE17451, GRO85051 -> FRO40251 = 1.0000000000
ELE20847, FRO92469 -> FRO40251 = 1.0000000000
```

---

3) a) no. of columns with m 1's out of n

$$\Rightarrow \binom{n}{m} = \frac{n!}{m!\,(n-m)!}$$

no. of col. having no 1's for selected k rows

$$\Rightarrow \binom{n-k}{m} = \frac{(n-k)!}{m!\,(n-k-m)!}$$

$$P_r(\text{don't know}) = \frac{\binom{n-k}{m}}{\binom{n}{m}} = \frac{(n-k)!\cdot m!\,(n-m)!}{n!\,m!\,(n-k-m)!}$$

$$= \frac{(n-k)!\,(n-m)!}{n!\,(n-k-m)!}$$

$P_r(\text{don't know}) \leq \left(\frac{n-k}{n}\right)^m$, so probability is at most $\left(\frac{n-k}{n}\right)^m$

3 b) To prove:

$$\left(\frac{n-k}{n}\right)^m \leq e^{-10}$$

Inference:-

$$\left(\frac{n-k}{n}\right)^m \leq e^{-10}$$

$$\Rightarrow \left(1-\frac{k}{n}\right)^m \leq e^{-10}$$

$$\Rightarrow \left[\left(1-\frac{k}{n}\right)^{n/k}\right]^{\frac{mk}{n}} \leq e^{-10}$$

$$\Rightarrow e^{\frac{-mk}{n}} \leq e^{-10}$$

$$\Rightarrow -\frac{mk}{n} \leq -10$$

$$\Rightarrow k \geq \frac{10n}{m}.$$ So, the smallest k is $\frac{10n}{m}$

c)

π_1	π_2	π_3
1	3	2
2	1	3
3	2	1

Cyclic Permutations

S_1	S_2
0	0
1	1
0	1

input matrix

S_1	S_2
2	2
1	1
3	1

Signature matrix

Jaccard Similarity between S_1, S_2 is $\frac{1}{2}$,

But for that the minhash value agree $= \frac{2}{3}$

4) a) (10)

Given:

$G = H^k$ is $(\lambda, c\lambda, p_1^k, p_2^k)$ sensitive.

$\Rightarrow$

So each $1 \le j \le L$ and $x \in T$ the $Pr[x \in T \cap w_j] \le p_2^k$

[as $k = \log_{1/p_2} n$ when $p_2^k = 1/n$]

Therefore $Pr[x \in T \cap w_j] \le 1/n$

$\Rightarrow \mathbb{E}[|x \in T \cap w_j|] \le 1.$

By $\mathbb{E}[\sum_{j=1}^{L} |T \cap w_j|] \le L$. (linearity of expectation)

By markov's inequality,

$$Pr[X \ge a] \le \frac{\mathbb{E}[X]}{a}.$$

So. $$Pr\left[\sum_{j=1}^{L} |T \cap w_j|\right] \le \frac{\mathbb{E}[\sum_{j=1}^{L} |T \cap w_j|]}{3L} < \frac{L}{3L} = \frac{1}{3}$$

b) Given

G_j is $(\lambda, c\lambda, p_1^k, p_2^k)$ sensitive for $1 \le j \le L$

$$P_r[g_j(x^*) = g_j(z)] \ge p_1^k \quad -(1)$$

By defn,

$$L = u^P \quad -(2)$$

$$P = \frac{\log 1/p_1}{\log 1/p_2} \quad -(3)$$

$$k = \log_{1/p_2} u \quad -(4)$$

from (4)

$$k = \log_{1/p_2} u \Rightarrow \log 1/p_2 = \frac{\log u}{k} \quad -(5)$$

from (3) subs in (1)

$$P = \frac{\log 1/p_1}{\log 1/p_2} = \frac{k \log 1/p_1}{k \log u} = \frac{k \log 1/p_1}{\log u}$$

$$\Rightarrow P \log u = k \log 1/p_1 \quad -(6)$$

$$\Rightarrow u^P = (1/p_1)^k$$

(6) in (2)

$$L = n^P = \left(\frac{1}{p_1}\right)^k \Rightarrow p_1^k = 1/4 \quad -(7)$$

From (3)

$$P_r\left[g_i(x^*) \neq g_i(2)\right] \leq 1 - p_1^t$$

$$= 1 - 1/2$$

Thus $P_r\left[\forall 1 \leq j \leq L, g_i(x^*) \neq g_i(2)\right] \leq (1 - 1/4)^L < \frac{1}{e}$

c) U be the set of (c, λ)-ANN,

Thus, $U = \{x \in A : d(x,2) \leq c\lambda\}$

(1) ⇒ None of (c, λ) ANN points are hashed in to the same bucket with 2, that means $\forall j$ $1 \leq j \leq L$, $\omega_j \cap U = \phi$.

$$P_r[E_1] \leq P_r\left[x^* \notin \bigcup_{j=1}^{L} \omega_j\right] \cdot P_r\left[\forall 1 \leq j \leq L, g_i(x^*) \neq g_i(2)\right] < \frac{1}{e}$$

(2) ⇒ Atleast one (c, λ) ANN point is x^* be hashed, but are more than 3L points at distance greater than $c\lambda$ in the union bucket. If are less than 3L points, the probability less than 1/3

$$p' = Pr[E_1 \cup E_2] \leq Pr[E_1] + Pr[E_2] \leq \frac{1}{e} + \frac{1}{3}.$$

Thus $p > 1 - \frac{1}{e} - \frac{1}{3}$

d) LSH => 0.406 second per Query

Linear Search took average 0.503 second per query

Error decreases and increase at 16 for L

Error increases for k > 5: