CSE 547: Machine Learning for Big Data
Homework 1

Academic Integrity We take academic integrity extremely seriously. We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions and the code independently. In addition, each student should write down the set of people whom they interacted with.

Discussion Group (People with whom you discussed ideas used in your answers):
Kaitlyn Ng

On-line or hardcopy documents used as part of your answers:

I acknowledge and accept the Academic Integrity clause.

*(Signed)*_____Ekin Ugurel

Answer to Question 1

I first read and parsed the data through the text file using the `textFile` function, processing the friendships into a suitable data structure (using the `map` function). Using `flatMap` and `itertools`, I then identified pairs of direct and mutual friends for each user and combined them into a single list. Then, using `reduceByKey`, I calculated the number of mutual friends for each user pair before grouping the results (with `groupByKey`) by user and sorting the recommendations by descending number of mutual friends (in the case of a tie, I put the smaller user ID first). Finally, I returned the top 10 recommendations for each user.
Recommendations:

- **User 924**: 439, 2409, 6995, 11860, 15416, 43748, 45881
- **User 8941**: 8940, 8943, 8944
- **User 8942**: 8939, 8940, 8943, 8944
- **User 9019**: 317, 9022, 9023
- **User 9020**: 317, 9016, 9017, 9021, 9022, 9023
- **User 9021**: 317, 9016, 9017, 9020, 9022, 9023
- **User 9022**: 317, 9016, 9017, 9019, 9020, 9021, 9023
- **User 9990**: 13134, 13478, 13877, 34299, 34485, 34642, 37941
- **User 9992**: 9987, 9989, 9991, 35667
- **User 9993**: 9991, 13134, 13478, 13877, 34299, 34485, 34642, 37941

Answer to Question 2(a)

This is a drawback in cases where B is not related to A. B may be a commonly occurring item irrespective of A. In this case, confidence may overestimate the significance of the association between B and A. Lift and conviction do not suffer from this problem as their formulation includes $S(B)$, which essentially estimates $Pr(B)$.

Answer to Question 2(b)

The confidence is NOT symmetrical, as the conditional probability is not symmetric by nature. Consider the case of Figure 6.1 from Leskovic et al. (2020):

> **Example 6.1:** In Fig. 6.1 are sets of words. Each set is a basket, and the words are items. We took these sets by Googling `cat dog` and taking snippets from the highest-ranked pages. Do not be concerned if a word appears twice in a basket, as baskets are sets, and in principle items can appear only once. Also, ignore capitalization.
>
> 1. {Cat, and, dog, bites}
> 2. {Yahoo, news, claims, a, cat, mated, with, a, dog, and, produced, viable, offspring}
> 3. {Cat, killer, likely, is, a, big, dog}
> 4. {Professional, free, advice, on, dog, training, puppy, training}
> 5. {Cat, and, kitten, training, and, behavior}
> 6. {Dog, &, Cat, provides, dog, training, in, Eugene, Oregon}
> 7. {"Dog, and, cat", is, a, slang, term, used, by, police, officers, for, a, male–female, relationship}
> 8. {Shop, for, your, show, dog, grooming, and, pet, supplies}
>
> Figure 6.1: Here are eight baskets, each consisting of items that are words

The confidence of $\{cat\} \to kitten$ is 1/6, whereas the confidence of $\{kitten\} \to cat$ is 1/1. Lift, on the other hand, is symmetrical:

$$\text{lift}(A \to B) = \frac{P(B|A)}{S(B)} = \frac{P(B \cap A)}{P(B)P(A)} = \frac{P(A|B)}{S(A)} = \text{lift}(B \to A)$$

Finally, conviction is not symmetrical as it is directional. Looking at another trivial example from Figure 6.1:

$$\text{conv}(\{cat, dog\} \to and) = \frac{1 - S(and)}{1 - \text{conf}(\{cat, dog\} \to and)} = \frac{1 - 6}{1 - (3/5)}$$

$$\text{conv}(and \to \{cat, dog\}) = \frac{1 - S(\{cat, dog\})}{1 - \text{conf}(and \to \{cat, dog\})} = \frac{1 - 5}{1 - (4/6)}$$

The two are clearly not equal, proving that conviction is not symmetrical.

Answer to Question 2(c)

Consider a case where $Pr(B|A) = 1$. In this case,

- $\text{conf}(A \rightarrow B) = 1$
- $\text{conv}(A \rightarrow B) \rightarrow \infty$

However, we cannot comment on $\text{lift}(A \rightarrow B)$ as it will depend on $S(B)$, which depends on how often B occurs in sets without A. Thus, confidence and conviction are *desirable* measures, while lift is not.

Answer to Question 2(d)

Top 5 pairs with the highest support:

1. ('DAI62779', 'ELE17451'): 1592
2. ('FRO40251', 'SNA80324'): 1412
3. ('DAI75645', 'FRO40251'): 1254
4. ('FRO40251', 'GRO85051'): 1213
5. ('DAI62779', 'GRO73461'): 1139

Top 5 rules with highest confidence:

1. 'DAI93865' → 'FRO40251': 1.0
2. 'GRO85051' → 'FRO40251': 0.999176276771005
3. 'GRO38636' → 'FRO40251' : 0.9906542056074766
4. 'ELE12951', → 'FRO40251': 0.9905660377358491
5. 'DAI88079', → 'FRO40251': 0.9867256637168141

Answer to Question 2(e)

Top 5 triples with highest support:

1. ('DAI75645', 'FRO40251', 'SNA80324'): 550
2. ('DAI62779', 'FRO40251', 'SNA80324'): 476
3. ('FRO40251', 'GRO85051', 'SNA80324'): 471
4. ('DAI62779', 'ELE92920', 'SNA18336'): 432
5. ('DAI62779', 'DAI75645', 'SNA80324'): 421

Top 5 rules with highest confidence:

1. {'DAI62779', 'DAI88079'} → 'FRO40251': 1.0
2. {'ELE17451', 'GRO85051'} → 'FRO40251': 1.0
3. {'ELE26917', 'GRO85051'} → 'FRO40251': 1.0
4. {'GRO85051', 'GRO38814'} → 'FRO40251': 1.0
5. {'GRO85051', 'GRO73461'} → 'FRO40251': 1.0

Answer to Question 3(a)

Since m denotes the number of rows with 1s, $\binom{n}{m}$ denotes the number of ways m rows of 1 can be chosen out of n rows. However, we care about the probability of no 1s sampled in the k selected rows, given by the combination $\binom{n-k}{m}$. The probability of getting k rows with all zeroes is thus $\binom{n-k}{m}$ / $\binom{n}{m}$. Using factorial notation

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

Expanding $(n-m)!$

$$(n-m)! = (n-m)(n-m-1)\cdots(n-m-k)(n-m-k-1)\cdots$$

but notice that $(n-k-m)! = (n-m-k)(n-m-k-1)\cdots$ cancels with the latter half of this such that we have

$$\frac{(n-k)!(n-m)(n-m-1)\cdots}{n!}$$

Similarly the factors of $n!$ starting from $(n-k)$ all cancel with $(n-k)!$ such that we simplify to

$$\frac{(n-m)(n-m-1)\cdots(n-m-k+1)}{n(n-1)\cdots(n-k+1)}$$

At this point, we would like to upper bound the above expression. To do this, we will show that $\frac{n-k}{n} \geq \frac{n-k-1}{n-1}$. Observe that

$$\begin{aligned}
\frac{n-k}{n} - \frac{n-k-1}{n-1} &\geq 0 \\
\frac{(n-k)(n-1) - n(n-k-1)}{n(n-1)} &\geq 0 \\
\frac{(n^2-nk-n+k) - (n^2-nk-n)}{n(n-1)} &\geq 0 \\
\frac{k}{n(n-1)} &\geq 0
\end{aligned}$$

For $n > 1$ and $k \geq 0$, both n and $n-1$ are positive therefore making the fraction non-negative. This proves that $\frac{n-k}{n} \geq \frac{n-k-1}{n-1}$ holds, and we can thus say

$$\frac{(n-m)(n-m-1)\cdots(n-m-k+1)}{n(n-1)\cdots(n-k+1)} \leq \left(\frac{n-k}{n}\right)^m$$

Answer to Question 3(b)

The hypothesis states that $Pr(h(S_m) = 0) \leq e^{-10} = (\frac{n-k}{n})^m$. In this case, some algebra will help us solve for k. Specifically,

$$\begin{aligned} e^{-10/m} &= \frac{n-k}{n} \\ -ne^{-10/m} + n &= k \\ n(1 - e^{-10/m}) &= k \\ n(1 - (1 - 10/m)) &\geq k \\ n(10/m) &\geq k \end{aligned}$$

Answer to Question 3(c)

Consider the two sets $S_1 = \{1, 0, 0\}^\top$ and $S_2 = \{1, 1, 0\}^\top$. In this case, $\text{Sim}(S_1, S_2) = 0.5$, as there are two unique elements and only one of them exist in both sets. However, out of the three possible permutations, the only one resulting in different minhash values would be if the middle row was chosen as the first permutation in the cycle. Thus, the probability that the minhash values $h(S_1)$ and $h(S_2)$ are equal is $2/3$.