cs246

Kuang Chen

January 2019

1 Question 1

Spark pipeline:
1. List all pair combinations of friends of user in the row. User of the row is the mutual friend of all these pairs.
2. Filter out all the pairs that are already friends with each other.
3. Sum up the occurrence of all the remaining pairs.
4. Sort the pairs by occurrence.

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

2 Question 2

(a)
If B is purchased in all of the baskets, $Pr(A \cup B)$ will be the same as $Pr(A)$ which results in the confidence having the same value of 1 regardless of what items A represent. Lift doesn't suffer from the drawback because $Support(B)$ is taken as the denominator and hence, the lift will be smaller if B appears in all of the basket. The same can be said for Conviction but now $Support(B)$ is taken as the numerator and conviction will get close to 0 if B appears often.

(b)
Confidence - We have established in the lecture notes that $conf(A \rightarrow B) = \frac{S(A \cap B)}{S(A)}$. While $S(A \cap B) = S(B \cap A)$, the denominator of the formula disallows the measure to be symmetrical for all values of S(A) and S(B).
Lift - From the formula, $lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)} = \frac{S(A \cap B)}{S(A)} * \frac{1}{S(B)} = \frac{S(A \cap B)}{(S(A) * S(B))}$. $lift(B \rightarrow A) = \frac{S(B \cap A)}{S(B)} * \frac{1}{S(A)} = \frac{S(A \cap B)}{(S(A) * S(B))} = lift(A \rightarrow B)$. Hence, this measure is symmetrical.
Conviction - $conv(A \rightarrow B)$ can be shown to be $\frac{[S(A) - S(B) * S(A)]}{[S(A) - S(BUA)]}$. While S(B)*S(A) and S(B U A) can be symmetrical, the other term 'S(A)' is unique to each input value and it will produce a different value if changed to S(B). Hence, the measure is not symmetrical.

(c)
The confidence measure has this property. To prove so, we can break down the formula into so: $conf(A \rightarrow B) = \frac{S(BUA)}{S(A)} = \frac{Pr(BUA)}{Pr(A)}$. In a perfect implication condition, item B will always exist in baskets where item(s) A exist. Hence, Pr(B U A) = Pr(A) = Pr(B). Using this equality statement, we can deduce that given a perfect implication of $A \rightarrow B, conf(A \rightarrow B) = \frac{Pr(A)}{Pr(A)} = 1$.

(d)
Top 5:
DAI93865 → FRO40251 1.0
GRO85051 → FRO40251 0.9992
ELE12951 → FRO40251 0.9907
GRO38636 → FRO40251 0.9906
DAI88079 → FRO40251 0.9867
(e)
Top 5:
(DAI23334,ELE92920) → DAI62779 1.0
(DAI31081,GRO85051) → FRO40251 1.0
(DAI55911,GRO85051) → FRO40251 1.0
(DAI62779,DAI88079) → FRO40251 1.0
(DAI75645,GRO85051) → FRO40251 1.0

3 Question 3

(a)
There are $\binom{n}{k}$ ways to choose k rows out of all n rows. There are $\binom{n-m}{k}$ ways to choose all 0's out of k rows. The probability of "don't know" is $\frac{\binom{n-m}{k}}{\binom{n}{k}}$. Expanding this equation, we get $\frac{(n-m)!}{k!*(n-m-k)!} * \frac{k!*(n-k)!}{n!}$. We can remove the k! and expand (n-k)! to $(n-k)(n-k-1)...(n-k-m+1)(n-k-m)!$. We can also expand n! to $(n)(n-1)...(n-m+1)(n-m)!$ Then, we can remove (n-k-m)! and (n-m)! from the equation. We are finally left with $\frac{(n-k)(n-k-1)...}{(n)(n-1)...}$. The numerator has m terms and is bounded by (n-k) while the denominator has m terms and is bounded by (n). Hence, we can prove that the probability is at most $(\frac{n-k}{n})^m$.

(b)
The probability of "don't know" can be simplified to $(1-\frac{k}{n})^m = (1-(\frac{k}{n})^{\frac{n}{k}})^{\frac{km}{n}}$. Now the inequality becomes $(1-(\frac{k}{n})^{\frac{n}{k}})^{\frac{km}{n}} \leq e^-10$. Since n is much larger than k, we can approximate $(1-(\frac{k}{n})^{\frac{n}{k}})$ to $\frac{1}{e}$. Hence, we have $\frac{mk}{n} \leq 10$. Expressing this in terms of k, we have $k \leq \frac{10n}{m}$.

(c)
The two columns are (1,1,0,0) and (1,0,1,0). The Jaccard similarity is $\frac{1}{4}$ but the probability that a random cyclic permutation yields the same minhash value for both S1 and S2 is $\frac{2}{4}$ since every row will differ except for the first and last row.

4 Question 4

(a)
$Pr(g_j(x) = g_j(z)|\forall 1 \leq j \leq L) leq (p_2)^k = \frac{1}{n}$. This means that the probability of having an point in T mapping to any same bucket as z is $\leq \frac{1}{n}$. Hence, $E[T \cap W_j] \leq \frac{1}{n} * n = 1$. By linearity of expectations, $E[\sum_{j=1}^{L} |T \cap W_j|] \leq L$. Then, by markov's inequality, we get $3L * Pr(\sum_{j=1}^{L} |T \cap W_j| \geq 3L) \leq L$ which can be reduced to $Pr(\sum_{j=1}^{L} |T \cap W_j| \geq 3L) \leq \frac{1}{3}$.

(b)
We know that $Pr(g_i(x^*) = g_i(z)) \geq P_1^k$ for any i. Hence, $Pr(g_i(x^*) \neq g_i(z)) \leq (1 - P_1^k)$ for any i. Thus, $Pr(\forall 1 \leq j \leq L, g_j(x^*) \neq g_j(z) < (1 - P_1^k)^L$. $(1 - P_1^k)$ can be expanded to $((1 - P_1^k)^{\frac{1}{P_1^k}})^(P_1^k)$ because the denominator of P_1^k is large. This can then be approximated to $(\frac{1}{e})^(L * P_1^k)$.
Next, to solve $(L * P_1^k)$, we try to express P_1^k in terms of L. $n^p = L \rightarrow \frac{log1/p_1}{log1/p_2} * logn = logL \rightarrow log(1/p_1)^{\frac{logn}{log1/p_2}} = logL$. Therefore, $p_1 = 1/L \rightarrow (L * P_1^k) = 1$. Thus, $Pr(\forall 1 \leq j \leq L, g_j(x^*) \neq g_j(z)) < \frac{1}{e}$.

(c)
There are two ways a reported point is not an actual $(c, \lambda) - ANN$. The false-positive case is if the point is an actual $(c, \lambda) - ANN$ but was not reported and this probability is $< \frac{1}{e}$ from (b). The false-negative case is if the point is not an actual $(c, \lambda) - ANN$ but was reported as it is and this probability is $\leq \frac{1}{3}$ from (a). Thus, the probability that a reported point is an actual $(c, \lambda) - ANN$ is simply $\geq 1 - \frac{1}{3} - \frac{1}{e}$.

(d)
Average search time for LSH: 0.367756605148 seconds
Average search time for Linear: 0.103164505959 seconds

It seems that error increases with greater values of K and decreases with greater values of L.

The images taken from the top 10 LSH has clearer defined features than the top 10 linear neighbors.

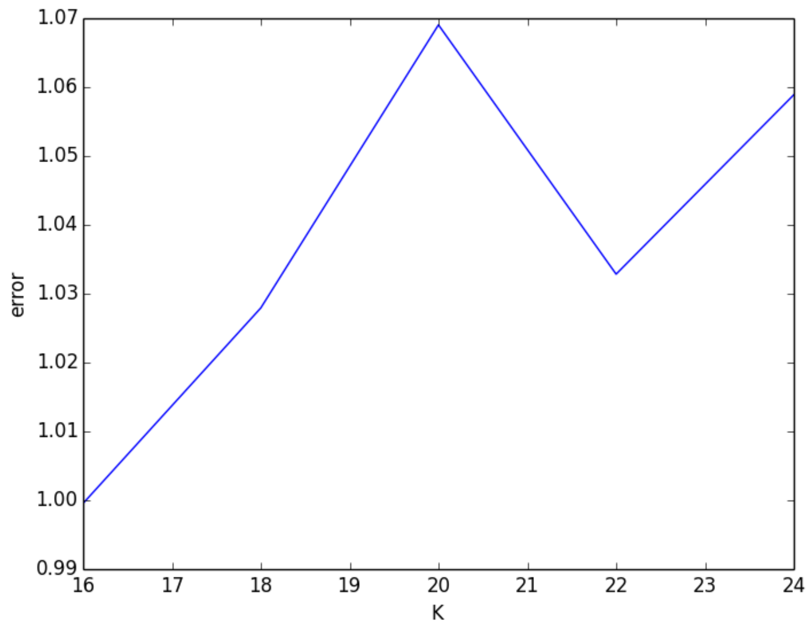

Figure 1: K to Error

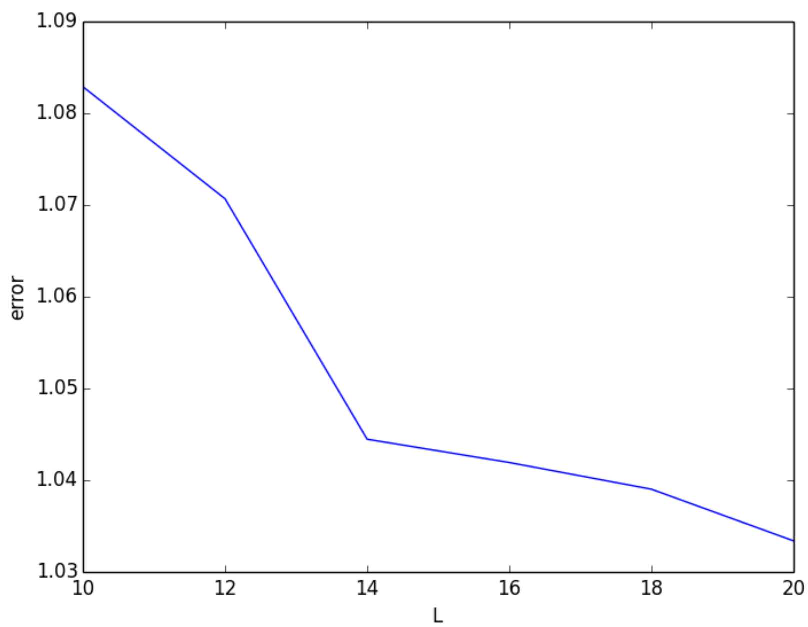

Figure 2: L to Error

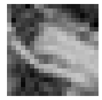

Figure 3: **original**

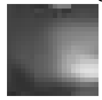

Figure 4: linear 1

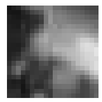

Figure 5: linear 2

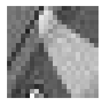

Figure 6: linear 3

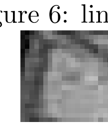

Figure 7: linear 4

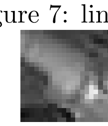

Figure 8: linear 5

Figure 9: linear 6

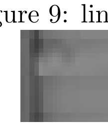

Figure 10: linear 7

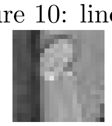

Figure 11: linear 8

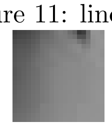

Figure 12: linear 9

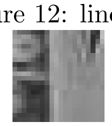

Figure 13: linear 10

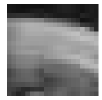

Figure 14: lsh 1

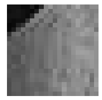

Figure 15: lsh 2

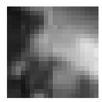

Figure 16: lsh 3

Figure 17: lsh 4

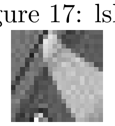

Figure 18: lsh 5

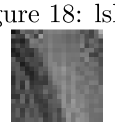

Figure 19: lsh 6

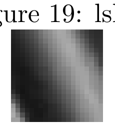

Figure 20: lsh 7

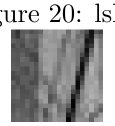

Figure 21: lsh 8

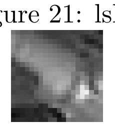

Figure 22: lsh 9

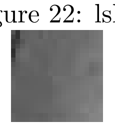

Figure 23: lsh 10