CS246: Mining Massive Data Sets

Assignment number: 1

Fill in and include this cover sheet with each of your assignments. It is an honor code violation to write down the wrong time. Assignments and code are due at 5:00 PM on Scoryst and SNAP respectively. Failure to include the coversheet with you assignment will be penalized by 2 points. Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Assignments are due on Thursdays, which means the first late period expires on the following Tuesday at 5:00 PM.) Once these late periods are exhausted, any assignments turned in late will be penalized 50% per late period. However, no assignment will be accepted more than one late period after its due date. (If an assignment is due to Thursday then we will not accept it after the following Thursday.)

Your name:Erli Zhou
Email: erlizhou@stanford.edu
SUNet ID:erlizhou

Collaborators:

I acknowledge and accept the Honor Code.

(Signed) Erli Zhou

Answer to Question 1

Algorithm:
Mapper: Read the input text file, then output the user ID as the intermediate key and his recommendation as the intermediate value.

For each user ID, output all of his immediate friends as the value with flag F, out one of his friend as the key and another as the value with flag R, also output all users who do not have a single friend as the value with flag N.

Reducer: Output the user ID as the key and his recommendations as the value.

For each key read in as user ID, if the user has no friends, output a blank line. If the value is flagged with F, update the map value to be -1 to indicate no recommendations. If the value is flagged with R, search whether the key exists in the map. If not, create a new key with value 1. If so, search whether the value equals to one. If not, increase the value by 1.

Recommendations for the users with following user IDs:
User ID: 924, Recommendations: 439, 2409, 6995, 11860, 15416, 43748, 45881
User ID: 8941, Recommendations: 8943, 8944, 8940
User ID: 8942, Recommendations: 8939, 8940, 8943, 8944
User ID: 9019, Recommendations: 9022, 317, 9023
User ID: 9020, Recommendations: 9021, 9016, 9017, 9022, 317, 9023
User ID: 9021, Recommendations: 9020, 9016, 9017, 9022, 317, 9023
User ID: 9022, Recommendations: 9019, 9020, 9021, 317, 9016, 9017, 9023
User ID: 9990, Recommendations: 13134, 13478, 13877, 34299, 34485, 34672, 37941
User ID: 9992, Recommendations: 9987, 9989, 35667, 9991
User ID: 9993, Recommendations: 9991, 13134, 13478, 13877, 34299, 34485, 34642, 37941

Answer to Question 2a

The drawback for ignoring $Pr(B)$ is that not all high-confidence rules are interesting. For example, the rule $X \rightarrow$ milk may have high confidence for many itemsets X, because milk is purchased very often and the confidence will be high. In this case, using Interest$(A \rightarrow B) = |\text{conf}(A \rightarrow B) - Pr(B)|$ is a good solution.

Lift does not suffer from this setback because $\text{lift}(A \rightarrow B) = \text{conf}(A \rightarrow B)/S(B)$. For example, if the occurrence of A does not impact B, $\text{lift}(A \rightarrow B) = 1$. If the occurrence of A increases the frequency of B, $\text{lift}(A \rightarrow B) > 1$. If the occurrence of A decreases the frequency of B, $\text{lift}(A \rightarrow B) < 1$.

Conviction does not suffer from this setback because $\text{conv}(A \rightarrow B) = (1 - S(B))/(1 - \text{conf}(A \rightarrow B))$. For example, if the occurrence of A does not impact B, $\text{conv}(A \rightarrow B) = 1$. If the occurrence of A increases the frequency of B, $\text{conv}(A \rightarrow B) > 1$. If the occurrence of A decreases the frequency of B, $\text{conv}(A \rightarrow B) < 1$.

Answer to Question 2b

Confidence is not symmetrical. $\text{conf}(A \rightarrow B) = Pr(B|A) = Pr(A \bigcap B)/Pr(A)$, while $\text{conf}(B \rightarrow A = Pr(A|B) = Pr(A \bigcap B)/Pr(B)$. Since $Pr(A)$ does not necessarily equal to $Pr(B)$, the measure is not symmetrical.

Lift is symmetrical. $\text{lift}(A \rightarrow B) = \text{conf}(A \rightarrow B)/S(B) = \text{support}(A \bigcup B)/(\text{support(A)}S(B)) = (N*\text{support}(A \bigcup B))/(\text{support(A)support(B)})$, while $\text{lift}(B \rightarrow A) = \text{conf}(B \rightarrow A)/S(A) = \text{support}(B \bigcup A)/(\text{support(B)}S(A)) = (N*\text{support}(A \bigcup B))/(\text{support(A)support(B)})$. Since $\text{lift}(A \rightarrow B) = \text{lift}(B \rightarrow A)$, the measure is symmetrical.

Conviction is not symmetrical. $\text{conv}(A \rightarrow B) = (1 - S(B))/(1 - \text{conf}(A \rightarrow B)) = (1 - Pr(B))/(1 - Pr(A \bigcap B)/Pr(A))$, while $\text{conv}(B \rightarrow A) = (1 - S(A))/(1 - \text{conf}(B \rightarrow A)) = (1 - Pr(A))/(1 - Pr(A \bigcap B)/Pr(B))$. Since $Pr(A)$ does not necessarily equal to $Pr(B)$, the measure is not symmetrical.

Answer to Question 2c

Confidence is desirable. Suppose $A \to B$ holds all the time, which means B occurs in every transaction in which A occurs. Then intuitively $\text{conf}(A \to B) = Pr(B|A) = Pr(A, B)/Pr(A) = 1$. Since $\text{conf}(A \to B)$ has a range of $[0, 1]$, the measure is desirable.

Lift is not desirable. Suppose $A \to B$ holds all the time. Then $\text{lift}(A \to B) = \text{conf}(A \to B)/S(B) = 1/S(B) = N/\text{support(B)}$. Suppose there exists another rule which always holds in the form of $C \to D$ and assume support(D)$\neq$ support(B).Then $\text{lift}(C \to D) = \text{conf}(C \to D)/S(D) = 1/S(D) = N/\text{support(D)}$. So $\text{lift}(A \to B) \neq \text{lift}(C \to D)$, which suggests the measure is not desirable.

Conviction is desirable. Suppose $A \to B$ holds all the time. Then $\text{conv}(A \to B) = (1 - S(B))/(1- \text{conf}(A \to B))$. Since $\text{conf}(A \to B) = 1$, $\text{conv}(A \to B) = \infty$. Since infinity is definitely maximal, the measure is desirable.

Answer to Question 2d

Top 5 confidence rules for itemsets of size two:
DAI93865 → FRO40251, confidence score of 1.0
GRO85051 → FRO40251, confidence score of 0.999176276771005
GRO38636 → FRO40251, confidence score of 0.9906542056074766
ELE12951 → FRO40251, confidence score of 0.9905660377358491
DAI88079 → FRO40251, confidence score of 0.9867256637168141

Answer to Question 2e

Top 5 confidence rules for itemsets of size three:
DAI23334, ELE92920 → DAI62779, confidence score of 1.0
DAI31081, GRO85051 → FRO40251, confidence score of 1.0
DAI55911, GRO85051 → FRO40251, confidence score of 1.0
DAI62779, DAI88079 → FRO40251, confidence score of 1.0
DAI75645, GRO85051 → FRO40251, confidence score of 1.0

Answer to Question 3a

Assume $x, y, z \in S, Pr[h(x) \neq h(z)] = Pr\{[h(x) \neq h(z)] \bigcap [h(x) = h(y)]\} + Pr\{[h(x) \neq h(y)] \bigcap [h(x) \neq h(y)]\} = Pr[h(y) \neq h(z)] + Pr\{[h(x) \neq h(y)] \bigcap [h(x) = h(z)]\} \leq Pr[h(y) \neq h(z)] + Pr[h(x) \neq h(y)]$
Which is the equivalent of 1 - $\text{sim}(x, z) \leq$ 1-$\text{sim}(y, z) + 1 - \text{sim}(x, y)$
Which leads to $d(x, z) \leq d(y, z) + d(x, y)$.

Answer to Question 3b

Let $A = \{1, 2, 3\}, B = \{4, 5, 6\}, C = \{3, 4, 5, 6\}$
Then $\text{sim}_{over}(A, B) = 0$, $\text{sim}_{over}(B, C) = 3/3 = 1$, $\text{sim}_{over}(A, C) = 1/3$
So $D(A, B) = 1 - 0 = 1, D(B, C) = 1 - 1 = 0, D(A, C) = 1 - 1/3 = 2/3$
$D(A, B) = 1 > 2/3 = D(B, C) + D(A, C)$

Answer to Question 3c

Let $A = \{1, 2, 3\}, B = \{4, 5, 6\}, C = \{3, 4, 5, 6\}$
Then $\text{sim}_{dice}(A, B) = 0$, $\text{sim}_{dice}(B, C) = 3/3.5 = 0.857$, $\text{sim}_{dice}(A, C) = 1/3.5 = 0.286$
So $D(A, B) = 1 - 0 = 1, D(B, C) = 1 - 0.857 = 0.14, D(A, C) = 1 - 0.286 = 0.71$
$D(A, B) = 1 > 0.85 = D(B, C) + D(A, C)$

Answer to Question 4a

Suppose $|T| = m \leq n$ and $x \in T$,
$Pr(g(x) = g(z)) = Pr[h(x) = h(z)]^k \leq {p_2}^k = {p_2}^{\log_{1/p_2} n} = {p_2}^{\log_{p_2} 1/n} = 1/n$
Thus,$E(T \bigcap W_j|) = mPr(g(x) = g(z)) \leq m/n \leq 1$
Since expectation is a linear function, $E|\Sigma_{j=1}^{L}(T \bigcap W_j)| \leq L$
Applying Markov's inequality, $Pr[\Sigma_{j=1}^{L}|T \bigcap W_j| \geq 3L] \leq E|\Sigma_{j=1}^{L}(T \bigcap W_j)|/3L \leq L/3L = 1/3$

Answer to Question 4b

Because $d(x^*, z) \leq \lambda, Pr[\forall 1 \leq j \leq L, g_j(x^*) = g_j(z)] \geq {p_1}^k$
$Pr[\forall 1 \leq j \leq L, g_j(x^*) \neq g_j(z)] = \{1 - Pr[\forall 1 \leq j \leq L, g_j(x^*) = g_j(z)]\} \leq 1 - {p_1}^k = (1 - p_1^{\log_{p_2} 1/n})^L = (1 - n^{-\rho})^L$
Since $1 - x \leq e^x, (1 - n^{-\rho})^L \leq e^{-L/n^\rho} = e^{-1} = 1/e$

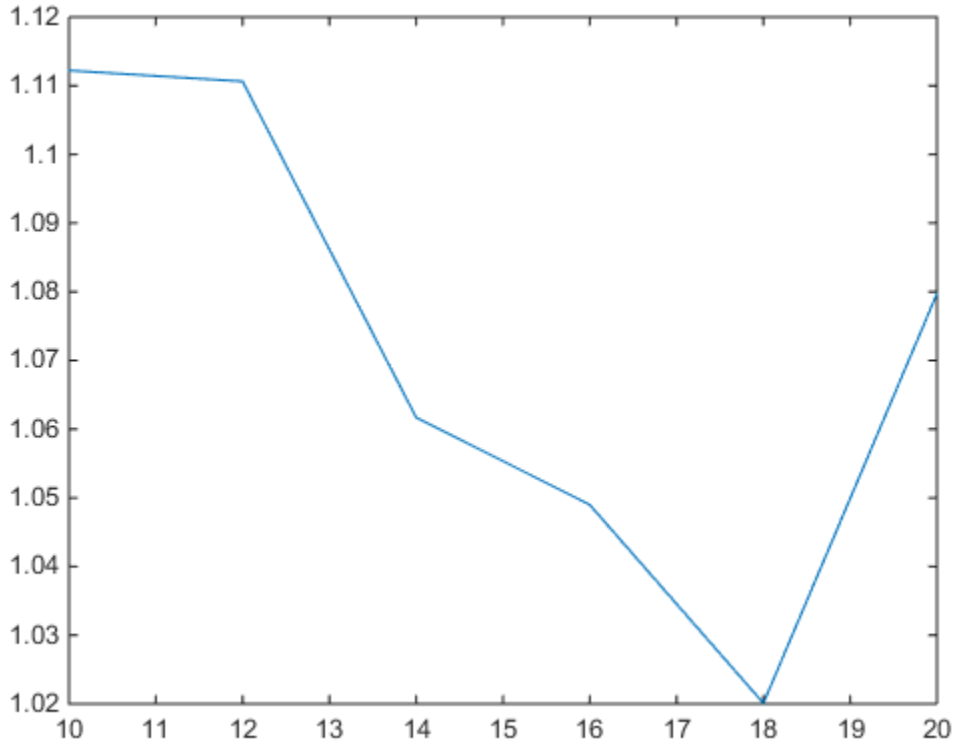
Answer to Question 4c

An error can occur if (1) there are more than 3L points in L buckets in $g_j(z)$ and all 3L sampled points are from set T
OR (2) a point $m \in A$ is not hashed to any of the L buckets in $g_j(z)$
Pr(error) = Pr(case 1) + Pr(case 2) = $Pr[\Sigma_{j=1}^{L}|T \bigcap W_j| \geq 3L] + Pr[\forall 1 \leq j \leq L, g_j(x^*) \neq g_j(z)] \leq 1/3 + 1/e$
So with probability greater than some fixed constant the reported point is an actual (c, λ)-ANN

Answer to Question 4d

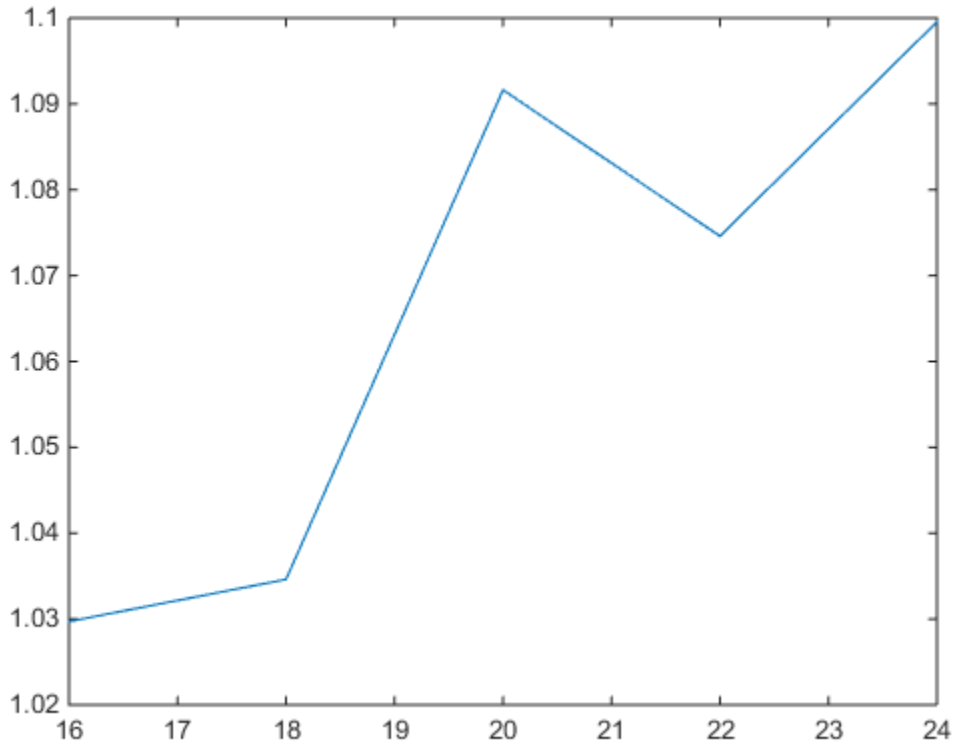
Average search time for LSH is 0.0123 seconds.
Average search time for linear search is 0.1211 seconds.

Error value for L = 10,12,14,..,20, with k = 24:



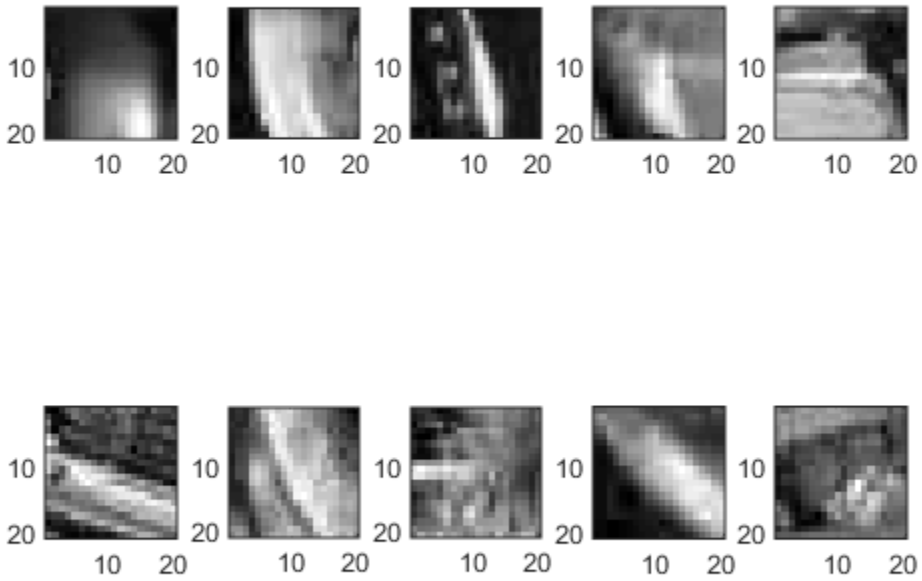
Given a set number of buckets, error generally decreases when the number of hash tables increases.

Error value for k = 16,18,20,22,24, with L = 10:

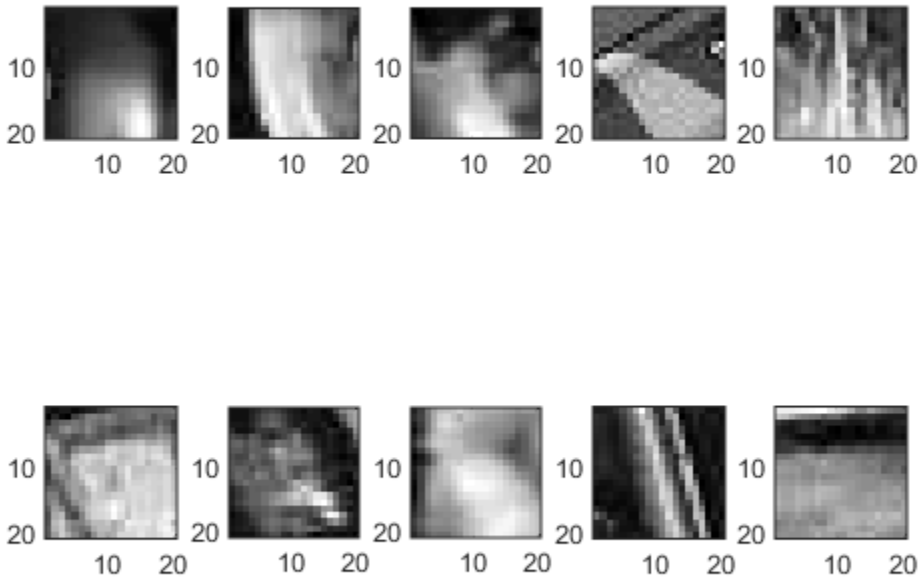


Given a set number of hash tables, error generally increases when the number of buckets increases.

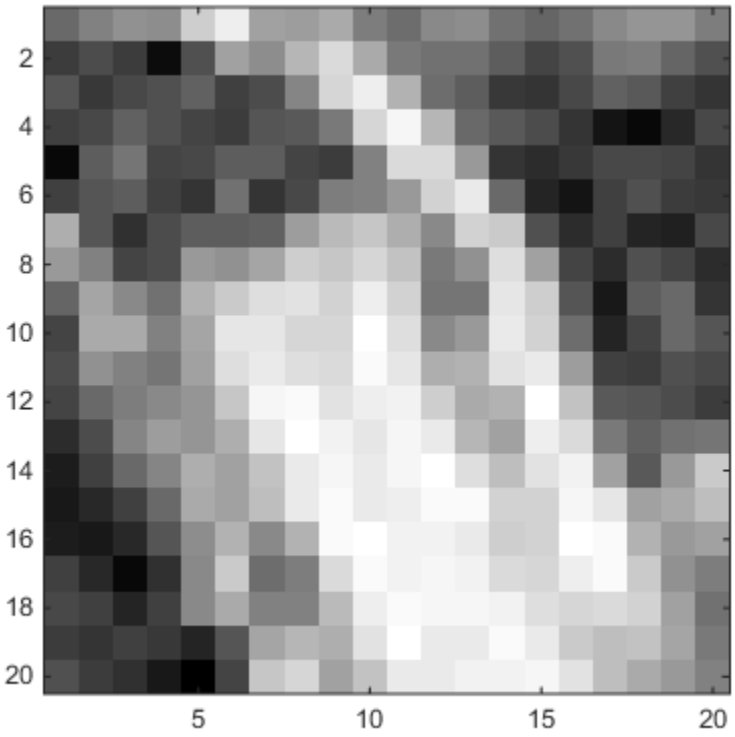
Top 10 nearest neighbours found by LSH method:



Top 10 nearest neighbours found by linear search method:



The image patch itself:



They look rather similar visually.