Philip Scuderi
Stanford Univeristy
Winter 2013
CS 246
Homework 1

1

The recommendations for the users specified in the homework follow. The output is in the format <User><Tab><Recommendations in decreasing order of number of friends>.

```
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(a)

An example of when the confidence formula ignoring P(B) being a drawback is the following:

If P(B) is very large then rules can be generated with high confidence but that do not tell us much. Suppose event B is the event that a person buys milk and everyone buys milk. Further suppose that event A is the event that a person buys some other item. In this case, we would have high confidence that B occurs given that A occurs. As you can see, this tells us almost nothing about consumer behavior that we did not already know.

Lift does not suffer the drawback that confidence does because lift does not ignore P(B). The problem states the following:

$$S(B) = \frac{Support(B)}{N}, where\ N\ is\ the\ total\ number\ of\ transactions$$

Support(B) is the number of transactions including B. The number of transactions including B divided by the total number of transactions is equivalent to the probability of including B in a transaction, or P(B). In other words:

$$S(B) = \frac{Support(B)}{N} = P(B)$$

The formula for lift is therefore equivalent to the following:

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)} = \frac{conf(A \rightarrow B)}{P(B)} = \frac{P(B|A)}{P(B)} = \frac{P(B \cap A)}{P(B)P(A)}$$

As you can see above, the definition of lift includes P(B) in its denominator. Thus, lift does not suffer the drawback of ignoring P(B).

Conviction does not suffer the drawback that confidence does because lift does not ignore P(B). As addressed above, S(B) = P(B). It follows that:

$$conv(A \rightarrow B) = \frac{1 - S(B)}{1 - conf(A \rightarrow B)} = \frac{1 - P(B)}{1 - P(B|A)} = \frac{1 - P(B)}{1 - \frac{P(B \cap A)}{P(A)}}$$

As you can see above, the definition of conviction includes P(B) in its numerator. Thus, lift does not suffer the drawback of ignoring P(B)

<u>2(b)</u>

No, all the measures here are not symmetrical.

Confidence is <u>not</u> symmetrical, as you can see below.

$$conf(A \rightarrow B) = P(B|A) = \frac{P(B \cap A)}{P(A)}$$

$$conf(B \rightarrow A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\frac{P(B \cap A)}{P(A)} \neq \frac{P(A \cap B)}{P(B)}$$

$$conf(A \rightarrow B) \neq conf(B \rightarrow A)$$

Lift <u>is</u> symmetrical, as you can see below.

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)} = \frac{conf(A \rightarrow B)}{P(B)} = \frac{P(B|A)}{P(B)} = \frac{P(B \cap A)}{P(B)P(A)}$$

$$lift(B \rightarrow A) = \frac{conf(B \rightarrow A)}{S(A)} = \frac{conf(B \rightarrow A)}{P(A)} = \frac{P(A|B)}{P(A)} = \frac{P(A \cap B)}{P(A)P(B)}$$

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

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

$$lift(A \rightarrow B) = lift(B \rightarrow A)$$

Conviction is not symmetrical, as you can see below.

$$conv(A \rightarrow B) = \frac{1 - S(B)}{1 - conf(A \rightarrow B)} = \frac{1 - P(B)}{1 - P(B|A)}$$

$$conv(B \rightarrow A) = \frac{1 - S(A)}{1 - conf(B \rightarrow A)} = \frac{1 - P(A)}{1 - P(A|B)}$$

$$\frac{1 - P(B)}{1 - P(B|A)} \neq \frac{1 - P(A)}{1 - P(A|B)}$$

$$conv(A \rightarrow B) \neq conv(B \rightarrow A)$$

2(c)

Confidence is a perfect implication. Lift and conviction are not perfect implications. See below for details.

We assume that $(A \rightarrow B)$ 100% of the time, which means that $(A \subseteq B)$. Therefore:

$$Support(A) = Support(B \cap A) \qquad (2.c.1)$$

$$conf(A \rightarrow B) = P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{S(B \cap A)}{S(A)} = \frac{\frac{Support(B \cap A)}{N}}{\frac{Support(A)}{N}} = \frac{Support(B \cap A)}{Support(A)}$$

$\frac{Support(B \cap A)}{Support(A)} = 1$ as shown in equation (2.c.1). Therefore:

$$conf(A \rightarrow B) = 1 \tag{2.c.2}$$

Accordingly, confidence is a perfect implication.

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{S(B)} = \frac{conf(A \rightarrow B)}{\frac{Support(B)}{N}} = \frac{N * conf(A \rightarrow B)}{Support(B)}$$

$conf(A \rightarrow B) = 1$ as shown in equation (2.c.2). Therefore:

$$lift(A \rightarrow B) = \frac{N}{Support(B)}$$

$N \geq Support(B)$, which means that $lift(A \rightarrow B)$ does not have to be maximal. Accordingly, lift is not a perfect implication.

$$conv(A \rightarrow B) = \frac{1 - S(B)}{1 - conf(A \rightarrow B)} = \frac{1 - \frac{Support(B)}{N}}{1 - conf(A \rightarrow B)}$$

$conf(A \rightarrow B) = 1$ as shown in equation (2.c.2). Therefore:

$$conv(A \rightarrow B) = \frac{1 - \frac{Support(B)}{N}}{1 - 1} = \frac{1 - \frac{Support(B)}{N}}{0}$$

Thus, in this case conviction is either undefined or tends towards infinity. It is my understanding that this means conviction is not a perfect implication.

2(d)

Itemset size = 2:
Rule = DAI93865 -> FRO40251 Confidence = 1
Rule = GRO85051 -> FRO40251 Confidence = 0.999176276771005
Rule = GRO38636 -> FRO40251 Confidence = 0.990654205607477
Rule = ELE12951 -> FRO40251 Confidence = 0.990566037735849
Rule = DAI88079 -> FRO40251 Confidence = 0.986725663716814

Itemset size = 3:
Rule = (DAI23334, ELE92920) -> DAI62779 Confidence = 1
Rule = (DAI31081, GRO85051) -> FRO40251 Confidence = 1

Rule = (DAI55911, GRO85051) -> FRO40251 Confidence = 1
Rule = (DAI62779, DAI88079) -> FRO40251 Confidence = 1
Rule = (DAI75645, GRO85051) -> FRO40251 Confidence = 1

2(e)

Itemset size = 2:
Rule = DAI18527 -> SNA44451 Lift = 67.6397014925373
Rule = DAI43868 -> SNA82528 Lift = 50.9434889434889
Rule = ELE28189 -> FRO17734 Lift = 50.102787764484
Rule = ELE88583 -> GRO30912 Lift = 47.4100609756098
Rule = ELE25077 -> GRO89004 Lift = 45.4186477748193

Itemset size = 3:
Rule = SNA53220 -> (FRO19221, SNA93860) Lift = 40.4831468371774
Rule = DAI42083 -> (DAI62779, DAI92600) Lift = 35.3443405415936
Rule = DAI42083 -> (DAI92600, ELE17451) Lift = 34.1711460446247
Rule = (DAI85309, ELE92920) -> SNA18336 Lift = 30.3351998821127
Rule = (DAI42083, DAI62779) -> DAI92600 Lift = 29.8515014397367

2(f)

Itemset size = 2:
Rule = DAI93865 -> FRO40251 Conviction = Infinity
Rule = GRO85051 -> FRO40251 Conviction = 1062.50860100963
Rule = GRO38636 -> FRO40251 Conviction = 93.6477926754767
Rule = ELE12951 -> FRO40251 Conviction = 92.772579659818
Rule = DAI88079 -> FRO40251 Conviction = 65.9327138462857

Itemset size = 3:
Rule = (DAI23334, ELE92920) -> DAI62779 Conviction = Infinity
Rule = (DAI31081, GRO85051) -> FRO40251 Conviction = Infinity
Rule = (DAI55911, GRO85051) -> FRO40251 Conviction = Infinity
Rule = (DAI62779, DAI88079) -> FRO40251 Conviction = Infinity
Rule = (DAI75645, GRO85051) -> FRO40251 Conviction = Infinity

2(g)

Itemset size = 2:
Rule = DAI93865 -> FRO40251 Confidence = 1

Itemset size = 3:
Rule = (DAI23334, ELE92920) -> DAI62779 Confidence = 1
Rule = (DAI31081, GRO85051) -> FRO40251 Confidence = 1
Rule = (DAI55911, GRO85051) -> FRO40251 Confidence = 1
Rule = (DAI62779, DAI88079) -> FRO40251 Confidence = 1
Rule = (DAI75645, GRO85051) -> FRO40251 Confidence = 1
Rule = (ELE17451, GRO85051) -> FRO40251 Confidence = 1
Rule = (ELE20847, FRO92469) -> FRO40251 Confidence = 1
Rule = (ELE20847, GRO85051) -> FRO40251 Confidence = 1
Rule = (ELE26917, GRO85051) -> FRO40251 Confidence = 1
Rule = (FRO53271, GRO85051) -> FRO40251 Confidence = 1
Rule = (GRO21487, GRO85051) -> FRO40251 Confidence = 1
Rule = (GRO38814, GRO85051) -> FRO40251 Confidence = 1
Rule = (GRO73461, GRO85051) -> FRO40251 Confidence = 1
Rule = (GRO85051, SNA45677) -> FRO40251 Confidence = 1
Rule = (GRO85051, SNA80324) -> FRO40251 Confidence = 1

2(h)

I would recommend FRO40251 because of the following rules:

Rule = GRO85051 -> FRO40251 Confidence = 0.999176276771005
Rule = GRO85051 -> FRO40251 Conviction = 1062.50860100963

99.92% of people who purchase GRO85051 also purchase FRO40251.

The rule (GRO85051 -> FRO40251) would be expected to be incorrect 1,062.5x more often if the relationship between GRO85051 and FRO40251 were due to random chance.

2(i)

I would recommend FRO40251 for the same reasons I addressed in 2(h), and also because of the following rule:

Rule = (ELE17451, GRO85051) -> FRO40251 Confidence = 1

100% of people who purchased ELE17451 and GRO85051 also purchased FRO40251.

I would recommend DAI42083 because of the following rule:

Rule = (DAI92600, ELE17451) -> DAI42083 Lift = 34.1711460446247

(DAI92600, ELE17451) occurs with DAI42083 34.17 times more often than would be expected to occur if (DAI92600, ELE17451) and DAI42083 were statistically independent.

I would also recommend DAI62779 because of the following rule:

Rule = DAI92600 -> (DAI42083, DAI62779)   Lift = 29.8515014397367

DAI92600 occurs with (DAI42083, DAI62779) 29.85 times more often than would be expected to occur if DAI92600 and (DAI42083, DAI62779) were statistically independent.

2(j)

I would recommend DAI62779 because of the following rules:

Rule = (DAI23334, ELE92920) -> DAI62779   Confidence = 1
Rule = (DAI23334, ELE92920) -> DAI62779   Conviction = Infinity

100% of people who purchased DAI23334 and ELE92920 also purchased DAI62779.

The rule ((DAI23334, ELE92920) -> DAI62779) would be expected to be incorrect much more often if the relationship between (DAI23334, ELE92920) and DAI62779 were due to random chance.

I would recommend FRO40251 because of the following rules:

Rule = DAI93865 -> FRO40251   Confidence = 1
Rule = DAI93865 -> FRO40251   Conviction = Infinity

Rule = DAI88079 -> FRO40251   Confidence = 0.986725663716814
Rule = DAI88079 -> FRO40251   Conviction = 65.9327138462857

100% of people who purchased DAI93865 also purchased FRO40251. The rule (DAI93865 -> FRO40251) would be expected to be incorrect much more often if the relationship between DAI93865 and FRO40251 were due to random chance.

98.67% of people who purchased DAI88079 also purchased FRO40251. The rule (DAI88079 -> FRO40251) would be expected to be incorrect 65.9x more often if the relationship between DAI88079 and FRO40251 were due to random chance.

3(a)

Given: $P_{h \in F}[h(x) = h(y)] = sim(x, y)$ where $sim(x, y) \in [0, 1]$ $\forall$ pairs of objects (x, y)

Theorem: $d(x, y) = 1 - sim(x, y)$ must satisfy $d(x, y) + d(y, z) \geq d(x, z)$

Proof:

$P_{h \in F}[h(x) = h(y)] = sim(x, y)$

$P_{h \in F}[h(x) \neq h(y)] = 1 - sim(x, y) = d(x, y)$

Because $P_{h \in F}[h(x) \neq h(y)] = d(x, y)$, it follows that $d(x, y) + d(y, z) \geq d(x, z)$ is equivalent to:

$P_{h \in F}[h(x) \neq h(y)] + P_{h \in F}[h(y) \neq h(z)] \geq P_{h \in F}[h(x) \neq h(z)]$

The statement immediately above must necessarily be true because when $h(x) \neq h(z)$ at least one of $h(x)$ and $h(z)$ must be different than $h(y)$.

3(b)

We are given the following:

$$\mathrm{sim}_{Over}(M, N) = \frac{|M \cap N|}{min(|M|, |N|)}$$

$$\mathrm{d}_{Over}(M, N) = 1 - \frac{|M \cap N|}{min(|M|, |N|)}$$

Let $X = \{a, b\}$

Let $Y = \{b, c\}$

Let $Z = \{a, b, c\}$;

where $a, b, c > 0$

$$\mathrm{sim}_{Over}(X, Y) = \frac{a + b}{a + b} = 1$$

$$\mathrm{d}_{Over}(X, Y) = 1 - \mathrm{sim}_{Over}(X, Y) = 1 - 1 = 0$$

$$\mathrm{sim}_{Over}(Y,Z) = \frac{b+c}{b+c} = 1$$

$$\mathrm{d}_{Over}(Y,Z) = 1 - \mathrm{sim}_{Over}(Y,Z) = 1 - 1 = 0$$

$$\mathrm{sim}_{Over}(X,Z) = \frac{b}{b + min(a,c)}$$

$0 < \mathrm{sim}_{Over}(X,Z) < 1$ because we know that $a, b, c > 0$

$$\mathrm{d}_{Over}(X,Z) = 1 - \mathrm{sim}_{Over}(X,Z)$$

$$0 < \mathrm{d}_{Over}(X,Z) < 1$$

$$\mathrm{d}_{Over}(X,Y) + \mathrm{d}_{Over}(Y,Z) = 0 < \mathrm{d}_{Over}(X,Z)$$

$$\mathrm{d}_{Over}(X,Y) + \mathrm{d}_{Over}(Y,Z) < \mathrm{d}_{Over}(X,Z) \tag{3.b.1}$$

The triangle inequality test requires $\mathrm{d}_{Over}(X,Y) + \mathrm{d}_{Over}(Y,Z) \geq \mathrm{d}_{Over}(X,Z)$. The Overlap similarity function fails the triangle inequality test, as shown by equation 3.b.1 above.

3(c)

We are given the following:

$$\mathrm{sim}_{Dice}(M,N) = \frac{|M \cap N|}{\frac{1}{2}(|M| + |N|)}$$

$$\mathrm{d}_{Dice}(M,N) = 1 - \frac{|M \cap N|}{\frac{1}{2}(|M| + |N|)}$$

Let $|X| = 1$

Let $|Z| = 1$

where $X \cap Z = \emptyset$

Let $Y = X \cup Z$

$$\mathrm{sim}_{Dice}(X,Y) = \frac{1}{\frac{1}{2}(1+2)} = \frac{2}{3}$$

$$\mathrm{d}_{Dice}(X,Y) = 1 - \mathrm{sim}_{Dice}(X,Y) = 1 - \frac{2}{3} = \frac{1}{3}$$

$$\mathrm{sim}_{Dice}(Y,Z) = \frac{1}{\frac{1}{2}(1+2)} = \frac{2}{3}$$

$$\mathrm{d}_{Dice}(Y,Z) = 1 - \mathrm{sim}_{Dice}(Y,Z) = 1 - \frac{2}{3} = \frac{1}{3}$$

$$\mathrm{sim}_{Dice}(X,Z) = \frac{0}{\frac{1}{2}(1+1)} = 0$$

$$\mathrm{d}_{Dice}(Y,Z) = 1 - \mathrm{sim}_{Dice}(Y,Z) = 1 - 0 = 1$$

$$\mathrm{d}_{Dice}(X,Y) + \mathrm{d}_{Dice}(Y,Z) = \tfrac{1}{3} + \tfrac{1}{3} = \tfrac{2}{3} < 1 = \mathrm{d}_{Over}(X,Z)$$

$$\mathrm{d}_{Dice}(X,Y) + \mathrm{d}_{Dice}(Y,Z) < \mathrm{d}_{Over}(X,Z) \qquad (3.c.1)$$

The triangle inequality test requires $\mathrm{d}_{Dice}(X,Y) + \mathrm{d}_{Dice}(Y,Z) \geq \mathrm{d}_{Over}(X,Z)$. The Dice similarity function fails the triangle inequality test, as shown by equation 3.c.1 above.

4(a)

$$P[x \in W_j] = P[g_j(x) = g_j(z)] = P[h_1^j(x) = h_1^j(z), h_2^j(x) = h_2^j(z), \ldots, h_k^j(x) = h_k^j(z)]$$

$$P[x \in W_j] = \textstyle\prod_{i=1}^{k} P[h_i^j(x) = h_i^j(z)] \qquad (4.a.1)$$

Because $\mathcal{H}$ is $(\lambda, c\lambda, p_1, p_2)$-sensitive and $T = \{x \in \mathcal{A} \mid d(x,z) > c\lambda\}$, it follows that:

$$P[h(x) = h(z)] \leq p_2 \; \forall x^* \in T \qquad (4.a.2)$$

From equations (4.a.1) and (4.a.2) and the fact that $k = \log_{1/p_2} n$, we obtain:

$$P[x \in T \cap W_j] \leq \prod_{i=1}^{k} p_2 = p_2{}^k = p_2{}^{\log_{1/p_2} n} = n^{-1}$$

$$P[x \in T \cap W_j] \leq n^{-1} \tag{4.a.3}$$

Let I_x^j be an indicator variable s.t. $I_x^j = 1$ when $x \in (T \cap W_j)$ and $I_x^j = 0$ when $x \notin (T \cap W_j)$.

From equation (4.a.3), it follows that:

$$P[I_x^j = 1] \leq n^{-1} \tag{4.a.4}$$

From equation (4.a.4), it follows that:

$$E[I_x^j] = 1 * P[I_x^j = 1] + 0$$

$$E[I_x^j] \leq n^{-1} \tag{4.a.5}$$

From equation (4.a.5) and the fact that $|T| \leq n$, it follows that:

$$E[\textstyle\sum_{x \in T} I_x^j] \leq |T| * n^{-1} = \frac{|T|}{n} \leq 1 \tag{4.a.6}$$

Because $E[|T \cap W_j|] = E[\sum_{x \in T} I_x^j]$, and equation (4.a.6), it follows that:

$$E[|T \cap W_j|] \leq 1 \tag{4.a.7}$$

Now using equation (4.a.7), we can determine an upper limit for $E[\sum_{j=1}^{L} |T \cap W_j|]$ as follows:

$$E\left[\sum_{j=1}^{L} |T \cap W_j|\right] = \sum_{j=1}^{L} E[|T \cap W_j|] \leq \sum_{j=1}^{L} 1 = L$$

$$E[\textstyle\sum_{j=1}^{L} |T \cap W_j|] \leq L \tag{4.a.8}$$

Now, we can use equation (4.a.8) and Markov's inequality to determine an upper bound for $P\left[\sum_{j=1}^{L}\left|T \cap W_j\right| > 3L\right]$, as follows:

$$P\left[\sum_{j=1}^{L}\left|T \cap W_j\right| > 3L\right] \leq \frac{E\left[\sum_{j=1}^{L}\left|T \cap W_j\right|\right]}{3L} \leq \frac{L}{3L} = \frac{1}{3}$$

$$P\left[\sum_{j=1}^{L}\left|T \cap W_j\right| > 3L\right] \leq \frac{1}{3}$$

4(b)

$$P\left[g_j(x^*) = g_j(z)\right] = P\left[h_1^j(x^*) = h_1^j(z), h_2^j(x^*) = h_2^j(z), \dots, h_k^j(x^*) = h_k^j(z)\right]$$

Because $\mathcal{H}$ is a LSH family, it follows that $h \in \mathcal{H}$ is selected uniformly at random. Therefore, ∀j:

$$P\left[g_j(x^*) = g_j(z)\right] = \prod_{i=1}^{k} P\left[h_i^j(x^*) = h_i^j(z)\right]$$

It follows that ∀j $\in [1, L]$:

$$P\left[g_j(x^*) \neq g_j(z)\right] = 1 - P\left[g_j(x^*) = g_j(z)\right] = 1 - \prod_{i=1}^{k} P\left[h_i^j(x^*) = h_i^j(z)\right] \qquad (4.b.1)$$

Because $\mathcal{H}$ is $(\lambda, c\lambda, p_1, p_2)$-sensitive, it follows that $\forall x^* \in A$ such that $d(x^*, z) \leq \lambda$ and $\forall h_i \in \mathcal{H}$:

$$P[h_i(x^*) = h_i(z)] \geq p_1$$

It follows that:

$$\prod_{i=1}^{k} P\left[h_i^j(x^*) = h_i^j(z)\right] \geq \prod_{i=1}^{k} p_1 = {p_1}^k \qquad (4.b.2)$$

Because $k = \log_{1/p_2} n$ and $p = \frac{\log 1/p_1}{\log 1/p_2}$, we can reduce ${p_1}^k$ as follows:

$${p_1}^k = {p_1}^{\log_{1/p_2} n} = n^{-\frac{\log 1/p_1}{\log 1/p_2}} = n^{-p} \qquad (4.b.3)$$

From equations (4.b.2) and (4.b.3), we obtain:

$$\prod_{i=1}^{k} P\left[h_i^j(x^*) = h_i^j(z)\right] \ge n^{-p}$$

$$1 - \prod_{i=1}^{k} P\left[h_i^j(x^*) = h_i^j(z)\right] \le 1 - n^{-p} \quad (4.b.4)$$

From equations (4.b.1) and (4.b.4), we obtain:

$$P\left[g_j(x^*) \ne g_j(z)\right] \le 1 - n^{-p} \quad (4.b.5)$$

$$P\left[g_j(x^*) \ne g_j(z)\right] \forall \mathrm{j} \in [1, L] = \sum_{j=1}^{L} P\left[g_j(x^*) \ne g_j(z)\right]$$

From the above and equation (4.b.5), we obtain:

$$P\left[g_j(x^*) \ne g_j(z)\right] \forall \mathrm{j} \in [1, L] = \sum_{j=1}^{L} P\left[g_j(x^*) \ne g_j(z)\right] \le (1 - n^{-p})^L$$

$$P\left[g_j(x^*) \ne g_j(z)\right] \forall \mathrm{j} \in [1, L] \le (1 - n^{-p})^{n^p} \quad (4.b.6)$$

From equation (4.b.3), we know that ${p_1}^k = n^{-p}$. It follows that:

$n^{-p} \in [0, 1]$ and $n^p \ge 1$ (4.b.7)

We know that $(1 - x)^{1/x} \forall \mathrm{x} \in (0, 1] \le \frac{1}{e}$ (4.b.8)

From equations (4.b.6), (4.b.7), and (4.b.8), we obtain:

$$P\left[g_j(x^*) \ne g_j(z)\right] \forall \mathrm{j} \in [1, L] \le \frac{1}{e}$$

<u>4(c)</u>

We determined the probability of a false positive in question 4(a) to be the following:

$$P\left[\sum_{j=1}^{L} |T \cap W_j| > 3L\right] \leq \frac{1}{3}$$

We determined the probability of a false negative in question 4(b) to be the following:

$$P\left[g_j(x^*) \neq g_j(z)\right] \forall \mathrm{j} \in [1, L] \leq \frac{1}{e}$$

The probability of an error (either a false positive or a false negative) can therefore be bounded to less than 1 as follows:

$$P[False\ Positive\ or\ False\ Negative] \leq \frac{1}{3} + \frac{1}{e} \approx 0.70 < 1$$

4(d)

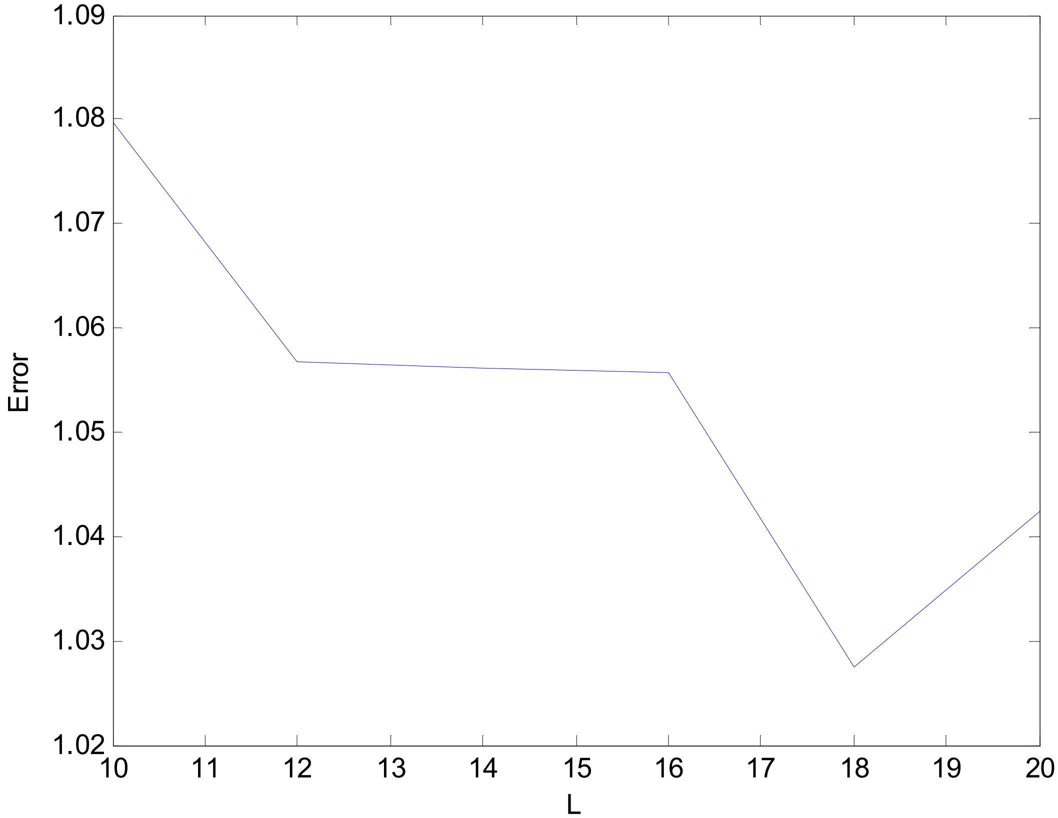
Note that I used the image patch in column 601 rather than column 600 due to the problem described at the following link: https://piazza.com/class#winter2013/cs246/222

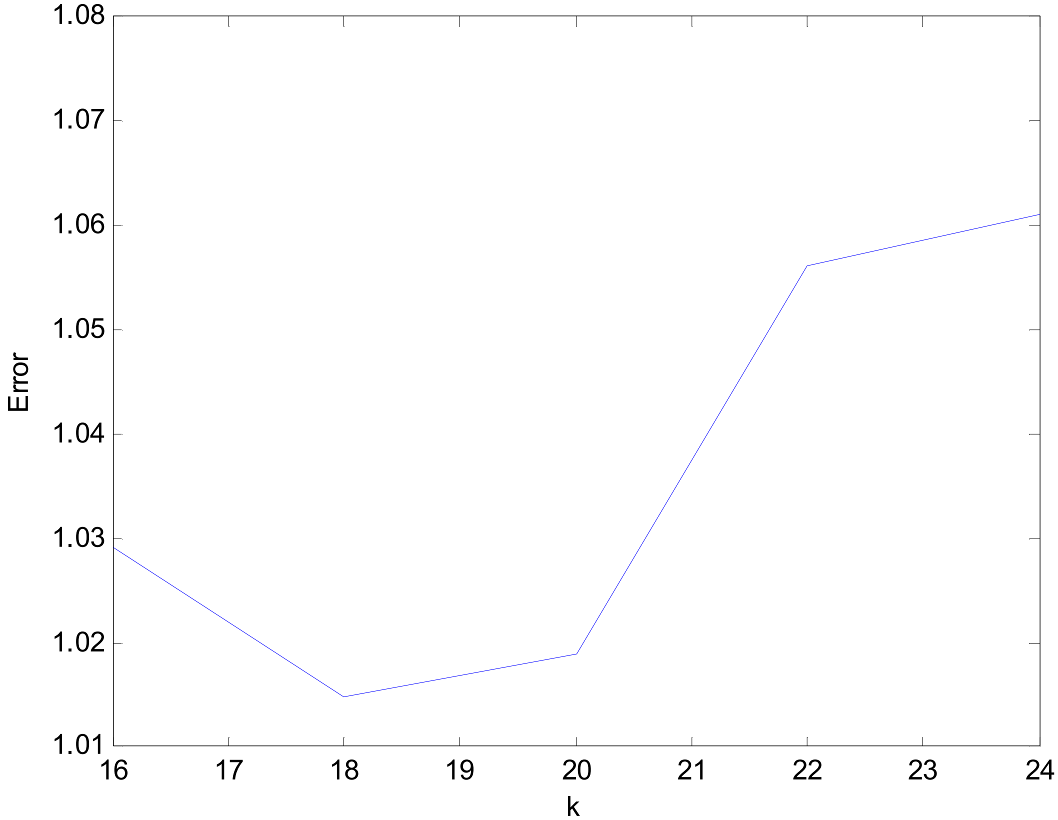
Average LSH Search Time = 0.014536

Average Exhaustive Search Time = 0.135361

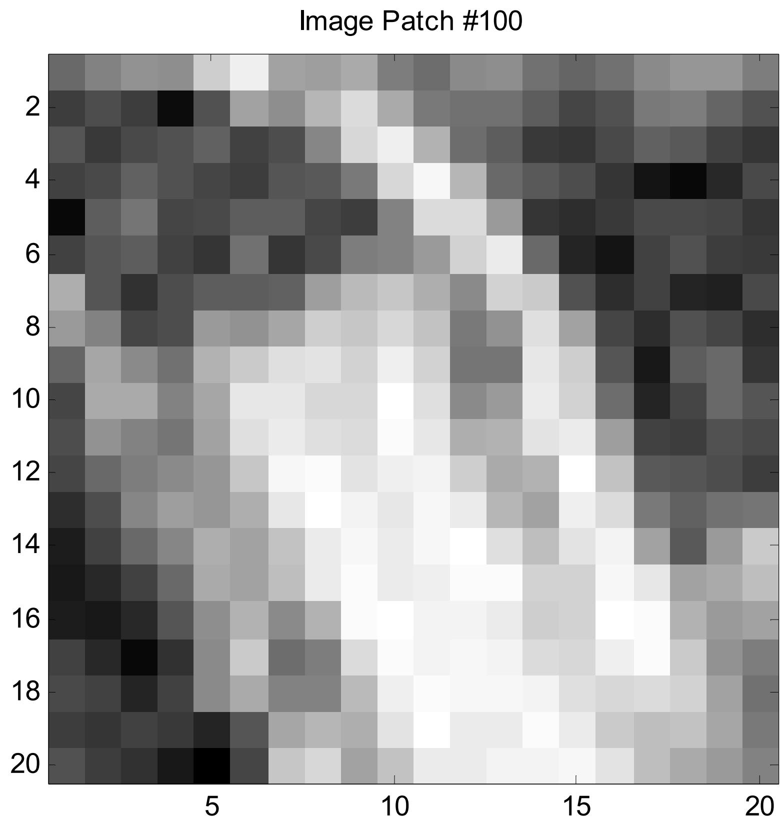
Exhaustive Search took 9.312306 times longer than LSH Search.

Error = 1.063324

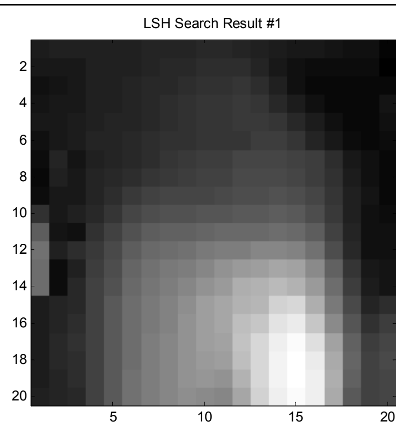
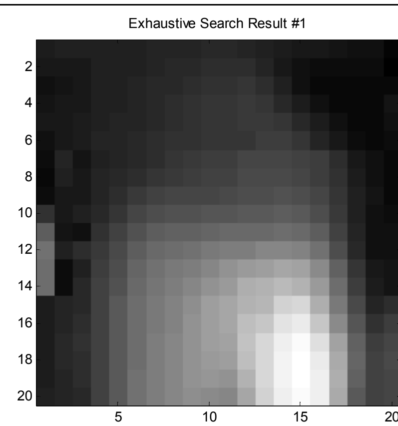

1.09
1.08
1.07
1.06
1.05
1.04
1.03
1.02
Error
10
11
12
13
14
15
16
17
18
19
20
L

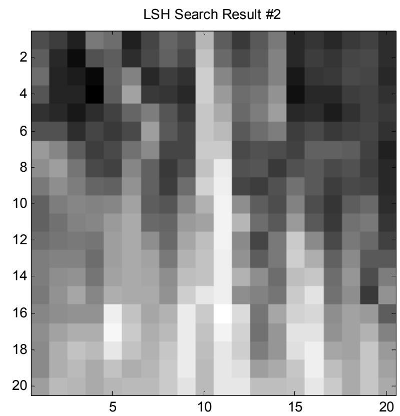


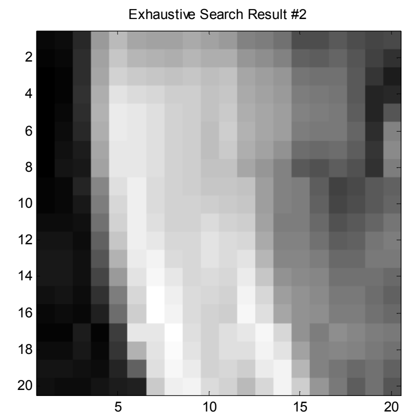
1.08
1.07
1.06
1.05
1.04
1.03
1.02
1.01
Error
16
17
18
19
20
21
22
23
24
k

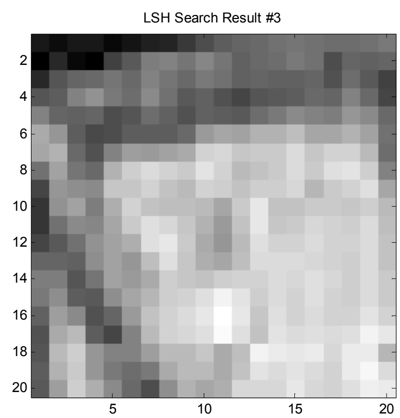


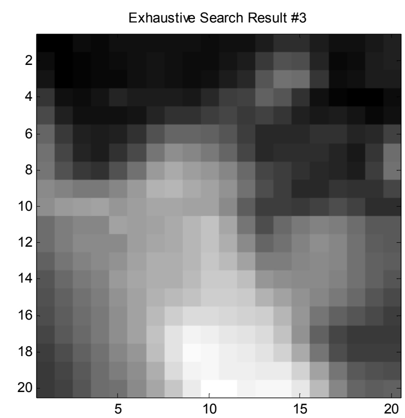
The following table shows the LSH search results (left) and the exhaustive search results (right). In general, the exhaustive search results don't strike me as being much closer to image #100 than the LSH search results.

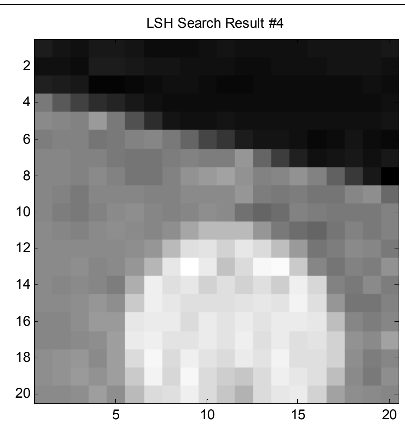
	

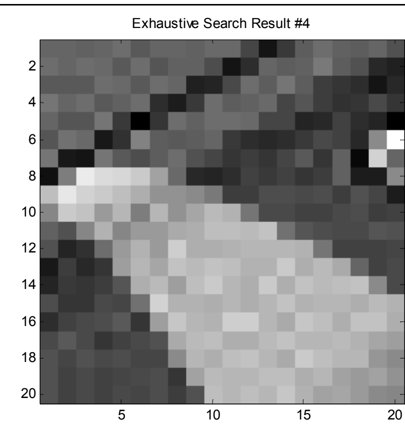

LSH Search Result #2
2 4 6 8 10 12 14 16 18 20
5 10 15 20

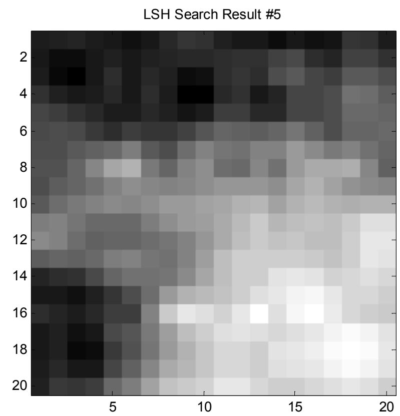

Exhaustive Search Result #2
2 4 6 8 10 12 14 16 18 20
5 10 15 20

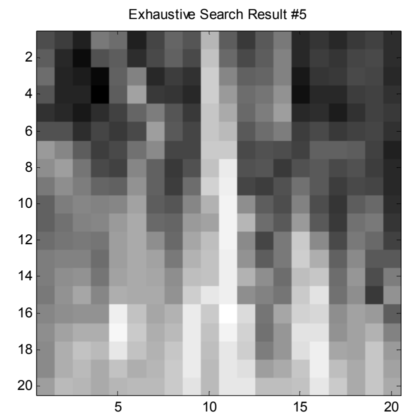

LSH Search Result #3
2 4 6 8 10 12 14 16 18 20
5 10 15 20

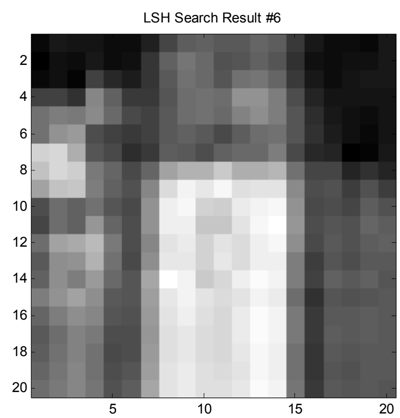

Exhaustive Search Result #3
2 4 6 8 10 12 14 16 18 20
5 10 15 20

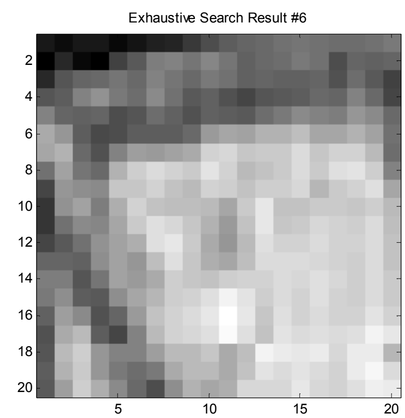

LSH Search Result #4
2 4 6 8 10 12 14 16 18 20
5 10 15 20

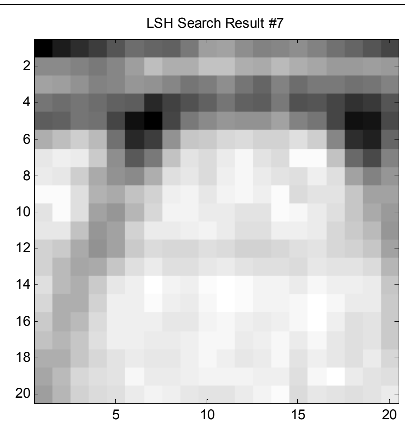

Exhaustive Search Result #4
2 4 6 8 10 12 14 16 18 20
5 10 15 20

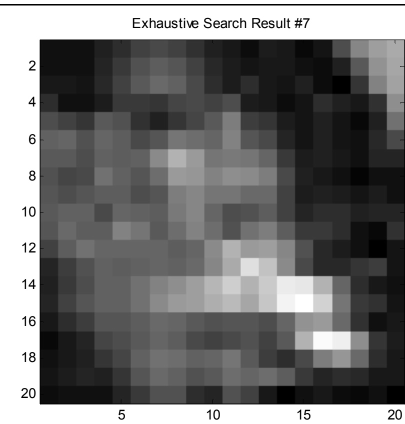

LSH Search Result #5
2 4 6 8 10 12 14 16 18 20
5 10 15 20

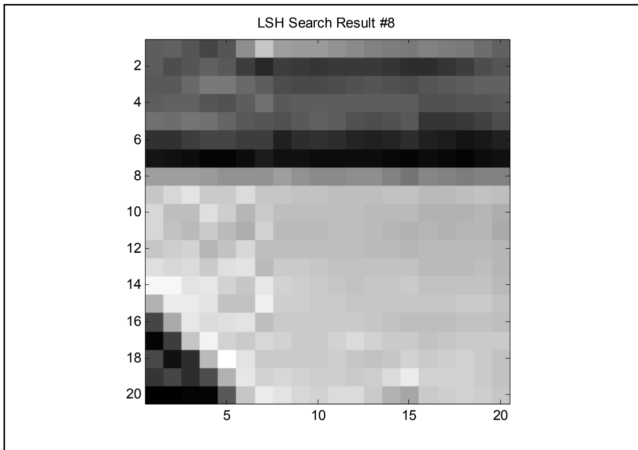

Exhaustive Search Result #5
2 4 6 8 10 12 14 16 18 20
5 10 15 20

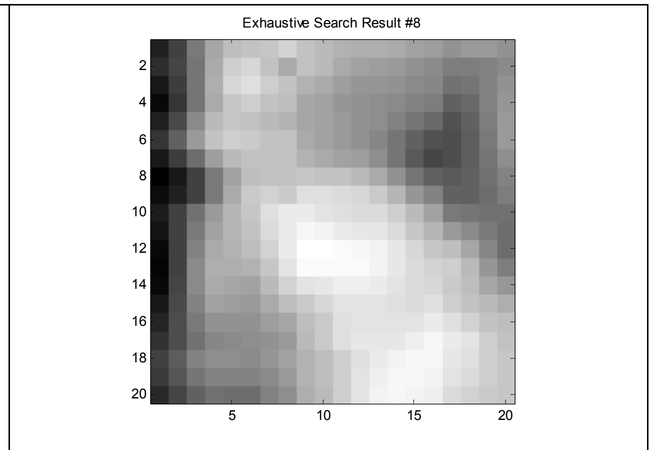

LSH Search Result #6
2 4 6 8 10 12 14 16 18 20
5 10 15 20

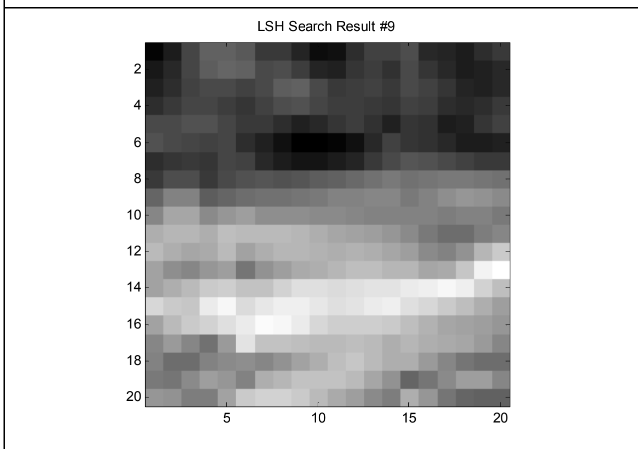

Exhaustive Search Result #6
2 4 6 8 10 12 14 16 18 20
5 10 15 20

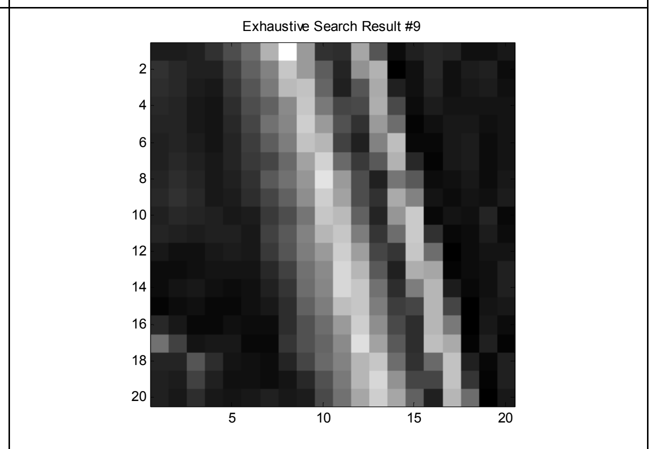

LSH Search Result #7
2 4 6 8 10 12 14 16 18 20
5 10 15 20

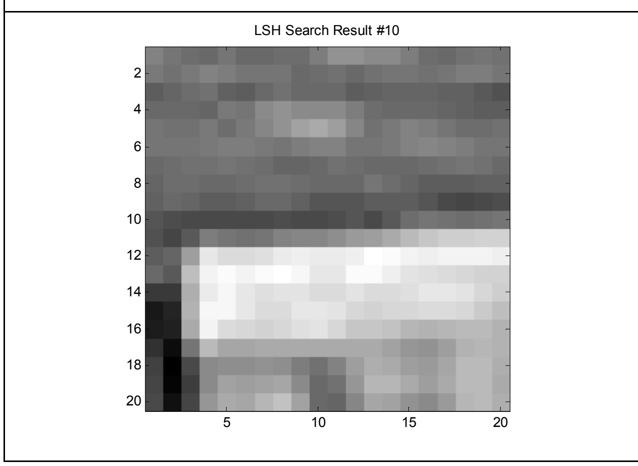

Exhaustive Search Result #7
2 4 6 8 10 12 14 16 18 20
5 10 15 20

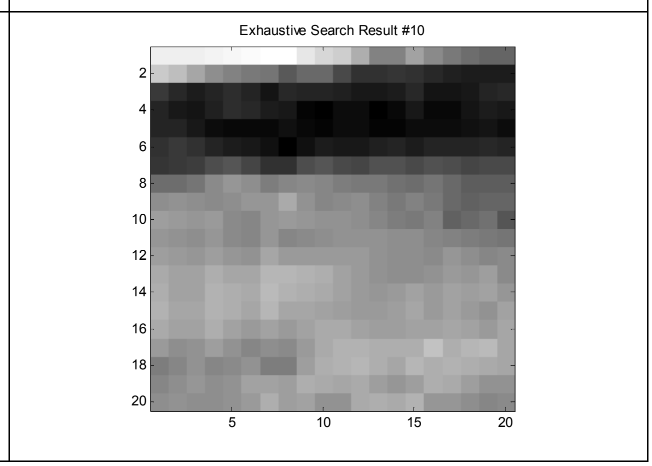

LSH Search Result #8


Exhaustive Search Result #8


LSH Search Result #9


Exhaustive Search Result #9


LSH Search Result #10


Exhaustive Search Result #10