Quiz 3

Online (In-class) Quiz –

1) [1 Point] Suppose we want to find similar items and we do so by min hashing the set 10 times and then applying LSH with 5 bands and 2 rows each. If two sets have Jaccard Similarity 0.5, what is the probability that they will be identified by LSH as candidate pairs? (^ represents power/exponential)

A.) (1-0.5^2) ^5
B.) (1-(1-0.5^2) ^5) ^10
C.) 1-0.5^2
D.) None of the above
Ans.) D

2) [1 Point] What are the 2-shingles for “abcdabc”? (1pt)

Answer –

The 2-shingles are {ab,bc,cd,da}

3) [0.5 Point] The Jaccard bag similarity of A(1, 1, 2, 2) and B(1, 1, 1, 2, 2, 3) is 2/3? True or False

Answer –

False

1.Jaccard(A,B) = |A∩B|/|A∪B| = 2/3 = 0.5

2.Jaccard-bag-similarity(A,B) = 4/10 = 0.4

4) [3 point] Consider the following characteristic matrix of two sets: S1 and S2.

Row #	S1	S2
0	1	0
1	1	1
2	1	1
3	0	1
4	1	1
5	1	0
6	0	1

Suppose the two hash functions: h1(x) = (x + 3) mod 8 and h2(x) = (2 * x) mod 5
a) Construct a signature for S1 and S2 based on the minhash values obtained from h1(x) and h2(x) above.
b) Estimate the Jaccard similarity of S1 and S2 using the signature.

Row #	S1	S2	h1(x) = (x+3) mod 8	h2(x) = (2*x) mod 5
0	1	0	3	0
1	1	1	4	2
2	1	1	5	4
3	0	1	6	1
4	1	1	7	3
5	1	0	0	0
6	0	1	1	2

Minhash Signature:

	S1	S2
h1	0	1
h2	0	1

J(S1, S2) = 0/2 = 0

5) [3 Point] Suppose that two sets are considered to be similar if their Jaccard similarity is greater than or equal to 0.6. Consider two sets S1 and S2. Suppose that their actual Jaccard similarity is 0.8.

Consider their minhash signatures S1' and S2', each having 100 minhash values. Suppose the signatures are divided into 20 bands with 5 rows in each band. That is, b = 20, r = 5. Locality-sensitive hashing (LSH) is then applied to the signatures to obtain candidate pairs of sets. What is the probability that S1 and S2 are not identified as a candidate pair (i.e., false negative rate)?

Probability that S1 and S2 are identified as a candidate pair in a single band = s^r = 0.8^5= 0.32768

Probability that S1 and S2 are not identified as a candidate pair in a single band = (1-s^r)

Probability that S1 and S2 are not candidate pair in any bands =>

False Negative rate = (1-s^r)^b

(1-0.8^5)^20 = 3.56X10-4

6) [1 point] What is the effect of following on False positive and False negative:

a. Increasing B, keeping r constant [0.5 points]
b. Increasing r, keeping b constant [0.5 points]

a. Decreases false negatives and Increases false positives [0.5, if both mentioned]

b. Increase False Negatives and Decreases False positives [0.5, if both mentioned]

7) [0.5 points] LSH is based on the idea that if 2 itemsets are similar their signatures SHOULD NOT be similar? True or False

Answer - False

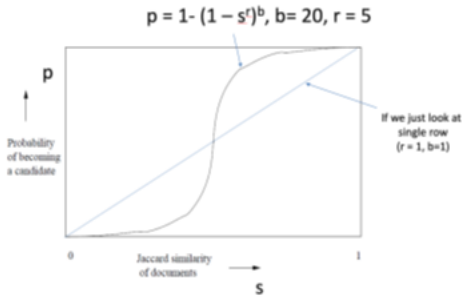
Offline (Take-home) Quiz –

1) [3 points] When we perform fingerprint matching with LSH, there are 2 kinds of problems that we discussed in class. Describe these 2 problems and how the technique of LSH is applied.

(1) Many to many problem: take an entire database of fingerprints and identify if there are any pairs that represent the same individual
1. Define a locality-sensitive family of hash functions:
Each function f in the family F is defined by 3 grid squares
Function f says "yes" for two fingerprints if both have minutiae in all three grid squares, otherwise, f says "no"
"Yes" means the two fingerprints are candidate pairs !
2. Sort of "bucketization"
Each set of three points creates one bucket
Function f sends fingerprints to its bucket that have minute in all three grid points of f
3. Compare all fingerprints in each of the buckets.

(2) Many to one problem: A fingerprint has been found at a crime scene, and we want to compare it with all fingerprints in a large database to see if there is a match
1. Could use many functions f from family F
2. Precompute their buckets of fingerprints to which they answer "yes" on the large database! 3. For a new fingerprint:
Determine which buckets it belongs to
Compare it with all fingerprints found in any of those buckets.

2) [3 points] LSH: explain what the curve in the figure below represents and show the steps to derive the curve function from s, r, and b (you also need to explain what s, r, and b represent) (2 pts) How would b and r change the curve and why do we care about them in terms of false positives and false negatives (1 pt)?



Answer –

What it represents (0.5 points) –

The graph represents the probability of becoming a candidate w.r.t the Jaccard Similarity of documents. Area under the curve is predicted positive and area above the curve is predicted negatives.

Steps to derive (1.5 points) -

b – number of bands (we divide signatures into b bands)

r – r rows per band

s - the probability the minhash signatures for these documents agree in any one particular row of the signature matrix

s^r is the probability of signatures agree on all rows in one band;

$(1- s^r)$ is the probability that they disagree on at least one row in a band;

$(1 – s^r)^b$ is the probability that they disagree on at least one row in all bands;

So,$1-(1 – s^r)^b$ is the probability that they agree on all rows in at least one band.

(explain what s, r, and b are: 0.5 points; proof – 1 point)

Changing b and r –

Increase r – Move curve right and down, Increase false negative, Reduce false positive

Increase b – Move curve left and up, Increase False Positive, Reduce false negative

3) [3 points] In one pass (from the top row to bottom row), generate the minhash signature for each set S. There are three minhash functions. The first minhash function is x + 1 mod 5, the second one is 3x + 1 mod 5, and the third one is x+2 mod 5. Compute the Jaccard similarity and Estimated similarity between (S1, S3), (S1, S4), and (S3, S4) (1 pt)

Row	S_1	S_2	S_3	S_4	$x+1 \mod 5$	$3x+1 \mod 5$
0	1	0	0	1	1	1
1	0	0	1	0	2	4
2	0	1	0	1	3	2
3	1	0	1	1	4	0
4	0	0	1	0	0	3

Answer-

Row	S1	S2	S3	S4	x + 1 mod 5	3x + 1 mod 5	x + 2 mod
0	1	0	0	1	1	1	2
1	0	0	1	0	2	4	3
2	0	1	0	1	3	2	4
3	1	0	1	1	4	0	0
4	0	0	1	0	0	3	1

Initially	S1	S2	S3	S4
h1(x)	∞	∞	∞	∞
h2(x)	∞	∞	∞	∞
h3(x)	∞	∞	∞	∞

Initially	S1	S2	S3	S4
h1(x)	1	∞	∞	1
h2(x)	1	∞	∞	1
h3(x)	2	∞	∞	2

Initially	S1	S2	S3	S4
h1(x)	1	∞	2	1
h2(x)	1	∞	4	1
h3(x)	2	∞	3	2

Initially	S1	S2	S3	S4
h1(x)	1	3	2	1
h2(x)	1	2	4	1
h3(x)	2	4	3	2

Initially	S1	S2	S3	S4
h1(x)	1	3	2	1
h2(x)	0	2	0	0
h3(x)	0	4	0	0

Initially	S1	S2	S3	S4
h1(x)	1	3	0	1
h2(x)	0	2	0	0
h3(x)	0	4	0	0

	Jaccard Similarity	Estimated Similarity
(S1, S3)	1/4	2/3
(S1, S4)	2/3	1
(S3, S4)	1/5	2/3

4) [1 Point] Suppose we want to find similar items and we do so by min hashing the set 10 times and then applying LSH with 5 bands and 2 rows each. If two sets have Jaccard Similarity 0.5, what is the probability that they will be identified by LSH as candidate pairs? (^ represents power/exponential)

A.) (1-0.5^2) ^5
B.) (1-(1-0.5^2) ^5) ^10
C.) 1-0.5^2
D.) None of the above
Ans.) D