02: Locality Sensitive Hashing

Number of questions:	5
Positive points per question:	3.0
Negative points per question:	1.0

Gradiance quiz on Locality Sensitive Hashing. You can attempt to answer the questions as many times as you like. Questions get randomly regenerated each time. The score of the *last* submission gets saved into our records (that is, once you get a perfect score, don't submit again with a bad one).

1. Consider the following three vectors u, v, w in a 6-dimensional space:
u = [1, 0.25, 0, 0, 0.5, 0]
v = [0.75, 0, 0, 0.2, 0.4, 0]
w = [0, 0.1, 0.75, 0, 0, 1]

Suppose we construct 3-bit sketches of the vectors by the random hyperplane method, using the randomly generated normal vectors r1, r2, and r3, in that order:
r1 = [1, -1, 1, -1, 1, -1]
r2 = [-1, -1, 1, 1, -1, 1]
r3 = [1, 1, 1, 1, 1, 1]

Construct the sketches of the three vectors u, v, w. Estimate the pairwise cosine similarities of u, v, and w from their 3-bit sketches. Which of the following estimates corresponds to the computed 3-bit sketches?

- a) cos(v,w) = 1
- b) cos(u,v) = 0.5
- c) cos(u,w) = 1
- d) cos(u,w) = -0.5

2. Suppose we have an LSH family h of $(d_1,d_2,.6,.4)$ hash functions. We can use three functions from h and the AND-construction to form a (d_1,d_2,w,x) family, and we can use two functions from h and the OR-construction to form a (d_1,d_2,y,z) family. Calculate w, x, y, and z, and then identify the correct value of one of these in the list below.

- a) w=.216
- b) z=.84
- c) x=.216
- d) z=.16

3. Consider the following matrix:

	C1	C2	C3	C4
R1	0	1	1	0
R2	1	0	1	1
R3	0	1	0	1
R4	0	0	1	0

R4	0	0	1	0
R5	1	0	1	0
R6	0	1	0	0

Perform a minhashing of the data, with the order of rows: R4, R6, R1, R3, R5, R2. Which of the following is the correct minhash value of the stated column? **Note**: we give the minhash value in terms of the original name of the row, rather than the order of the row in the permutation. These two schemes are equivalent, since we only care whether hash values for two columns are equal, not what their actual values are.

- a) The minhash value for C4 is R3
- b) The minhash value for C4 is R5
- c) The minhash value for C2 is R3
- d) The minhash value for C2 is R1

4. Here is a matrix representing the signatures of seven columns, C1 through C7.

C1	C2	C3	C4	C5	C6	C7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Suppose we use locality-sensitive hashing with three bands of two rows each. Assume there are enough buckets available that the hash function for each band can be the identity function (i.e., columns hash to the same bucket if and only if they are identical in the band). Find all the candidate pairs, and then identify one of them in the list below.

- a) C1 and C3
- b) C2 and C4
- c) C3 and C6
- d) C1 and C2

5. Suppose we have computed signatures for a number of columns, and each signature consists of 24 integers, arranged as a column of 24 rows. There are N pairs of signatures that are 50% similar (i.e., they agree in half of the rows). There are M pairs that are 20% similar, and all other pairs (an unknown number) are 0% similar.

We can try to find 50%-similar pairs by using Locality-Sensitive Hashing (LSH), and we can do so by choosing bands of 1, 2, 3, 4, 6, 8, 12, or 24 rows. Calculate approximately, in terms of N and M, the number of false positive and the number of false negatives, for each choice for the number of rows. Then, suppose that we assign equal cost to false positives and false negatives (an atypical assumption). Which number of rows would you choose if M:N were in each of the following ratios: 1:1, 10:1, 100:1, and 1000:1? Identify the correct choice from the list below.

- a) For M = N, pick $r = 2$.
- b) For M = 1000N, pick $r = 4$.
- c) For M = 10N, pick $r = 4$.
- d) For M = 100N, pick $r = 4$.