DAT630 Fall 2017

Exercises on Locality Sensitive Hashing.

Vinay Setty (vinay.j.setty@uis.no)

Exercise 1) Minhashing:

Element	S1	S2	S3	S4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

a. Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x)=5x+2 \bmod 6$.

b. Which of these hash functions are true permutations?

c. How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard

Exercise 2)

In the lecture, you learned how to compute minhash signatures for a given binary matrix representation of documents and shingles. Suppose we want to use a MapReduce framework to compute minhash signatures. If the matrix is stored in chunks that correspond to some columns, then it is quite easy to exploit parallelism. Each Map task gets some of the columns and all the hash functions, and computes the minhash signatures of its given columns. However, suppose the matrix were chunked by rows, so that a Map task is given the hash functions and a set of rows to work on. Design Map and Reduce functions to exploit MapReduce with data in this form. You can write the map and reduce functions as a pseudocode or you can also just describe what mappers and reducers do in your own words.

Exercise 3) Prove that if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity

Exercise 4)

- **For any random permutation π on the binary shingle matrix Prove that $Pr[h_\pi(C_1) = h_\pi(C_2)] = sim(C_1, C_2)$**

Exercise 5) Distance Functions:

Which of the following distance functions are metrics? Provide a proof for your claim.

a. Jaccard distance
b. Cosine similarity
c. Hamming distance
d. max(x, y) = the larger of x and y.
e. sum(x,y) = x+y.