Machine Learning for Descriptive Problems

Exercises 2

Minhashing

Exercise 3.1.1: Exercise 3.1.1 : Compute the Jaccard similarities of each pair of the following three sets: {1, 2, 3, 4}, {2, 3, 5, 7}, and {2, 4, 6}.
Exercise 3.1.3: Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T ?
Exercise 3.3.2: In Fig. 3.5 is a matrix with six rows.

A. Compute the minhash signature for each column if we use the following three hash functions: h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x)=5x+2 mod 6.
B. Which of these hash functions are true permutations?
C. How close are the estimated Jacquard similarities for the six pairs of columns to the true Jaccard similarities?

Element	S_1	S_2	S_3	S_4
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

Locality Sensitive Hashing

Exercise Chapter 3.4: Print the S-curve $1 - (1 - s^r)^b$ for different values of `r` and `b` with `r*b = 100` (approximately), and find the best combination or `r`,`b` for (a) t=0.3, (b) t=0.8, (c) t=0.9. The best combination is such that its S-curve plot has a value 0.5 at position t (see lecture 5, slide 32).

Exercise 5: Suppose we want to implement a parallel version of Locality-Sensitive Hashing by Spark on Hadoop. Assume the input files are already distributed across the computational nodes.

A. How do you parallelize each step? (Shingling, Minhashing and LSH)
B. Describe what data you need to send between the parallel machines.

Exercise 6: Please give your feedback about hands-on programming assignments during lectures. Do you want to have such tasks? What can be improved, or made differently?