Data Science and Big Data

Summer Term 2019

5 - Locality Sensitive Hashing

submission deadline: May 29, 2019 before the lecture

solutions will be discussed: May 29, 2019

Tutor: Tobias Elvermann (t.elvermann@gmail.com)
Grading: Rajkumar Ramamurthy (ramamurt@cs.uni-bonn.de)

1 Hamming Spaces [15 points]

Prove that for all $c > 1$, $\mathcal{H}_{\text{Hamming}}$ is a

$$(r, cr, 1 - \frac{r}{d}, 1 - \frac{cr}{d})\text{-sensitive}$$

family of hash functions for the Hamming distance.

2 AND-Amplification Lemma [10 points]

Prove the AND-Amplification lemma on slide 18 of lecture 2019–05–29.

3 OR-Amplification Lemma [15 points]

Prove the OR-Amplification lemma on slide 19 of lecture 2019–05–29.

4 The (ϵ, δ, r)-PLEB LSH Algorithm [20 points]

Implement the algorithm on slides 24 and 25 of lecture 2019–05–29 for the Hamming distance and the corresponding family of hash functions $\mathcal{H}_{\text{Hamming}}$. Run the algorithm for parameters $\epsilon = 0.2$, $\delta = 0.05$, and $r = 20$ on a dataset P of 10000 randomly generated vectors of dimension 100.

Create two query sets of 1000 vectors each: (1) randomly, (2) by taking the first 1000 vectors from P and flipping 10 random bits in each of them. Look at the distribution of the exits of the query algorithm (i.e., whether it returns in line 4, 7, or 8) for both query sets and interpret the results.

Hint: `numpy.random.randint(low=0, high=2, size=[10000,100])` nicely generates the required set P if you are using python.