Data Science and Big Data

Summer Term 2019

2 - Data Streams II, Metrics

submission deadline: May 8, 2019 before the lecture

solutions will be discussed: May 8, 2019

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

1. Bloom Filter [10 points]

Construct the Bloom filter for the set $\{9, 11\}$ for $n = 5$ and $k = 2$. Use the hash functions

$$\begin{aligned} h_1(x) &= x \bmod 5 \\ h_2(x) &= 2x + 3 \bmod 5 \end{aligned}$$

2. Union of Bloom Filters [25 points]

Let B_1 and B_2 be Bloom filters of size n that represent the sets $S_1, S_2 \subseteq U$, respectively. Suppose both Bloom filters use the same k hash functions. Construct the Bloom filter B for $S_1 \cup S_2$ from B_1 and B_2. What can you say about the probability of a false positive for B?

3. Distances [25 points]

(i) Show that the Minkowski distance D_p violates the triangle inequality for $p < 1$.

(ii) Show that the Jaccard distance is a metric.

(iii) Show that the edit distance is a metric.