Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Julian Busch, Daniyal Kazempour

Munich, 20.12.2017

Big Data Management and Analytics

WS 2017/18

Tutorial 8: Stream Clustering and Text Processing

Assignment 8-1 *CluStream*

Given the following series of data points.

Time t	1	2	3	4	5	6	7	8	9	10	11	12
Data point p	$\binom{1}{1}$	$\binom{2}{1}$	$\binom{4}{9}$	$\binom{4}{8}$	$\binom{10}{4}$	$\binom{9}{3}$	$\binom{2}{3}$	$\binom{11}{3}$	$\binom{12}{12}$	$\binom{12}{11}$	$\binom{11}{12}$	$\binom{4}{2}$

Perform the online steps of the CluStream algorithm on the data point series with the following settings:

- $initPoints = 6$
- $q = 3$
- factor of clu radius $t = 5$

Assignment 8-2 *Finding similar items*

Suppose that the universal set is given by $\{1, \ldots, 10\}$. Construct minhash signatures for the following sets:

(a) $S_1 = \{3, 6, 9\}$

(b) $S_2 = \{2, 4, 6, 8\}$

(c) $S_3 = \{2, 3, 4\}$

1. Construct the signatures for the sets using the following list of permutations:
   - (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
   - (10, 8, 6, 4, 2, 9, 7, 5, 3, 1)
   - (4, 7, 2, 9, 1, 5, 3, 10, 6, 8)
2. Suppose that instead of using particular permutations to construct signatures for the threse sets, we use hash functions. The three hash functions we use are:
   - $h_1(x) = x \mod 10$
   - $h_2(x) = (2x + 1) \mod 10$
   - $h_3(x) = (3x + 2) \mod 10$
3. How does the estimated Jaccard similarity, derived from $(1.)$ and $(2.)$ compare with the true Jaccard similarity of the original data? How to reduce deviations in the approximated Jaccard similarities?