Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

SS 2019

Exercise 9: Sequential Data

Exercise 9-1 Manhattan Distance and Edit Distance

Given an alphabet $A = \{a_1, \ldots, a_n\}$, the histogram of a sequence $S = (s_1, \ldots, s_l)$ is defined as $H(S) = (h_1(S), \ldots, h_n(S))$ with $h_k(S) = |\{s_i | i \in \{1, \ldots, l\}, s_i = a_k\}|$

Given two sequences $S = (s_1, \ldots, s_l)$ and $T = (t_1, \ldots, t_r)$, **prove** or **disprove**:

(a) The Manhattan Distance $L_1(H(S), H(T))$ is a lower bound for the Edit Distance $D_{edit}(S, T)$.

(b) The modified Manhattan Distance

$$D(H(S), H(T)) = \sum_{i=1}^{n} \begin{cases} h_i(S) - h_i(T) & , \quad if\ h_i(S) > h_i(T) \\ 0 & , \quad else \end{cases}$$

is a lower bound for the Edit Distance $D_{edit}(S, T)$.

Exercise 9-2 Edit Distance and LCSS

Given two sequences: **CLASSIFY** and **CLUSTER**, compute the edit distance and the longest common subsequence similarity using dynamic programming way.

		C	L	A	S	S	I	F	Y
	0								
C									
L									
U									
S									
T									
E									
R									

		C	L	A	S	S	I	F	Y
	0								
C									
L									
U									
S									
T									
E									
R									