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

Knowledge Discovery in Databases II

SS 2017

Exercise 4: Sequential Data

Exercise 4-1 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 4-2 Implementing Edit Distance (Optional)

Compute the edit distance between the words **CLASSIFICATION** and **CLUSTERING** by implementing the dynamic programming approach introduced in the lecture.

Exercise 4-3 Normalized Time Series

(a) For a given time series $X = (3, 5, 10, 4, 1, 7, 7, 9, 1, 3)$, compute the z-score normalization $\hat{X}$ of X.

(b) Prove or disprove the following statement for a z-score normalized time series $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)$:

$$\sum_{i=1}^{n} \hat{x}_i = 0$$

(c) Prove or disprove the following statement for a z-score normalized time series $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)$:

$$\sum_{i=1}^{n} \hat{x}_i^2 = n$$

Exercise 4-4 Uniform and Dynamic Time Warping

Given the following two time series: $X = (3, 5, 9, 2, 3, 6, 3)$ and $Y = (3, 4, 6, 10, 1, 3, 2, 7, 4)$, compute the following distances:

(a) Uniform Time Warping Distance D^2_{UTW}

(b) Dynamic Time Warping DTW^2

(c) k-Dynamic Time Warping Distance D^2_{k-DTW} where $k = 3$ (Optional)

Visualize the optimal alignment between the time series.