Kiel University
Christian-Albrechts-Universität zu Kiel

Faculty of Engineering

Intelligent Systems

Excersice 5 – Similarities

Simon Reichhuber, Ingo Thomsen

January 6, 2020

University of Kiel, Winter Term 2020

TABLE OF CONTENT

Similartiy Measures

1. A

A. **What is the Minkowski distance? How can it be applied on time series data?**

B. **Calculate the distance of the two time series in the Figure according to the following distance measures:**
   - Manhattan distance
   - Euclidean distance
   - Cosine distance
   - Hamming distance

C. Find two time series, similiar as in the Figure with a Cosine distance of 0.

1. A Minkowski Distance

- Element-wise distance
- For time series data only applicable on series of same length (adapt interpolation) or on their feature vectors
  - $D_p(X, Y) = (\sum_{i=1}^{N} |x_i - y_i|^p)^{\frac{1}{p}}$
- p = 1 - Manhattan Distance
- p = 2 - Euclidean Distance
- ...

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

A. What is the Minkowski distance? How can it be applied on time series data?

B. Calculate the distance of the two time series in the Figure according to the following distance measures:
- Manhattan distance
- Euclidean distance
- Cosine distance
- Hamming distance

C. Find two time series, similiar as in the Figure with a Cosine distance of 0.

1. B Similarity calculations I

time series A: (3,0,4,1)
time series B: (1,4,2,5)
time series C: (-1,-4,-2,-5)

- Manhattan distance:
  - $D_{A,B} = |3-1| + |0-4| + |4-2| + |1-5| = 12$
  - $D_{A,C} = |3+1| + |0+4| + |4+2| + |1+5| = 20$
  - $D_{B,C} = |1+1| + |4+4| + |2+2| + |5+5| = 24$
- Euclidean distance:
  - $D_{A,B} = \sqrt{|3-1|^2 + |0-4|^2 + |4-2|^2 + |1-5|^2} = 6.324$
  - $D_{A,C} = \sqrt{|3+1|^2 + |0+4|^2 + |4+2|^2 + |1+5|^2} = 10.198$
  - $D_{B,C} = \sqrt{|1+1|^2 + |4+4|^2 + |2+2|^2 + |5+5|^2} = 13.564$

1. B Similarity calculations II

time series A: (3,0,4,1)
time series B: (1,4,2,5)
time series C: (-1,-4,-2,-5)

- Cosine distance:
  - $D_{A,B} = \frac{3*1+0*4+4*2+1*5}{\sqrt{3^2+0^2+4^2+1^2}*\sqrt{1^2+4^2+2^2+5^2}} = 0.463$
  - $D_{A,C} = -0.463$
  - $D_{B,C} = -1$
- Hamming distance:
  - $D_{A,B} = 4$
  - $D_{A,C} = 4$
  - $D_{B,C} = 4$

A. What is the Minkowski distance? How can it be applied on time series data?
B. Calculate the distance of the two time series in the Figure according to the following distance measures:
   - Manhattan distance
   - Euclidean distance
   - Cosine distance
   - Hamming distance
C. **Find two time series, similiar as in the Figure with a Cosine distance of** 0**.**

1. C Cosine distance

The Cosine distance is between -1 (exact mirror image) and 1 (exact same). The value 0 represents orthogonality (decorrelation)and values in between gradually similarity or non similarity.

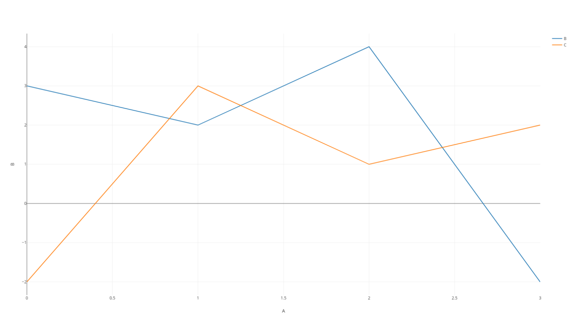


Figure 1: Cosinus distance of 0

Dynamic Similarity Measures

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

A. **What is the benefit of dynamic similarity measures?**
B. Calculate the *LCSS* of the the following two series:
   - Sequenz A = z,e,i,t,r,e,i,h,e
   - Sequenz B = r,e,i,t,z,e,i,t
C. Explain the steps of the *LCSS* on time series in your own words.
D. In the Figure, there are given two time series. Calculate the *DTW* path with the means of a *DTW* matrix and the backtracking algorithm by using the distance $|x - y|^2$
E. Which problems can arise from backtracking? Which conditions guarantee reasonable paths during backtracking?

2. A Benefit of dynamic similarity measures?

- Customised similarity measures for time series regarding dynamic relations
- Time series data of unequal length comparable
- Dynamical adaptation of various scale or translational variations

A. What is the benefit of dynamic similarity measures?

B. Calculate the *LCSS* of the the following two series:

- Sequenz A = z,e,i,t,r,e,i,h,e
- Sequenz B = r,e,i,t,z,e,i,t

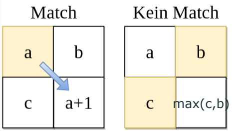
B. Explain the steps of the *LCSS* on time series in your own words.

C. In the Figure, there are given two time series. Calculate the *DTW* path with the means of a *DTW* matrix and the backtracking algorithm by using the distance $|x - y|^2$

D. Which problems can arise from backtracking? Which conditions guarantee reasonable paths during backtracking?

1) Calculate the *LCSS* matrix according to the following rules:

		z	e	i	t	r	e	i	h	e
		0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	1	1	1	1	1
e	0	0	1	1	1	1	2	2	2	2
i	0	0	1	2	2	2	2	3	3	3
t	0	0	1	2	3	3	3	3	3	3
z	0	1	1	2	3	3	3	3	3	3
e	0	1	2	2	3	3	4	4	4	4
i	0	1	2	3	3	3	4	5	5	5
t	0	1	2	3	4	4	4	5	5	5



2. B LCSS II

2) Find a way from start to end under compliance of the following rules:

- Areas shall only be left via bridges
- Allowed moving sequences are: **leftwards then upwards** or **upwards then leftwards** (must be consistent).

		z	e	i	t	r	e	i	h	e
		0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	1	1	1	1	1
e	0	0	1	1	1	1	2	2	2	2
i	0	0	1	2	2	2	2	3	3	3
t	0	0	1	2	3	3	3	3	3	3
z	0	1	1	2	3	3	3	3	3	3
e	0	1	2	2	3	3	4	4	4	4
i	0	1	2	3	3	3	4	5	5	5
t	0	1	2	3	4	4	4	5	5	5

2. B LCSS III

		z	e	i	t	r	e	i	h	e
		0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	1	1	1	1	1
e	0	0	1	1	1	1	2	2	2	2
i	0	0	1	2	2	2	2	3	3	3
t	0	0	1	2	3	3	3	3	3	3
z	0	1	1	2	3	3	3	3	3	3
e	0	1	2	2	3	3	4	4	4	4
i	0	1	2	3	3	3	4	5	5	5
t	0	1	2	3	4	4	4	5	5	5

Result: e,i,t,e,i

A. What is the benefit of dynamic similarity measures?
B. Calculate the *LCSS* of the the following two series:
   - Sequenz A = z,e,i,t,r,e,i,h,e
   - Sequenz B = r,e,i,t,z,e,i,t
C. **Explain the steps of the *LCSS* on time series in your own words.**
D. In the Figure, there are given two time series. Calculate the *DTW* path with the means of a *DTW* matrix and the backtracking algorithm by using the distance $|x - y|^2$
E. Which problems can arise from backtracking? Which conditions guarantee reasonable paths during backtracking?

Explain the steps of the *LCSS* on time series in your own words.

- **Step 1:** Atomic Matching
  - Compare subsequence of same length with sliding window
- **Step 2:** Merge subsequences
  - Merging conditions:
  - 1. No overlapping and distance less or equal *n*
  - 2. Overlapping on both time series by the same length
- **Step 3:** Find the largest common subsequence
  - 1. Merged subsequences must be non overlapping
  - 2. The total length of all subsequences is bounded

A. What is the benefit of dynamic similarity measures?
B. Calculate the *LCSS* of the the following two series:
   - Sequenz A = z,e,i,t,r,e,i,h,e
   - Sequenz B = r,e,i,t,z,e,i,t
C. Explain the steps of the *LCSS* on time series in your own words.
D. **In the Figure, there are given two time series. Calculate the *DTW* path with the means of a *DTW* matrix and the backtracking algorithm by using the distance** $|x - y|^2$
E. Which problems can arise from backtracking? Which conditions guarantee reasonable paths during backtracking?

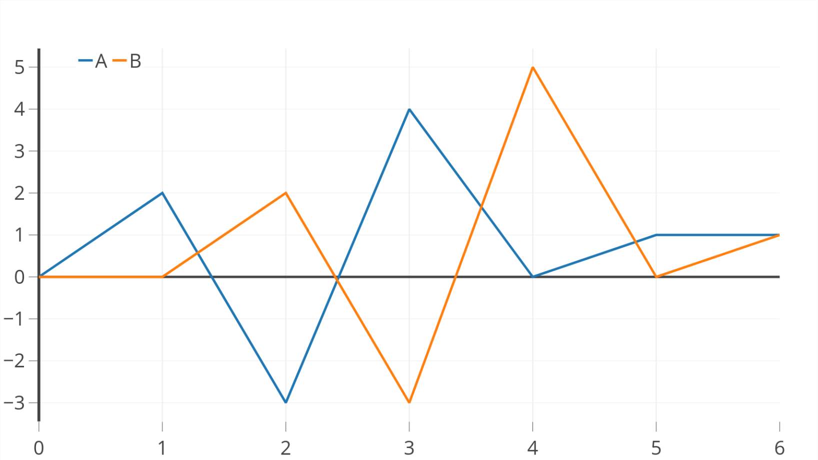


Figure 2: zeitreihen

Distance matrix:

	0	**2**	**-3**	**4**	**0**	**1**	**1**
0	0	4	9	16	0	1	1
0	0	4	9	16	0	1	1
2	4	0	25	4	4	1	1
-3	9	25	0	49	9	16	16
5	25	9	64	1	25	16	16
0	0	4	9	16	0	1	1
1	1	1	16	9	1	0	0

Table 1: Distanz Matrix

Accumulated distance matrix

	0	**2**	**-3**	**4**	**0**	**1**	**1**
0	0	4	13	29	29	30	31
0	0	4	13	29	29	30	31
2	4	0	25	17	21	22	23
-3	13	25	0	49	26	37	38
5	38	22	64	1	26	42	53
0	38	26	31	17	1	2	3
1	39	27	42	26	2	1	1

Table 2: Accumulated distance matrix

Backtracking path: [[6, 6], [5, 6], [4, 5], [3, 4], [2, 3], [1, 2], [0, 1], [0, 0]]

A. What is the benefit of dynamic similarity measures?
B. Calculate the *LCSS* of the the following two series:
   - Sequenz A = z,e,i,t,r,e,i,h,e
   - Sequenz B = r,e,i,t,z,e,i,t
C. Explain the steps of the *LCSS* on time series in your own words.
D. In the Figure, there are given two time series. Calculate the *DTW* path with the means of a *DTW* matrix and the backtracking algorithm by using the distance $|x - y|^2$
E. Which problems can arise from backtracking? Which conditions guarantee reasonable paths during backtracking?

- **Problem:** Path can degenerate
  - Optimal path: along the diagonal; worst case: along the border
- **Solution:** DTW path is ruled by side conditions
  - **Side conditions**: Path has a fixed start end end point. First and last data point has to be equal.
  - **Continuity**: Every wrapping path is continuous, i.e. every single element of the warping path is taken from one of both time series.
  - **Monotonicity:** The temporal ordering must not be violated.
  - **Additional side conditions to harm deviations from the diagonal path possible!**

Segmentation

3. A

A. **Name four different application scenarios where segmentation of time series plays a role.**
B. What is the difference between offline learning and online learning?
C. Name three criteria for segmentation.
D. Explain the offline and online segementation techniques for segmentation in your own words.

Time series often represent sequences of discrete segments. For different domains data can be segmented by:

1. Different events in the stock market (tax wars with the US, war in the Near East)
2. Single words or characters in recordings of handwritting
3. The speaker of recordings in a conference
4. State of motion found in acceleration sensor data (standing, walking, running, sleeping, etc.)

3. B

A. Name four different application scenarios where segmentation of time series plays a role.
B. What is the difference between offline learning and online learning?
C. Name three criteria for segmentation.
D. Explain the offline and online segementation techniques for segmentation in your own words.

3. B Offline vs. Online

Offline:

- Global view possible: The full time series is available

Online:

- Only local view: Sequence possibly of infinite length (or for example the recording or receiving of the data stream is not yet finished while segmenting)

3. C

A. Name four different application scenarios where segmentation of time series plays a role.
B. What is the difference between offline learning and online learning?
C. **Name three criteria for segmentation.**
D. Explain the offline and online segementation techniques for segmentation in your own words.

3. C Segmentation criteria

1. According a certain duration (e.g. sampling with a frame rate of 1 sec, segmentation summarize the data to pieces of one day each)
2. Meaning of certain values (e.g. exceeding a predefined threshold)
3. Maining of certain segments (e.g. fixed number of phases repeated during a heartbeat)

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

A. Name four different application scenarios where segmentation of time series plays a role.
B. What is the difference between offline learning and online learning?
C. Name three criteria for segmentation.
D. Explain the offline and online segementation techniques for segmentation in your own words.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Offline:

- **Top down approach:**
  - Step 1: Separate whole time series in two subsegments with minimal approximation error.
  - Step 2: Separate the subsegments with an approximation error above a predefined threshold further into smaller subsegments by repeating Step 1 and 2.
- **Bottom up approach:**
  - Analogously to the top down approach, but start with small subsegments and merge them if the approximation error of the merged segement is below a predefined threshold.

Online:

- **Equidistant:**
  - Segmentation into segments of equal length
  - Applicable if the semantics of the time series is known
  - *For example*: One measuring per hour
- **Sliding Window:**
  - Sliding window will be shifted over the time series and segmented according a certain criterion
  - *For example*: until the approximation error exceeds a certain threshold
  - *Or*: Until the gradient changes its sign and the curvature exceeds a certain threshold

1. **Growing Window:**
   - Segments will be expanded stepwise
   - Segmentation until a certain criterion is fulfilled
   - *For example:* Ongoing approximation creating new segments only if the approximation error exceeds threshold.
2. **Sliding Window and Bottom-Up (SWAB):**
   - Combination of sliding window, growing window, and bottom up strategy
   - Offline approach (bottom up segmentation) is now applicable also for data streams

Comparing Time Series with Python