Università di Pisa	A.A. 2018-2019

Data Mining II April 4th, 2019

Exercise 1 - Sequential patterns **(16 points)**

A) **(8 points)** Given the following input sequence

<	{A}	{A,B,C}	{C}	{A,E,F}	{G}	{A}	{A,B}	{A,E} >
	t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7

show all the occurrences (there can be more than one or none, in general) of each of the following subsequences in the input sequence above. Repeat the exercise twice: the first time considering no temporal constraints (left column): the second time considering max-gap = 2 (i.e. gap <= 2, right column). Each occurrence should be represented by its corresponding list of time stamps, e.g..: <0,2,3> = <t=0, t=2, t=3>.

Solutions are highlighted in yellow:

	Occurrences	*Occurrences with max-gap = 2*
***ex.:* <{B}{E}>**	**<1,3> <1,7> <6,7>**	**<1,3> <6,7>**
w_1 = <{C} {A} {E} >	**<1,3,7> <1,5,7> <1,6,7> <2,3,7> <2,5,7> <2,6,7>**	**none**
w_2 = <{E}{A}{E}>	**<3,5,7> <3,6,7>**	**<3,5,7>**
w_3 = <{A} {A} {A}>	**NOT NEEDED**	**<0,1,3> <1,3,5> <3,5,6> <3,5,7><5,6,7>**

B) **(7 points)** For a given dataset of sequences, the GSP algorithm at the **second iteration** found the frequent 3-sequences shown below:

Frequent 3-sequences

{ A } -->{ C } -->{ A } { A } -->{ C , D } { A , B } -->{ A } { A , B } -->{ C } { A , B } -->{ D }	{ B } -->{ C } -->{ A } { B } -->{ C , D } { C } -->{ C , D } { C , D } -->{ A } { D } -->{ C , D }

Show which candidates the GSP will generate at the **third** iteration, and which of them are removed by pruning.

Answer:

Candidates

{ A } { C , D } { A } **PRUNED**
{ A , B } { C } { A }
{ A , B } { C , D }
{ B } { C , D } { A } **PRUNED**
{ C } { C , D } { A } **PRUNED**
{ D } { C , D } { A } **PRUNED**

C) (1 point) Which of the candidates generated above are pruned if we are using GSP with a max-gap constraint?

Answer: exactly the same as in point B), since all significant subsequences if the pruned candidates are contiguous.

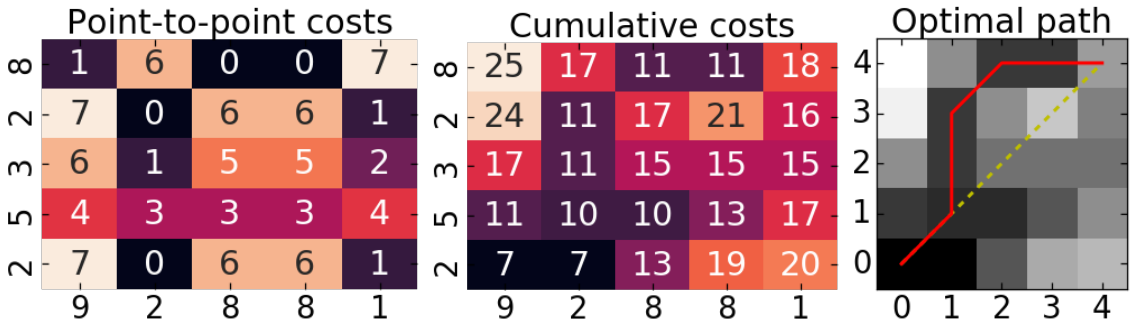
Exercise 2 - Time series / Distances **(16 points)**

A) **(6 points)** Given the following input time series:

t1	< 9, 2, 8, 8, 1 >
t2	< 2, 5, 3, 2, 8 >

compute the distance between "t1" and "t2", using the DTW with distance between points computed as d(x,y) = |x – y|.

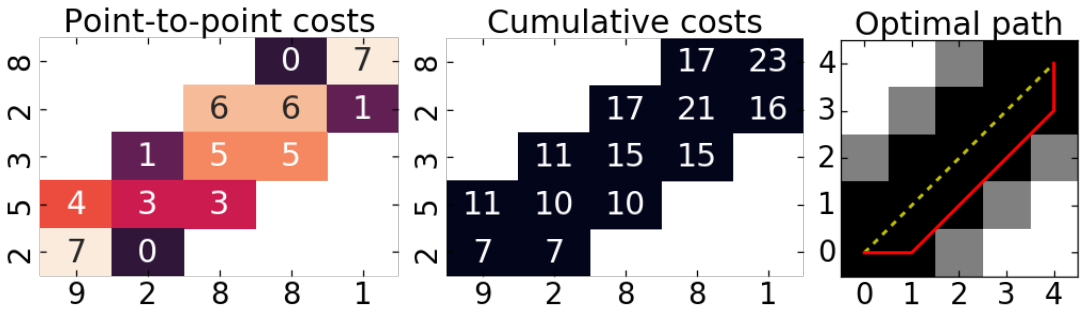
Answer:



Result: 18

B) **(3 points)** If we repeat the computation of point (A) above, this time with a Sakoe-Chiba band of size r=1, might the result change? In positive case, recompute the value of DTW.

Answer: Yes. Because the DTW optimal path is outside the band. Result:



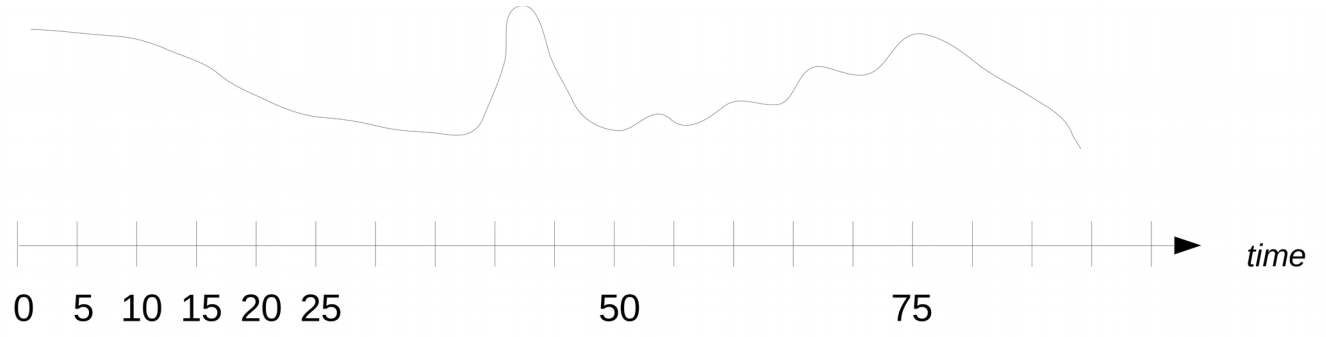
C) **(0 points)** We are working with a dataset of "uncertain time series", i.e. for each point of a time series we know an interval of its possible values but not the real one. E.g. < [2,7], [0,3],[3,10] > means that the real time series is < X1, X2, X3 > with $X1 \in [2,7], X2 \in [0,3]$ and $X3 \in [3,10]$. Can we modify the DTW algorithm to compute the "uncertain DTW" between two uncertain time series, yielding the range of possible values that the DTW can have on the real (single valued) time series? How?

Answer: Yes. Replace values with intervals, starting from the point-to-point matrix. E.g. | [3,4] - [5,8] | = [1,5]. Then, when computing accumulated costs, we consider the arrival cost in the best and in the worst case. Notice that the uncertain DTW will yield the same result as normal DTW when intervals are all of the form [x,x], i.e. data is not uncertain.

D) **(6 points)** A large city is divided into several hundred sections, and for each of them we know the average price of houses (€ per square meter), recorded every month in the last 50 years. We want to create clusters of sections whose price evolved in a similar way. What kind of similarity measure and clustering method would you adopt? Motivate your answer.

Answer: This is a very open question and several solutions are possible and acceptable. Here is one: in total we have time series of length 600 points, yet a montly update of prices is probably unnecessary, and the real changes happen 1-2 times per year. Lengths reduce to 100, and it looks reasonable to compare time series by shape. DTW with a tight time constraint (e.g. 4 = 2 years) is probably better, since different areas might react to external factors with slight delays. Clustering might be performed with virtually any algorithm, though probably a hierarchical one is preferrable, in order to obtain a dendogram and decide later what might be a meaningful number of clusters. Yet, results might suggest to move to other approaches.

E) **(1 point)** We want to analyze the following time series in search of anomalies. In particular, we consider two different prediction methods as basis: (A) a next value is equal to the average of the last 10 values of the time series; (B) the next value is the linear interpolation of the previous 10 values, where the weights of the interpolation are computed on the whole history seen so far (i.e. all the time series from the start to the present point). At which time instants the two methods will reveal an anomaly on the given time series, assuming that the maximum discrepancy tolerated is the standard deviation of the last 10 values?



Answer: Though difficult to identify exactly the outcome, Approach A) will mostly find values that show quick changes w.r.t. previous ones, in particular in the range 40-45, probably also at 65-70 and at 75. Approach B) will most likely find changes in direction, e.g. 35-40, 45-50 and 75.