Università di Pisa	A.A. 2016-2017

Data Mining II April 7th, 2017

Exercise 1 - Sequential patterns **(12 points)**

A) **(6 points)** Given the following input sequence

<	{A}	{B,F}	{E}	{A,B}	{A,C,D}	{F}	{B,E}	{C,D}	>
	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 = 3 (i.e. gap <= 3, 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 = 3*
***ex.:* <{B}{E}>**	**<1,2> <1,6> <3,6>**	**<1,2><3,6>**
w_1 = <{A} {F} {E} >	**<0,1,2> <0,1,6> <0,5,6> <3,5,6> <4,5,6>**	**<0,1,2> <3,5,6> <4,5,6>**
w_2 = <{A}{D}>	**<0,4> <0,7> <3,4> <3,7> <4,7>**	**<3,4> <4,7>**
w_3 = <{C,D}>	**<4> <7>**	**<4> <7>**

B) **(6 points)** Given the following dataset on the left and a minimum support threshold set to 40%, the GSP algorithm at the **second iteration** found the frequent 3-sequences on the right:

Dataset

{ A B } → { A } → { C } → { D }
{ A } → { B } → { C D } → { C }
{ D } → { C } → { B C } → { D }
{ A B } → { D } → { C } → { C D } → { E }

Frequent 3-sequences

{ A B } → { C }	{ A } → { D } → { C }
{ A B } → { D }	{ B } → { C } → { C }
{ A } → { C D }	{ B } → { C } → { D }
{ B } → { C D }	{ B } → { D } → { C }
{ A } → { C } → { C }	{ D } → { C } → { C }
{ A } → { C } → { D }	{ D } → { C } → { D }

Simulate the execution of GSP at the **third** iteration, showing all the phases, from candidate generation to the resulting frequent 4-sequences.

Candidates list (striked = pruned)	support	
{ A B } → { C D }	1/4	
{ A B } → { C } → { C }	1/4	
{ A B } → { C } → { D }	2/4	← Only frequent 4-sequence
{ A B } → { D } → { C }	1/4	
{ A } → { D } → { C } → { C }	1/4	
~~{ A } → { D } → { C } → { D }~~	----	
{ B } → { D } → { C } → { C }	1/4	
~~{ B } → { D } → { C } → { D }~~	----	

Exercise 2 - Time series / Distances **(12 points)**

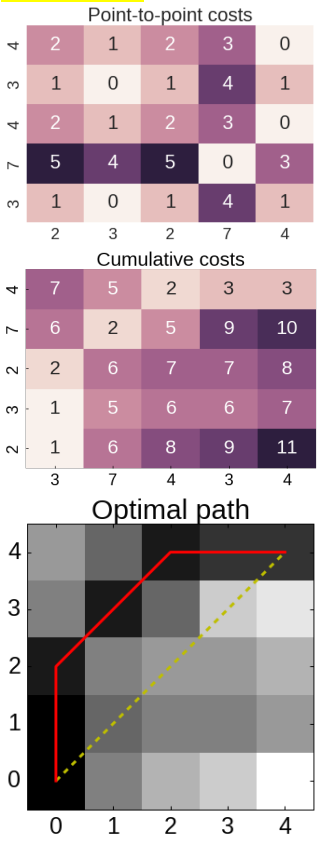
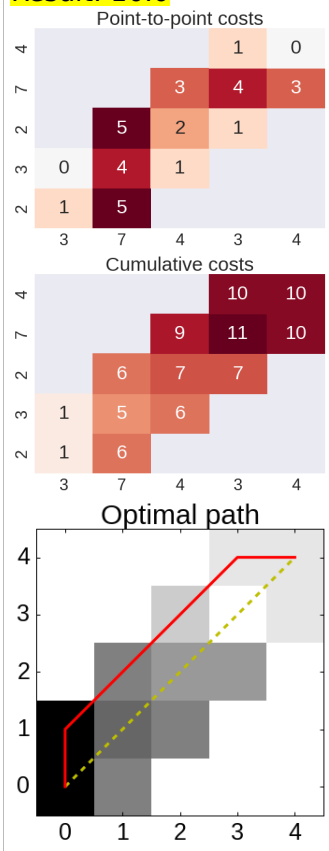
1) **(10 points)** Given the following input time series:

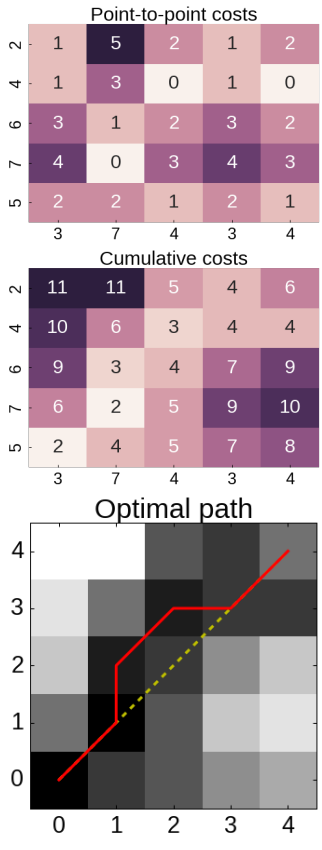
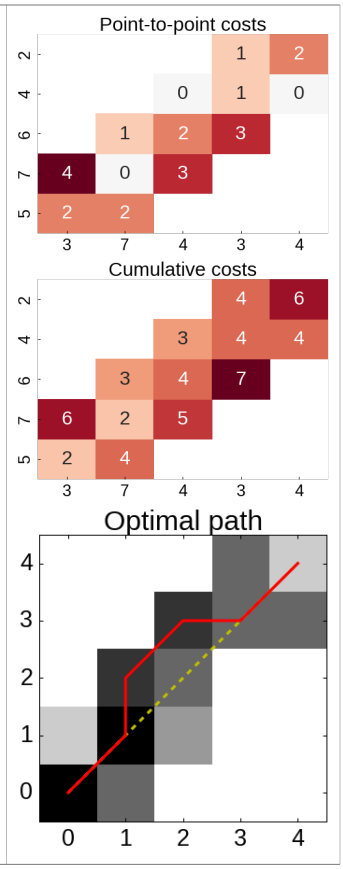
t1	< 2, 3, 2, 7, 4 >
t2	< 5, 7, 6, 4, 2 >

and

q	< 3, 7, 4, 3, 4 >

compute the distance between "q" and the two time series "t1" and "t2", using: (i) Euclidean distance, (ii) DTW, and (iii) DTW with Sakoe-Chiba band of size r=1 (i.e. all cells at distance <= 1 from the diagonal are allowed).

	Euclidean	DTW	DTW with band r=1
D(q, t1)	Sqrt(37) = 6.08...	Result: 3.0 	Result: 10.0 
D(q, t2)	Sqrt(13) = 3.60...	Result: 6.0	Result: 6.0

For each case, show the computation performed, if needed.

2) **(2 points)** We would like to cluster the customers of a retail seller based on their time series of purchases, represented by weekly aggregates over 10 years. What distance between customers' time series would you apply? Why?

We have first to decide if it is better to apply a shape-based distance (Euclidean or DTW) or a structure-based one. 10 years of data means that our timeseries are quite long and contain >500 values, which seems to suggest a structure-based approach (choosing a shape-based approach is not out of the question, but one should provide some motivation for that). Any specific approach is in principle possible, for instance using some fixed features (mean, variance, etc.), using a frequency decomposition (Fourier transform or the like), or even the default solution based on compression.

Exercise 3 - Analysis process & CRISP-DM **(8 points)**

A large retail sales company is going to introduce a new product in a few months. In order to better advertise it, the company decides to send personalized invitations to try the product to 1000 customers that are most likely to appreciate the product and start to buy it (… and hopefully spread the word).

How can they select these well-targeted customers?

The company has hundreds of thousands customers, deals with several thousands different products, and has more than 100 sales points (e.g. supermarket, etc.). The company is willing to use all the data they collected in the last 5 years, in particular:

- All demographic information on its customers

- The details of all transactions (timestamp, product, cost, customer ID, whether the product was under promotion)
- The details of all products (product category and all the associated hierarchy of products, features of the product [bio, gluten free, cost, package type, etc.], etc.)

Briefly describe a project plan to answer the requests mentioned above, (loosely) following the CRISP-DM methodology. Clearly remark the choices and assumptions made in the process.

Key ideas: one possible way to approach the problem is to identify customers that in the past bought products similar to the new one, and then build a classification model for recognizing such customers. Alternatively, following a less direct approach, a customer segmentation on various attributes (demographic and purchase-based) can be performed, trying later to identify the segments/clusters that contain high percentages of customers who were positive to the products similar to the new one. All the project should be described through the CRISP-DM phases. In particular, the first one should clearly state what we want to do in terms of DM analyses.