Università di Pisa	A.A. 2016-2017

Data Mining II **July 4th, 2017**

Exercise 1 - Sequential patterns **(6 points)**

a) **(3 points)** Given the following input sequence

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

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 = 4 (i.e. gap <= 4, 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>.

	Occurrences	*Occurrences with max-gap =4*
w_1 = <{B} {E} >		
w_2 = <{B}{D}>		
w_3 = <{F}{B}{C,D}>		

Answer:

	Occurrences	*Occurrences with max-gap =4*
w_1 = <{B} {E} >	**<0,4> <0,5> <2,4> <2,5>**	**<0,4> <2,4> <2,5>**
w_2 = <{B}{D}>	**<0,3> <0,6>** **<2,3> <2,6>** **<5,6>**	**<0,3>** **<2,3> <2,6>** **<5,6>**
w_3 = <{F}{B}{C,D}>	**<0,2,3> <0,2,6> <0,5,6>** **<3,5,6>**	**<0,2,3> <0,2,6>** **<3,5,6>**

b) **(3 points)** Simulate the execution of the GSP algorithm on the following dataset of sequences, assuming a minimum support threshold of 60%.

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

Answer:

Output Sequential patterns

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

remark: also the following 3-sequences were generated, but then discarded:
{ A } -> { C } -> { D } ←removed by PRUNING
{ B } -> { C } -> { C } ←infrequent
{ B } -> { C } -> { D } ←infrequent
{ C } -> { C } -> { C } ←infrequent

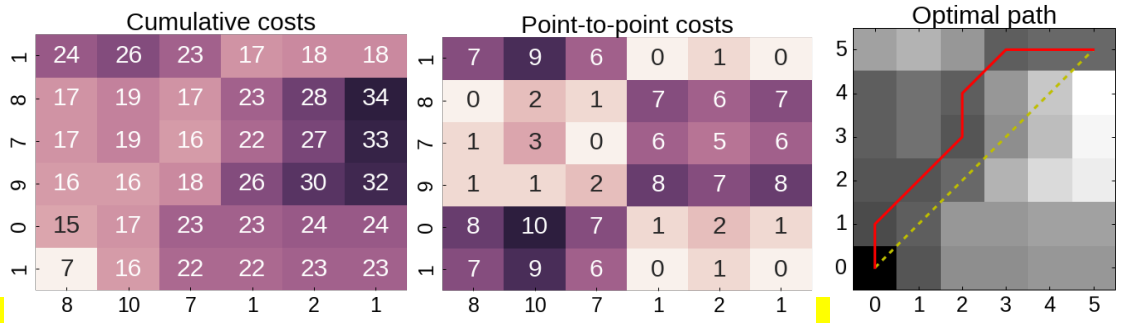
Exercise 2 - Time series / Distances (**6 points)**

Given the following time series:

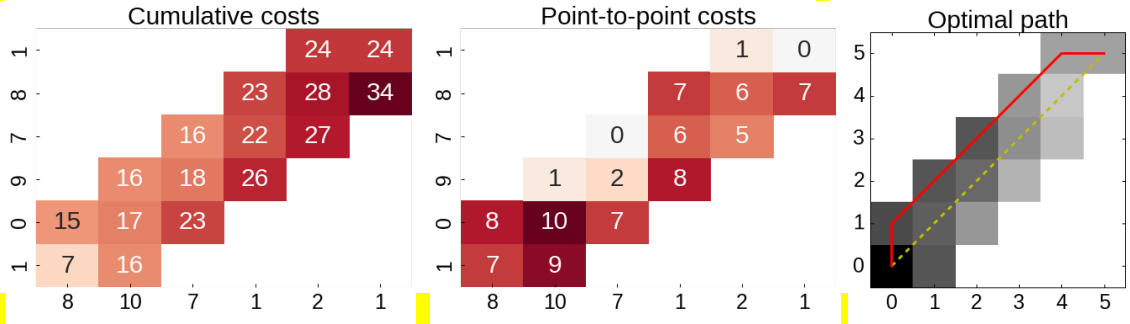
t = < 1, 0, 9, 7, 8, 1 >
q = < 8, 10, 7, 1, 2, 1 >

compute (i) their DTW, and (ii) their DTW with Sakoe-Chiba band of size r=1 (i.e. all cells at distance <= 1 from the diagonal are allowed). Show the cost matrices and the optimal paths found.

Answer:



DTW: 18



DTW r=1: 17

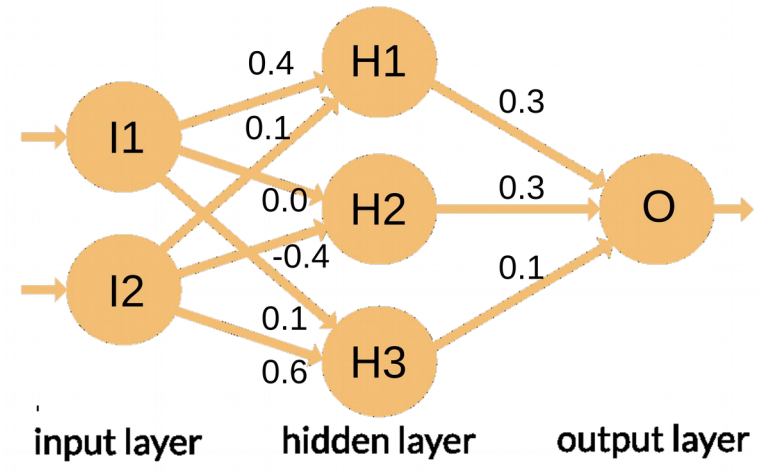
Exercise 3 - Analysis process & CRISP-DM **(4 points)**

A large toy manufacturer is going to produce a new board game. They received several alternative proposals from their game designers, too many to scrutinize manually, therefore they decided to choose the most appealing candidate (the one which has best chances of being successful) based on data.
The data available include all previous games produced, with their characteristics (number of players, minimum age, etc.) and the performances on the market (how much it sold, in which countries, etc.).
Briefly describe a project plan to help the company, (loosely) following the CRISP-DM methodology. Clearly remark the choices and assumptions made in the process.

Answer:
Key steps: after deciding what successful means (e.g. more than 2000 pieces are sold in Italy and at least 1000 in the rest of the world), we identify and synthesize the features describing each game. The rest of the process is basically identical to a target marketing case, where now we are targeting products instead of people.

Exercise 4 - Classification **(6 points)**

a) Neural Networks (**3 points**)
Given the neural network below (on the left), apply it to the test set provided (on the right). The weights are reported beside each connection, while the activation function is simply f(S) = sign(S), i.e. -1 for positive values, +1 for positive ones and 0 for S=0. For each case, show the output also of the nodes on the hidden layer.



I1	I2	O
+0	-1	
+1	+0	
-1	+1	
+1	+1	
+1	-1	

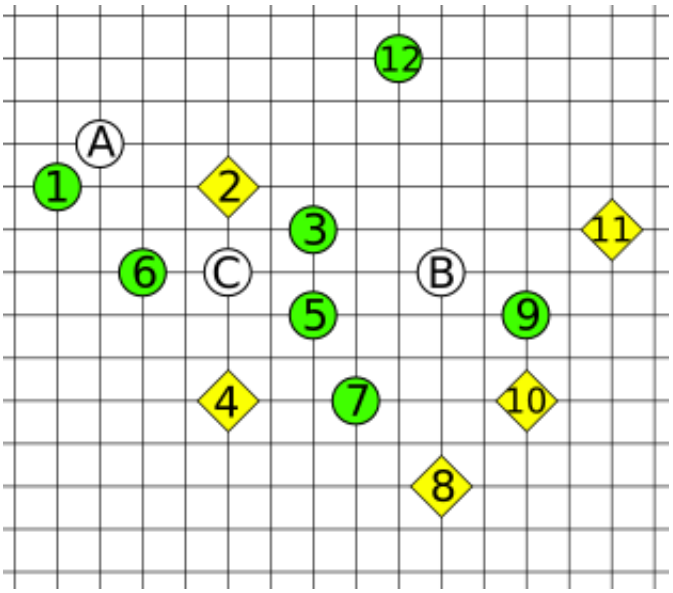
Answer:

I1	I2	O
+0	-1	-1
+1	+0	+1
-1	+1	-1
+1	+1	+1
+1	-1	+1

Input1	0	1	-1	1	1
Input2	-1	0	1	1	-1
H1	-1	1	-1	1	1
H2	1	0	-1	-1	1
H3	-1	1	1	1	-1
Output	-1	1	-1	1	1

b) k-NN (**3 points**)

Given the training set on the right, composed of elements numbered from 1 to 12, and labelled as circles and diamonds, use it to classify the remaining 3 elements (letters A, B and C) using a k-NN classifier with k=3.
For each point to classify, list the points of the dataset that belong to its k-NN set.
Notice: A, B and C belong to the test set, not to the training set. Also, the Euclidean distance should be used.



Answer:
kNN(A) = { 1, 2, 6 } → CIRCLE
kNN(B) = { 3, 5, 9 } → CIRCLE
kNN(C) = { 2, 3, 5, 6 } → CIRCLE

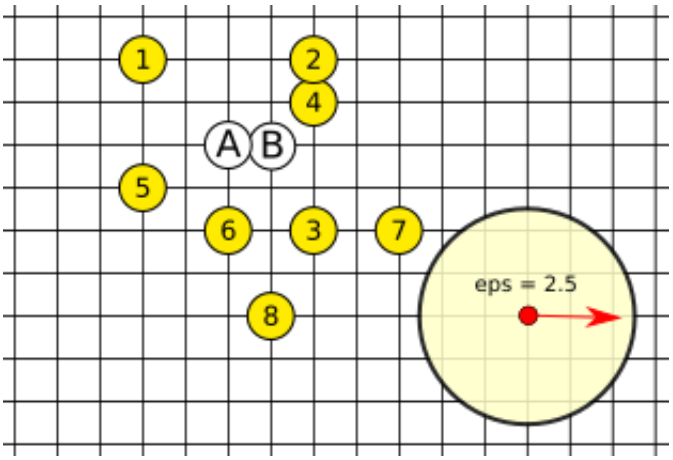
Exercise 5 - Outlier Detection **(6 points)**

Given the dataset of 10 points below, consider the outlier detection problem for points A and B, adopting the following three methods:

a) Distance-based: DB(ε,π) (**2 points**)
Are A and/or B outliers, if thresholds are forced to ε = 2.1 and π = 0.15 ? The point itself should not be counted.

b) Density-based: LOF (**2 points**)
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2 NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

c) Depth-based (**2 points**)
Compute the depth score of all points.



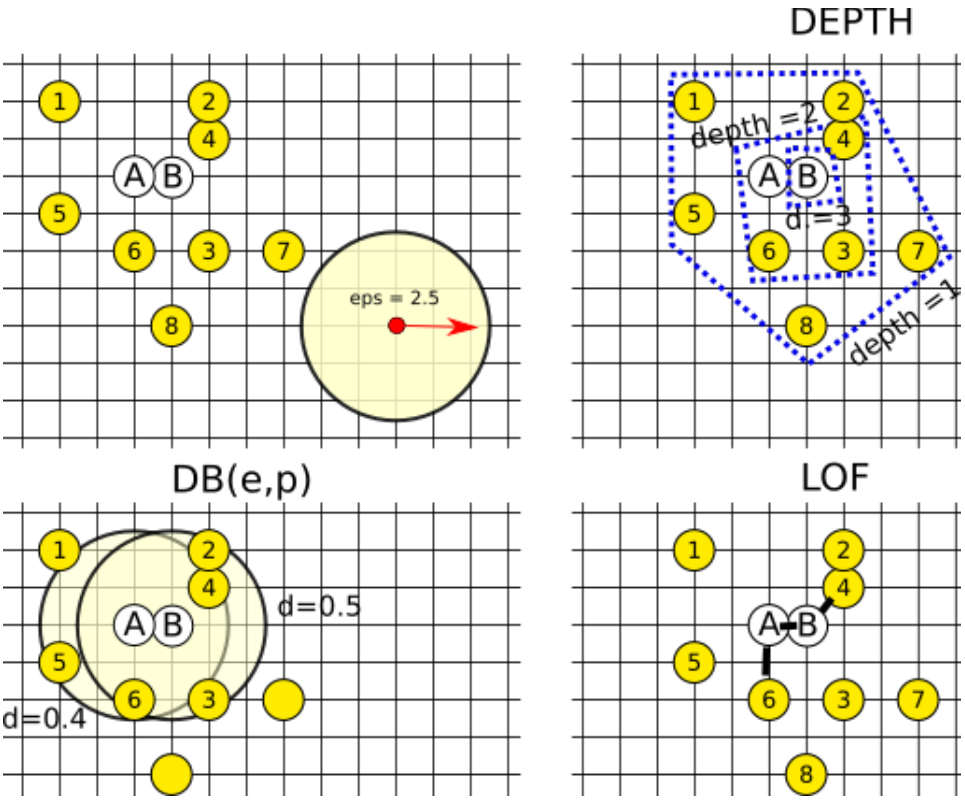
Answer:

For a) and c), see the figure below.
Solution for b):

LRD(A) = 1/ [(1 + 2)/2] = 0.666
LRD(B) = 1/ [(1 + √2)/2] = 0.828
LRD(6) = 1/ [(2 + 2)/2] = 0.500
LOF(A) = ([LRD(B) + LRD(6)]/2) / LRD(A) = [(0.828 + 0.500) / 2] / 0.666 = 1.003

LRD(4) = 1/ [(1 + √2)/2] = 0.828
LOF(B) = ([LRD(A) + LRD(4)]/2) / LRD(B) = [(0.666 + 0.828) / 2] / 0.828 = 0.902

Both are smaller or very close to 1, so they are most likely no outliers.



Exercise 3 - Validation **(3 points)**

ROC & AUC (**3 points**)

On a given test set below, our classification model provided the predictions and associated confidences reported on the "Predicted" column of the table. Draw the corresponding ROC curve and compute it Area Under the Curve. Show the process followed to achieve that.

Record	Real Class	Predicted	
row 1	Y	N	0.33
row 2	N	Y	0.98
row 3	N	Y	0.58
row 4	N	Y	0.65
row 5	N	Y	0.05
row 6	Y	Y	0.04
row 7	Y	Y	0.8
row 8	Y	Y	0.37
row 9	Y	N	0.62
row 10	N	N	0.95

Record	Real Class	Predicted		Score
row 1	Y	N	0.33	0.67
row 2	N	Y	0.98	0.98
row 3	N	Y	0.58	0.58
row 4	N	Y	0.65	0.65
row 5	N	Y	0.05	0.05
row 6	Y	Y	0.04	0.04
row 7	Y	Y	0.8	0.8
row 8	Y	Y	0.37	0.37
row 9	Y	N	0.62	0.38
row 10	N	N	0.95	0.05

SORTED		TPR	FPR	AUC partial	
Real Class	Score	0	0		
N	0.98	0	1	0	
Y	0.8	1	1	0	
Y	0.67	2	1	0	
N	0.65	2	2	2	
N	0.58	2	3	2	
Y	0.38	3	3	0	
Y	0.37	4	3	0	
N	0.05	4	4	4	
N	0.05	4	5	4	
Y	0.04	5	5	0	

AUC	**12**
Normalized	**0.48**