Università di Pisa	A.A. 2016-2017

Data Mining II June 13th, 2017

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

a) **(3 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 min-gap = 1 (i.e. gap > 1, 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 min-gap =1*
***ex.:* <{B}{E}>**	**<1,2> <1,6> <3,6>**	**<1,6><3,6>**
w_1 = <{A} {B} {E} >		
w_2 = <{B}{D}>		
w_3 = <{F}{E}{C,D}>		

Answer:

	Occurrences	*Occurrences with min-gap =1*
***ex.:* <{B}{E}>**	**<1,2> <1,6> <3,6>**	**<1,6><3,6>**
w_1 = <{A} {B} {E} >	**<0,1,2> <0,1,6> <0,3,6>**	**<0,3,6>**
w_2 = <{B}{D}>	**<1,4> <1,7> <3,4> <3,7> <6,7>**	**<1,4> <1,7> <3,7>**
w_3 = <{F}{E}{C,D}>	**<1,2,4> <1,2,7> <1,6,7> <5,6,7>**	**none**

b) **(3 points)** Running the GSP algorithm on a dataset of sequences, at the end of the second iteration it found the frequent 3-sequences on the left, and at the next iteration it generated (among the others) the candidate 4-sequences on the right. Which of the candidates will be **pruned**, and why?

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 }

Candidates

1. { A B } → { C D }
2. { A } → { D } → { C } → { D }
3. { B } → { D } → { C } → { D }
4. { A B } → { D } → { C }
5. { A B } → { C } → { D }

Answer:

Candidates

1. { A B } → { C D }
2. **{ A } → { D } → { C } → { D }** ← **PRUNED**
3. **{ B } → { D } → { C } → { D }** ← **PRUNED**
4. { A B } → { D } → { C }
5. { A B } → { C } → { D }

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

Given the following time series:

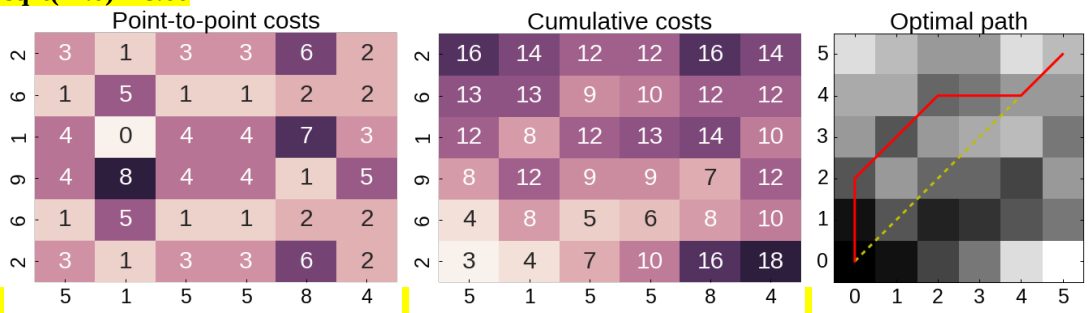
t = < 2, 6, 9, 1, 6, 2 >

q = < 5, 1, 5, 5, 8, 4 >

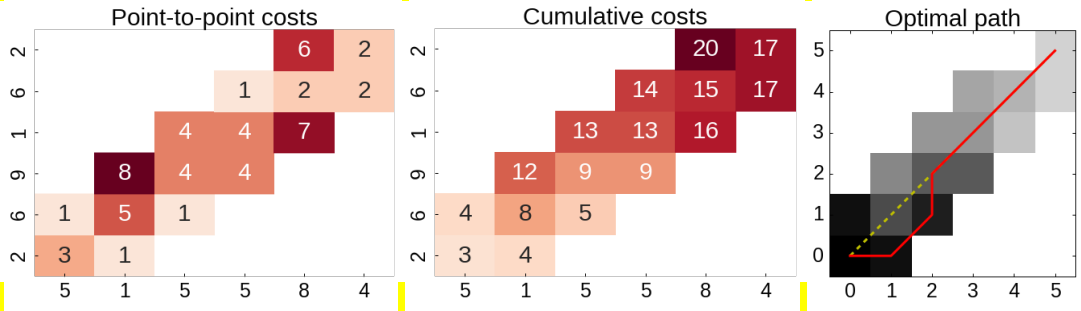
compute (i) their Euclidean distance, (ii) their DTW, and (iii) their DTW with Sakoe-Chiba band of size r=1 (i.e. all cells at distance <= 1 from the diagonal are allowed). For points (ii) and (iii) show the cost matrix and the optimal path found.

Answer:

Euclidean: sqrt(74.0) = 8.60



DTW: 14



DTW r=1: 17

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

A large telecom operator realized that in recent years it has been losing also several of its long-standing (more than 10-years) customers due to competitors. In order to fight this phenomenon, the company needs both to identify potential churners and understand what makes (or will make) them to leave. Notice that at this stage the operator is not interested in the recent customers – they will be processed apart, with different initiatives.

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.

The operator has the following information:

- All demographic information on its customers
- The details of all their transactions (calls, data traffic usage, etc.) and contracts (including type of contract, start date and termination)

Answer:

Key steps: select long-standing customers; label customers as churners vs. non-churners; extract features about customer activity (frequency of class, traffic data usage rates) and contracts (most frequent type? Variability?). Build a classification model – better a decision tree, for readibility reasons.

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

a) Naive Bayes (**3 points**)

Given the training set below, build a Naive Bayes classification model (i.e. the corresponding table of probabilities) using (i) the normal formula and (ii) using Laplace formula. What are the main effects of Laplace on the models?

A	B	class
no	green	N
no	red	Y
yes	green	N
no	red	N
no	red	Y
no	green	Y
yes	green	N

Answer:

Normal

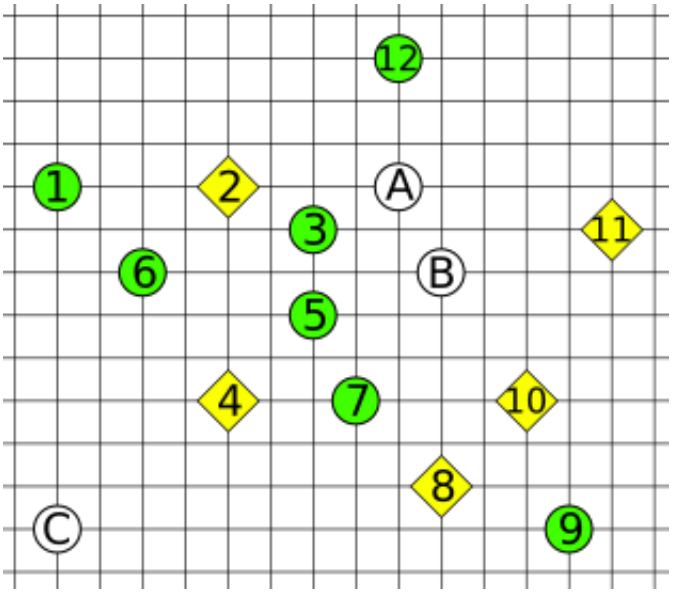
	Y	N		Y	N
	3	4		0.43	0.57
	A | Y	A | N		A | Y	A | N
yes	0	2	yes	0.00	0.50
no	3	2	no	1.00	0.50
	B | Y	B | N		B | Y	B | N
green	1	3	green	0.33	0.75
red	2	1	red	0.67	0.25

Laplace

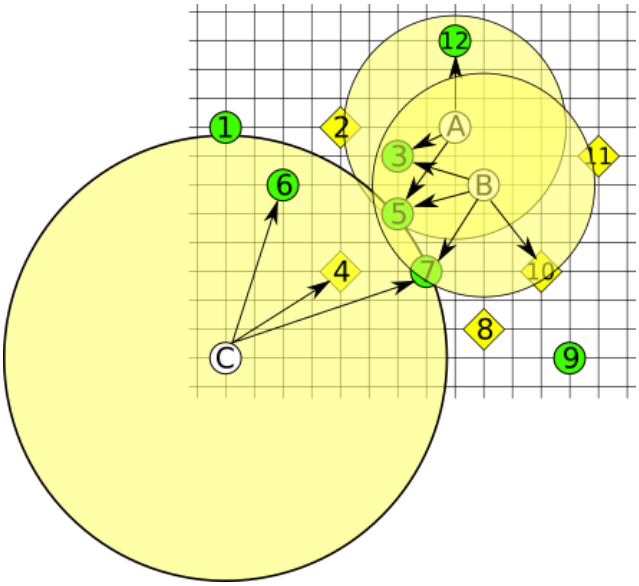
	Y	N		Y	N
	3	4		0.43	0.57
	A | Y	A | N		A | Y	A | N
yes	0	2	yes	0.20	0.50
no	3	2	no	0.80	0.50
	B | Y	B | N		B | Y	B | N
green	1	3	green	0.40	0.67
red	2	1	red	0.60	0.33

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.

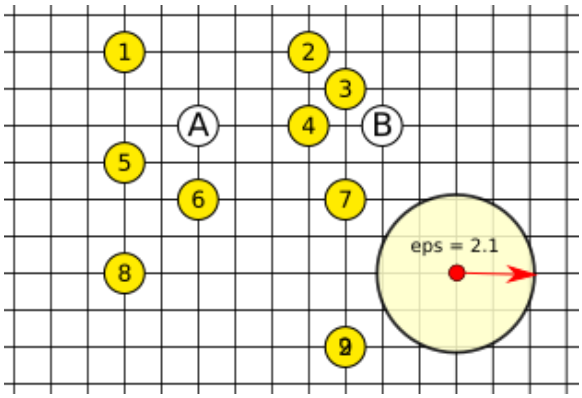


Answer:
kNN(A) = {3, 5, 12} → CIRCLE
kNN(B) = { 3, 5, 7, 10 } → CIRCLE
kNN(C) = { 4, 6, 7 } → CIRCLE



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:

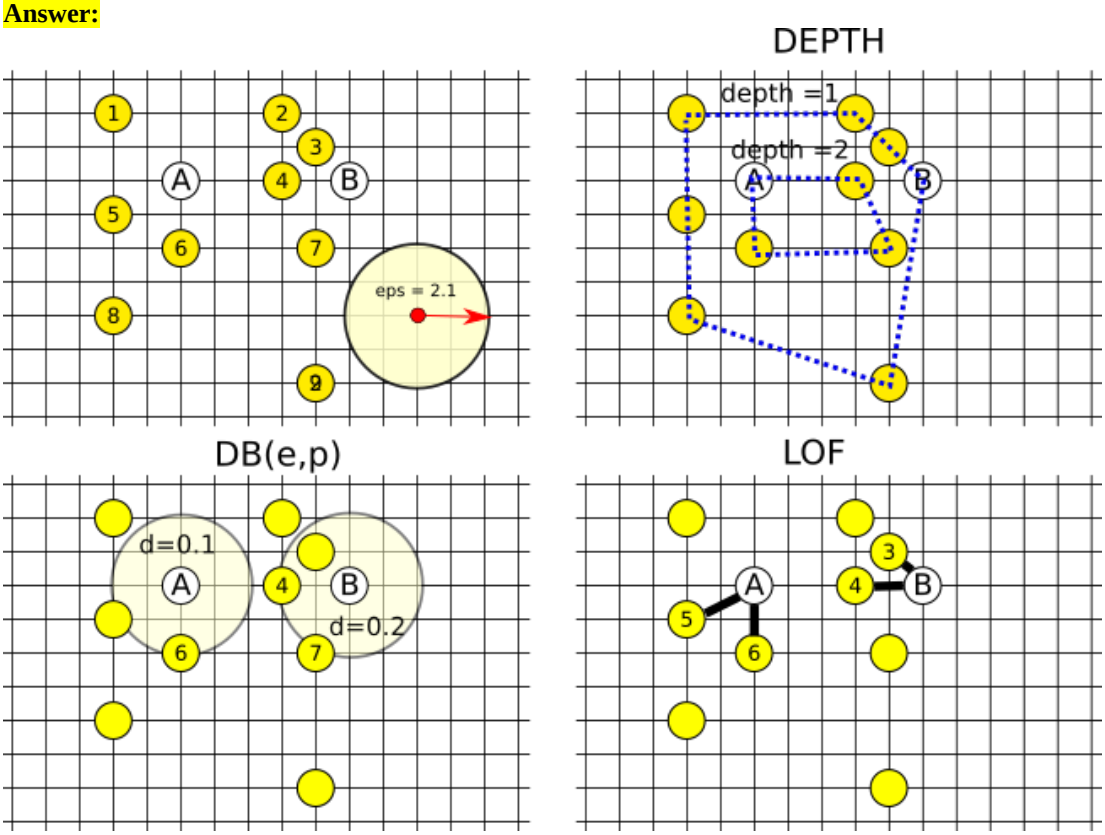


a) Distance-based: DB(ε,π) (**2 points**)
Are A and/or B outliers, if thresholds are forced to ε = 2.1 and π = 0.15 ? The point itself should not be counted.

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. Are A and/or B outliers of depth 1?

Answer:

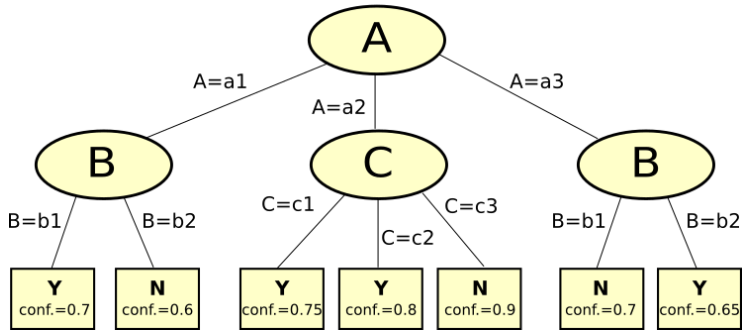


LRD(A) = 1/ [(2 + √5)/2] = 0.472
LRD(5) = 1/ [(√5 + √5)/2] = 0.447
LRD(6) = 1/ [(2 + √5)/2] = 0.472
LOF(A) = ([LRD(5) + LRD(6)]/2) / LRD(A) = [(0.472 + 0.447) / 2] / 0.472 = 0.973

LRD(B) = 1/ [(2 + √2)/2] = 0.586
LRD(3) = 1/ [(√2 + √2 + √2)/3] = 0.707
LRD(4) = 1/ [(2 + 2 + √2)/3] = 0.554
LOF(B) = ([LRD(3) + LRD(4)]/2) / LRD(B) = [(0.707 + 0.554) / 2] / 0.586 = 0.929

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

Lift chart (**3 points**)
Given the following decision tree on left, where the leaves also show the confidence of each prediction, and given the test set on the right, build the corrisponding Lift chart. Show the process you followed.



A	B	C	class
a1	b1	c2	Y
a2	b2	c3	N
a2	b1	c1	Y
a3	b2	c1	N
a3	b1	c2	Y

Answer:

Predictions

A	B	C	class	predicted
a1	b1	c2	Y	Y 0.7
a2	b2	c3	N	N 0.9
a2	b1	c1	Y	Y 0.75
a3	b2	c1	N	Y 0.65
a3	b1	c2	Y	N 0.7

Sorted

A	B	C	class	predicted	p(Y)	FPR	TPR	Positive cases
a2	b1	c1	Y	Y 0.75	0.75	0	0	0
a1	b1	c2	Y	Y 0.7	0.7	0	1	1
a3	b2	c1	N	Y 0.65	0.65	1	2	2
a3	b1	c2	Y	N 0.7	0.3	1	3	3
a2	b2	c3	N	N 0.9	0.1	2	3	4
								5