Università di Pisa	A.A. 2016-2017

Data Mining II **September 6th, 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 = 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>.

	Occurrences	*Occurrences with max-gap =3*
w_1 = <{B} {B} >		
w_2 = <{F}{B}>		
w_3 = <{B}{F}{C,D}>		

Answer:

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

Answer:

Output Sequential patterns

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

remark: also the following 3-sequences were generated, but then discarded:

{ A } -> { C } -> { B } ← removed by PRUNING
{ B } -> { C } -> { B } ← removed by PRUNING
{ B } -> { C } -> { C } ← infrequent
{ C } -> { C } -> { C } ← infrequent

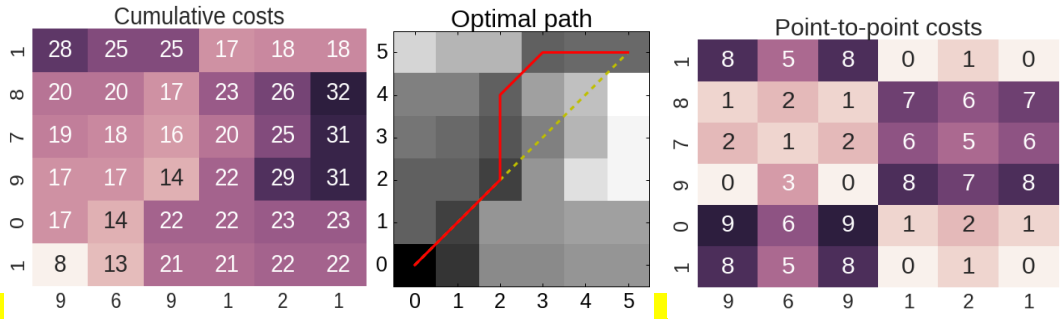
Exercise 2 - Time series / Distances (**6 points)**

Given the following time series:

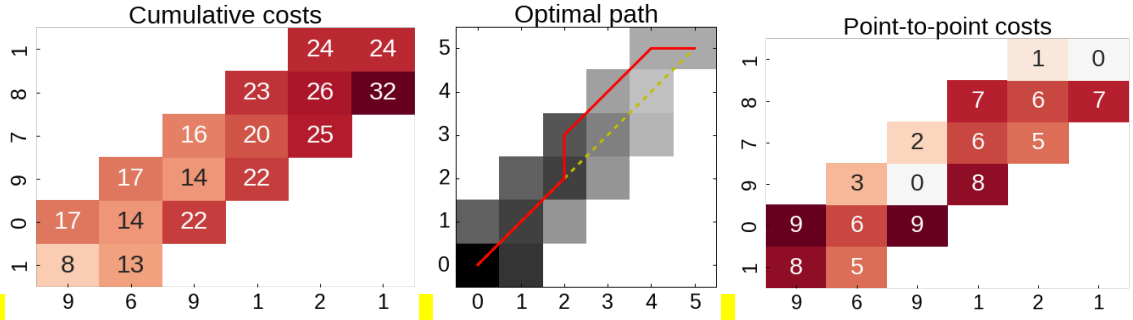
$$t = < 1, 0, 9, 7, 8, 1 >$$
$$q = < 9, 6, 9, 1, 2, 1 >$$

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: 24

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

A large on-line book seller wants to provide its on-line users a social networking service that automatically forms groups ("reading clubs") of users that have similar tastes and reading habits, in order to put them in contact.
The data available by the book seller are the records of books bought by each customer, a small description of the customers (including age, gender and education level) and of the books (including author, genre, number of pages, cost).
Briefly describe a project plan to help the company to organize such a service, (loosely) following the CRISP-DM methodology. Clearly remark the choices and assumptions made in the process.

Answer:
Key steps: the problem looks like a customer segmentation, thus approachable by clustering. The features to consider include all those about the customer (age, etc.) plus some aggregates derived from purchase history, for instance: frequency of the various genre, average length of books, number of books per year.

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	high	N
no	low	Y
yes	low	N
yes	high	N
no	high	Y
no	low	Y
yes	low	N

Answer:
Normal

	Y	N
	3	4
	A | Y	A | N
yes	0	3
no	3	1
	B | Y	B | N
high	1	2
low	2	2

	Y	N
	0.43	0.57
	A | Y	A | N
yes	0.00	0.75
no	1.00	0.25
	B | Y	B | N
high	0.33	0.50
low	0.67	0.50

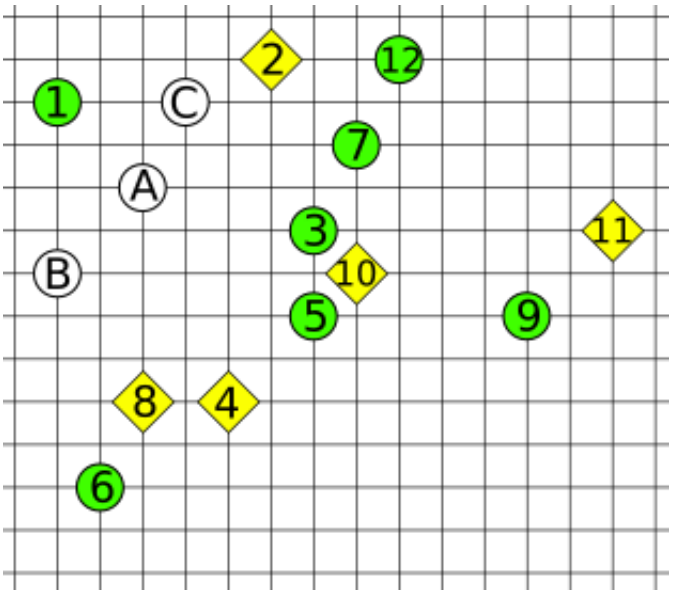
Laplace

	Y	N
	3	4
	A | Y	A | N
yes	0	3
no	3	1
	B | Y	B | N
high	1	2
low	2	2

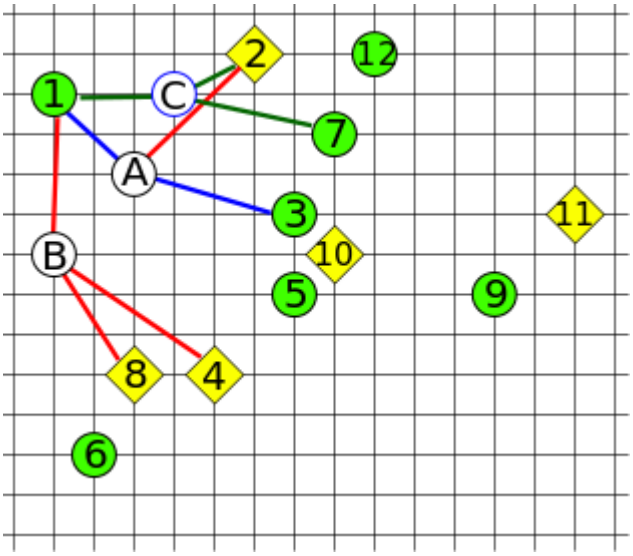
	Y	N
	0.43	0.57
	A | Y	A | N
yes	0.20	0.67
no	0.80	0.33
	B | Y	B | N
green	0.40	0.50
red	0.60	0.50

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.

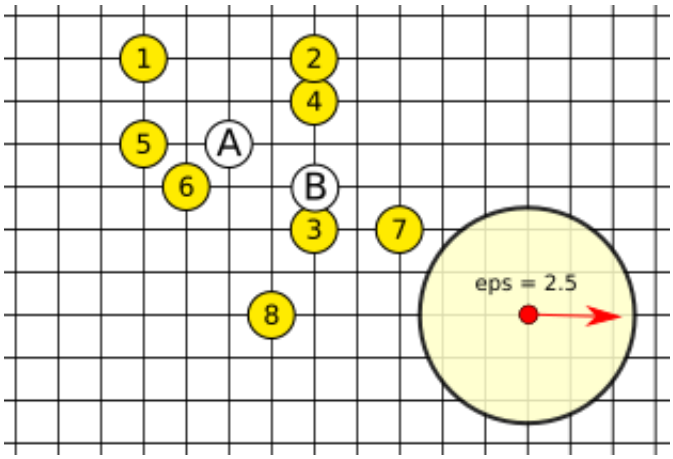


Answer:
kNN(A) = { 1, 2, 3} → CIRCLE
kNN(B) = { 1,4,8 } → SQUARE
kNN(C) = { 1,2,7 } → CIRCLE



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

Given the dataset of 10 points below (A, B, 1, 2, …, 8), consider the outlier detection problem for points A and B, adopting the following three methods:

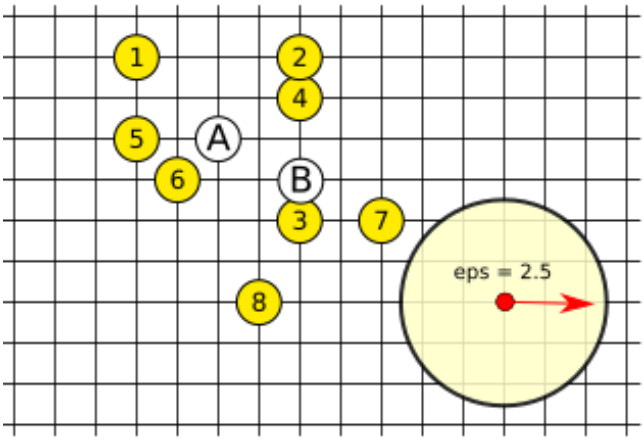


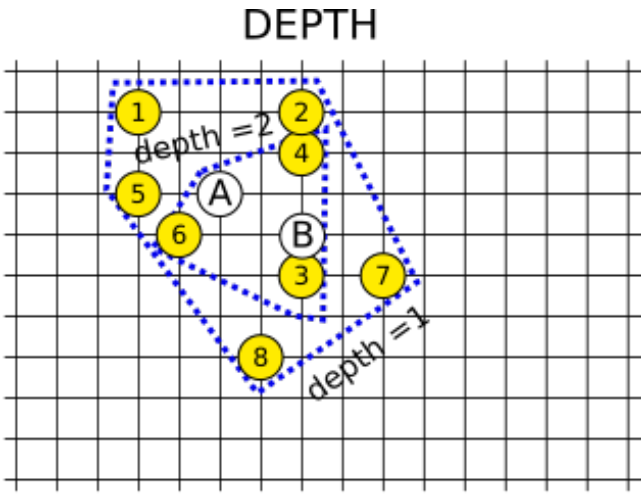
a) Distance-based: DB(ε,π) (**2 points**)
Are A and/or B outliers, if thresholds are forced to ε = 2.5 and π = 0.35 ? The point itself should not be counted.

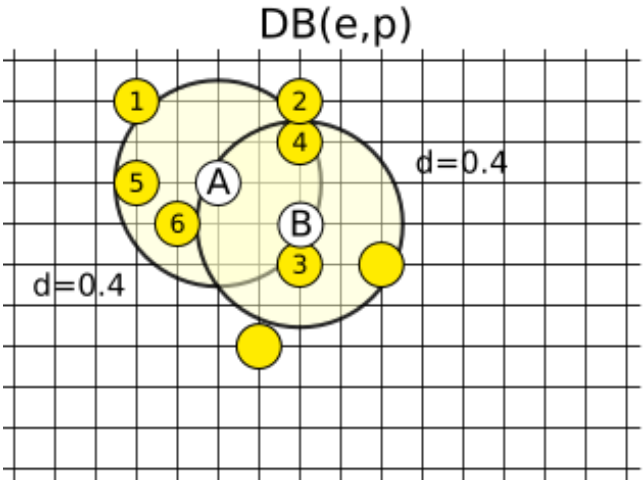
b) Density-based: LOF (**2 points**)
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2 NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

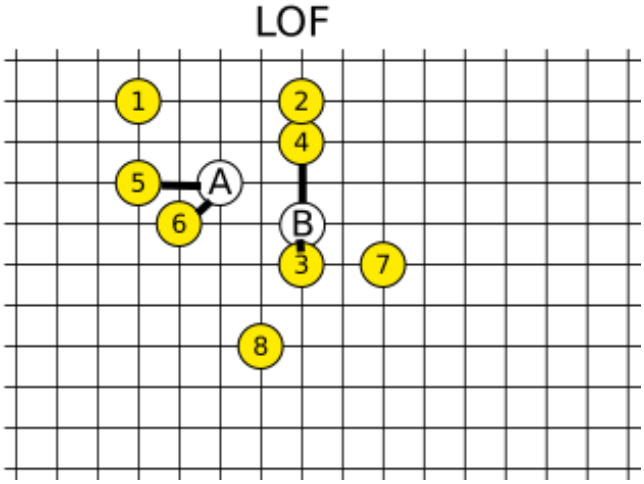
c) Depth-based (**2 points**)
Compute the depth score of all points.

Answer:









2-NN(A) = { 5,6 }

LRD(A) = 1/ [(2 + √2)/2] = 0.586
LRD(5) = 1/ [(√2 + 2 + 2)/3] = 0.554
LRD(6) = 1/ [(√2 + √2)/2] = 0.707
LOF(A) = ([LRD(5) + LRD(6)]/2) / LRD(A) = [(0.554 + 0.707) / 2] / 0.586 = 1.076 close to 1, no outlier

2-NN(B) = { 3,4 }

LRD(B) = 1/ [(1 + 2)/2] = 0.666
LRD(3) = 1/ [(1 + 2)/2] = 0.666
LRD(4) = 1/ [(1 + 2)/2] = 0.666
LOF(B) = ([LRD(3) + LRD(4)]/2) / LRD(B) = [(0.666 + 0.666) / 2] / 0.666 = 1.000 no outiler

---

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	N	N	0.8
row 2	Y	N	0.69
row 3	Y	N	0.44
row 4	N	Y	0.25
row 5	Y	N	0.9
row 6	Y	Y	0.79
row 7	N	Y	0.75
row 8	Y	N	0.81
row 9	Y	N	0.79
row 10	N	N	0.43

Answer:

Record	Real Class	Predicted		Score
row 1	N	N	0.8	0.2
row 2	Y	N	0.69	0.31
row 3	Y	N	0.44	0.56
row 4	N	Y	0.25	0.25
row 5	Y	N	0.9	0.1
row 6	Y	Y	0.79	0.79
row 7	N	Y	0.75	0.75
row 8	Y	N	0.81	0.19
row 9	Y	N	0.79	0.21
row 10	N	N	0.43	0.57

SORTED		TPR	FPR	AUC partial
Real Class	Score	0	0	
Y	0.79	1	0	0
N	0.75	1	1	1
N	0.57	1	2	1
Y	0.56	2	2	0
Y	0.31	3	2	0
N	0.25	3	3	3
Y	0.21	4	3	0
N	0.2	4	4	4
Y	0.19	5	4	0
Y	0.1	6	4	0

AUC **9**

Normalized **0.375**