Università di Pisa	A.A. 2016-2017

Data Mining II **October 30th, 2017**

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

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

<	{A,B,F}	{A,B}	{B,C}	{C,D,F}	{E}	{C,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 min-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>.

	Occurrences	*Occurrences with min-gap = 2*
w_1 = <{B} {B} >		
w_2 = <{F}{B}>		
w_3 = <{B}{F}{C,D}>		

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

Candidates

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

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

Given the following time series:

$$\mathbf{t} = \langle 2, 6, 1, 2, 1, 1 \rangle$$
$$\mathbf{q} = \langle 1, 0, 2, 7, 4, 0 \rangle$$

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.

Exercise 3 - Analysis process & CRISP-DM **(5 points)**

Large retail sellers usually have a significant fraction of customers that own a loyalty card, which enables the seller to track their purchases. However, they are insterested in providing some form of CRM also to the other customers, that we will call "card-free customers", for which the company has no personal profile and is also unable to link together their different baskets.

One of these large retail sellers is going to introduce several new products, and would like to to offer samples of them to a selected subset of customers. The selection should be performed: (i) in such a way that the customer most likely will appreciate the poduct and will start buying it; and (ii) should include card-free customers.

The offer will be proposed to the customer right after she payed for her basket, therefore the system can exploit the knowledge of what she bought.

To set up the service, it is possible to use also all the historical data of purchases performed in the past by the custmers.

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.

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

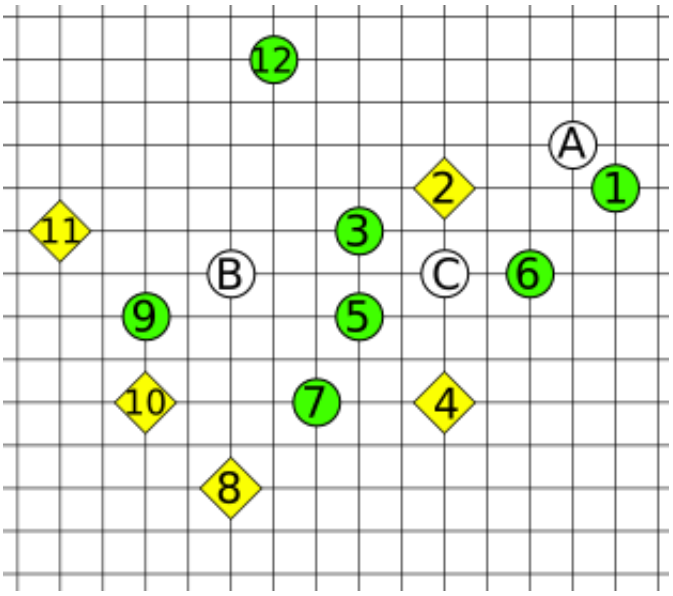
a) Naive Bayes (**3 points**)
Given the training set on the left, build a Naive Bayes classification model and apply it to the test set on the right.

A	B	C	class
high	no	green	Y
medium	no	red	Y
low	yes	green	N
high	no	red	N
low	yes	red	Y
high	no	green	Y
medium	yes	green	N

A	B	C	class
low	no	red	
high	yes	green	
medium	yes	red	

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.

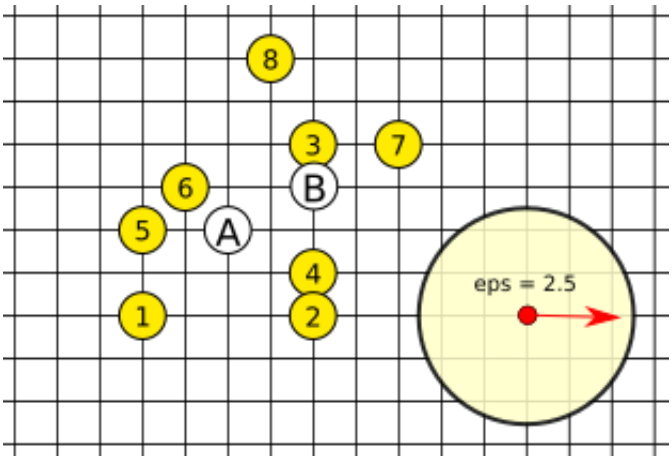


Exercise 5 - Outlier Detection **(6 points)**

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) (3 **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) Depth-based (3 **points**)
Compute the depth score of all points.



Exercise 6 - Validation **(3 points)**

ROC & AUC

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	0.03
row 2	Y	Y	0.43
row 3	Y	Y	0.36
row 4	N	Y	0.02
row 5	Y	Y	0.74
row 6	Y	Y	0.49
row 7	N	N	0.17
row 8	Y	Y	0.96
row 9	N	Y	0.95
row 10	N	Y	0.14