Università di Pisa	A.A. 2017-2018

Data Mining II **July 3, 2018**

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

Given the input sequences listed in the table below (column 1), show for each of them **all the occurrences** of subsequences {A} → {C} → {D} and {A,B} → {C}, and finally write its total support. Repeat the exercise twice: the first time **considering no temporal constraints** (columns 2 and 4); the second time **considering min-gap = 1** (i.e. gap > 1) (columns 3 and 5). Each occurrence should be represented by its corresponding list of time stamps, e.g..: <0,2,3> = <t=0, t=2, t=3>.

column 1	*column 2*	*column 3*	*column 4*	*column 5*
	{A} → {C} → {D}		**{A,B} → {C}**	
	No constraints	*min-gap = 1*	*No constraints*	*min-gap = 1*
< {A,B,F} {C} {C,D,F} {E} {C,D} > t=0 t=1 t=2 t=3 t=4				
< {A,B} {C} {A,B} {C,D} > t=0 t=1 t=2 t=3				
< {F} {A,B,F} {A,B,C} {D} {E} {C} > t=0 t=1 t=2 t=3 t=4 t=5				
< {A,F} {B,C} {A,B} {E} {D} > t=0 t=1 t=2 t=3 t=4				
< {A,B,F} {A,C} {A,B,D} {C} {C,D} > t=0 t=1 t=2 t=3 t=4				
Total support:				

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

Given the following time series:

t = < 7, 1, 0, 0, 0,1 >

q = < 8, 2, 1, 7, 3, 0 >

compute (i) their DTW, and (ii) their DTW with Sakoe-Chiba band of size r=2 (i.e. all cells at distance <= 2 from the diagonal are allowed). Show the cost matrices and the optimal paths found.

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

A retail seller wants to start some promotions made of a set of "convenience baskets", i.e. pre-defined groups of products (each group must contain at least 3 products, and no more than 6) that can be bought all together at a convenient price. One example might be the following: Convenient-Basket-1 = { milk, cookies, cereals, orange-juice }. The retail seller still has to define such "convenience baskets", and would like to choose some that contain popular combinations of products and such that the overall cost does not exceed 50€.
Briefly describe a project plan to help the company to do that, (loosely) following the CRISP-DM methodology, and providing concrete comments about how to solve the problem.

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

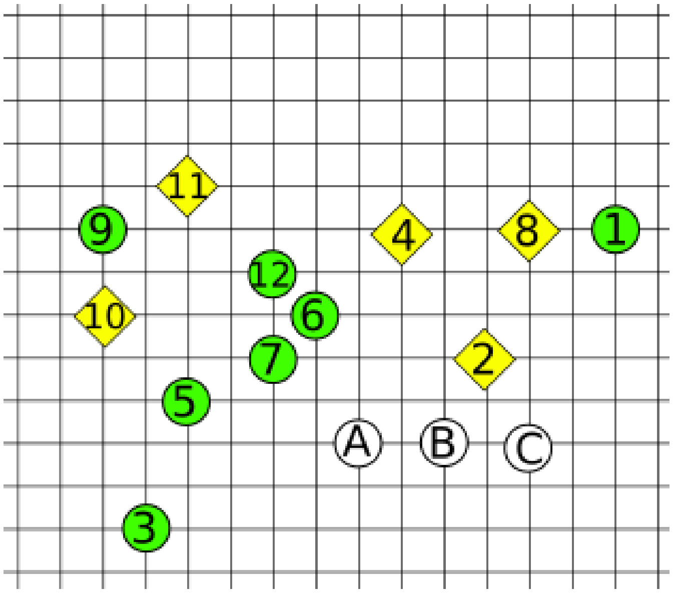
a) Naive Bayes (**3 points**)
Given the training set on the left, build a Naive Bayes classification model and apply it to the test set on the right.

SCORE	FIRST-TRY	FACULTY	class
good	no	science	Y
medium	yes	science	N
bad	yes	science	N
bad	yes	humanities	Y
good	no	humanities	N
good	no	science	Y
medium	no	humanities	Y

SCORE	FIRST-TRY	FACULTY	class
bad	no	humanities	
good	yes	science	
medium	yes	humanities	

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.



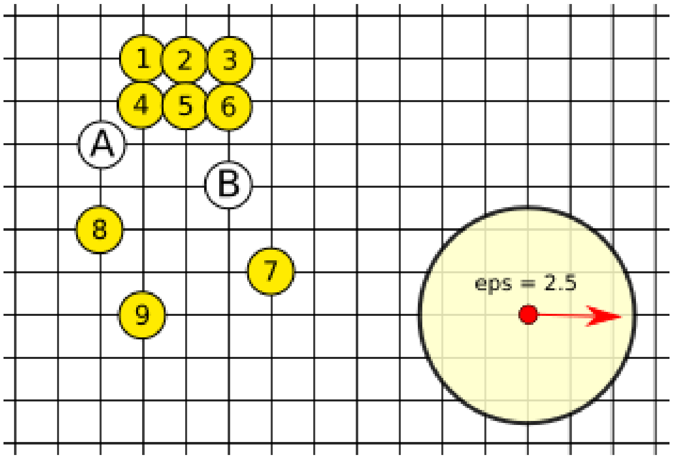
Exercise 5 - Outlier Detection **(6 points)**

Given the dataset of 11 points below (A, B, 1, 2, …, 9), 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 = 0.3? Show the density of the two points. (Notice: in computing the density of a point P, P itself should not be counted as neighbour).

b) Density-based: LOF (3 **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 (1 **points**)
Compute the depth score of all points.



Exercise 6 - 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 its AUC. Show the process followed to achieve that.

Record	Real Class	Predicted	
row 1	Y	N	0.70
row 2	Y	N	0.73
row 3	Y	Y	0.77
row 4	N	N	0.80
row 5	N	N	0.97
row 6	Y	N	0.76
row 7	N	N	0.65
row 8	N	N	0.87
row 9	Y	Y	0.93
row 10	Y	N	0.66