Data Mining II **April 4th, 2016**

Exercise 1 - Classification – alternative methods **(11 points)**

Given the training dataset below, predict the class of the below new test data by using k-Nearest Neighbor for k=3. For similarity measure use a simple match of attribute values: Similarity(A,B) that is computed by the following formula

$$\sum_{i=1}^{4} w_i * \partial(a_i, b_i) \Big/ 4$$

where $\partial(a_i, b_i)$ is 1 if a_i equals b_i and 0 otherwise. The only exception is the attribute Income, for which we have that ∂(Low,Medium) = ∂(Medium, Low) = ∂(High, Medium) = ∂(Medium, High) = 0.5, while for all the other cases the same rule apply as above, i.e. $\partial(a_{income}, b_{income})$ is 1 if if a_{income} equals b_{income} and 0 otherwise. Weights are all 0.3, except for Age, which is 0.1.

Training Data

Income	**Student**	**Age**	**Sex**	**Credit**
Low	yes	Young	F	No
High	yes	Young	M	Yes
Low	no	Young	M	No
High	yes	Old	F	Yes
Medium	yes	Young	M	Yes
Low	no	Old	F	No
Medium	no	Young	M	No
High	yes	Old	M	No

Test Data

Income	**Student**	**Age**	**Sex**	**Credit**
High	yes	Young	F	
High	no	Old	F	

Exercise 2 - Sequential patterns **(11 points)**

Given the following input sequence

< {A,C}	{B,C}	{D,E}	{A,E}	{A,C,D}	{F}	{B,E}	{C,D} >
t=0	t=1	t=2	t=3	t=4	t=5	t=6	t=7

A) 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 = 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>.

B) list all the subsequences that contain the event F, and satisfy both min-gap=3 (i.e. all gaps must be >3) and max-span=5.

	Occurrences	*Occurrences with max-gap=2*
***ex.:* <{C}{D}>**	**<0,2> <0,4> <0,7> <1,2> <1,4> <1,7> <4,7>**	**<0,2> <1,2>**
w_1 = <{A} {C} >		
w_2 = <{A}{A}>		
w_3 = <{A} {E} {D}>		

Exercise 3 - Time series / Distances **(10 points)**

Given the following dataset of time series (on the left):

ID	Time series
W	< 6, 11, 13, 15 >
X	< 10, 7, 7, 12, 14, 17 >
Y	< 9, 11, 14, 13, 20 >

	W	X	Y
W			
X			
Y			

compute the matrix of distances among all pairs of time series (on the right) adopting a Dynamic Time Warping distance, and computing the distances between single points as d(x,y) = | x – y |. For each pair of time series compared also show the matrix used to compute the final result.

W – X

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	4	5	6	12	20	31
[2,]	5	8	9	7	10	16
[3,]	8	11	14	8	8	12
[4,]	13	16	19	11	9	10

W – Y

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	3	8	16	23	37
[2,]	5	3	6	8	17
[3,]	9	5	4	4	11
[4,]	15	9	5	6	9

X – Y

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	2	6	9	19
[2,]	3	5	9	12	22
[3,]	5	7	12	15	25
[4,]	8	6	8	9	17
[5,]	13	9	6	7	13
[6,]	21	15	9	10	10