Università di Pisa	A.A. 2015-2016

Data Mining II July 8th, 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 ∂(a ,b) is 1 if a_i equals b_i and 0 otherwise. a_i and b_i are either age, sex, height or weight. In the above formula weights have the following values: $w_1 = w_4 = 0.4$, $w_2 = w_3 = 0.2$.

Training Data

Height	**Weight**	**Age**	**Sex**	**Disease**
Short	High	Young	F	No
Short	Low	Young	F	Yes
Short	Low	Old	M	No
Short	Medium	Young	M	Yes
Short	Medium	Old	M	Yes
Tall	Medium	Old	F	Yes
Tall	Low	Young	M	No
Short	High	Young	F	No
Tall	High	Old	M	No
Short	Medium	Old	M	Yes

Test Data

Height	**Weight**	**Age**	**Sex**	**Disease**
Short	High	Old	F	
Tall	Medium	Old	M	

Exercise 2 - Sequential patterns **(11 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

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 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>.

B) list all the subsequences containing at least 3 events and that satisfy min-gap=5 (i.e. all gaps must be >5).

{A} {B,E}

{B,F} {C,D}
{F} {C,D}
{B} {C,D}
{B,F} {D}
{B,F} {C}

	Occurrences	*Occurrences with min-gap=2*
***ex.:* <{B}{E}>**	**<1,2> <1,6> <3,6>**	**<1,6>**
w_1 = <{A} {F} {D} >	**<0,1,4> <0,1,7> <0,5,7> <3,5,7> <4,5,7>**	**-**
w_2 = <{A}{E}>	**<0,2> <0,6> <3,6><4,6>**	**<0,6> <3,6>**
w_3 = <{B} {C,D}>	**<1,4> <1,7> <3,4><3,7><6,7>**	**<1,4> <1,7> <3,7>**

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

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

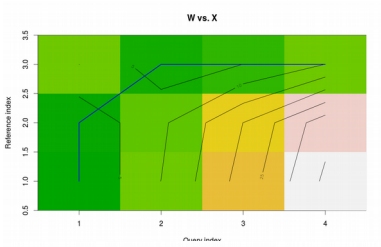
W	< 1, 11, 13,15 >
X	< 1, 2, 10 >
Y	< 9, 8, 1, 13, 1 >
Z	< 0,1,2,3 >

	W	**X**	**Y**	**Z**
W				
X				
Y				
Z				

1) 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.

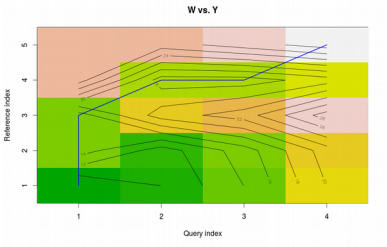
DTW(W,X)

```
     [,1] [,2] [,3]
[1,]    0    1   10
[2,]   10    9    2
[3,]   22   20    5
[4,]   36   33   10
```



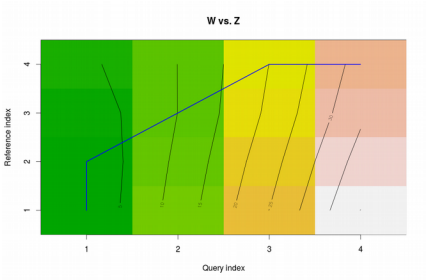
DTW(W,Y)

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    8   15   15   27   27
[2,]   10   11   21   17   27
[3,]   14   15   23   17   29
[4,]   20   21   29   19   31
```



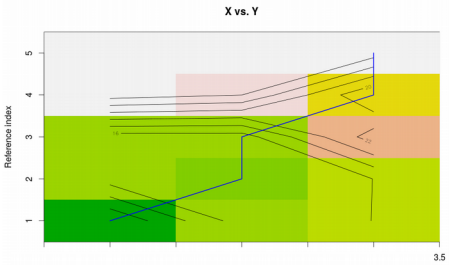
DTW(W,Z)

```
     [,1] [,2] [,3] [,4]

[1,]    1    1    2    4
[2,]   12   11   10   10
[3,]   25   23   21   20
[4,]   40   37   34   32
```



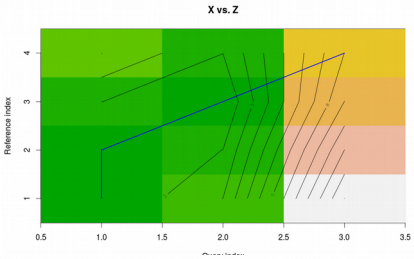
DTW(X,Y)

```
     [,1] [,2] [,3] [,4] [,5]

[1,]    8   15   15   27   27
[2,]   15   14   15   26   27
[3,]   16   16   23   18   27
```



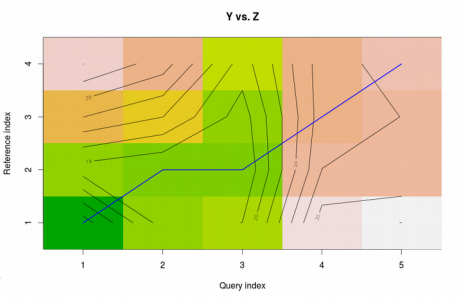
DTW(X,Z)

```
     [,1] [,2] [,3] [,4]
[1,]    1    1    2    4
[2,]    3    2    1    2
[3,]   13   11    9    8
```



DTW(Y,Z)

```
     [,1] [,2] [,3] [,4]

[1,]    9   17   24   30
[2,]   17   16   22   27
[3,]   18   16   17   19
[4,]   31   28   27   27
[5,]   32   28   28   29
```



2) Which distances will change if we constrain the DTW with a "Sakoe-Chiba Band " of size r=1, i.e. the maximum misalignment allowed in the matching is of 1 position? **ANS: "W vs Y" only**