| Università di Pisa | A.A. 2013- 2014 |
|---|---|

**Data Mining II** June 30th, 2014

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

Given the training dataset below (on the left), build and apply a Naïve Bayes Classifier to predict the "class" attribute on the test dataset (on the right).

| outlook | Temp. | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

**Training set**

| outlook | temp. | humidity | windy |
|---|---|---|---|
| sunny | hot | high | false |
| overcast | cool | high | true |
| rainy | hot | normal | true |
| sunny | hot | high | true |

**Test set**

Exercise 2 - Sequential patterns **(10 points)**

Given the following input sequence

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

show all the occurrences (there can be more than one or none, in general) of each of the following su-sequences in the input sequence above. Repeat the exercise twice: the first time considering no temporal constraints (left column): the second time considering min-gap = 1 (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=1* |
|---|---|---|
| *ex.:* <{C}{F}{D}> | **<1,2,3> <1,2,4> <1,2,7> <1,6,7> <2,6,7>** | **none** |
| $w_1$ = <{B} {C} {A} > | | |
| $w_2$ = <{A}{A}> | | |
| $w_2$ = <{C}{B}> | | |

Exercise 3 - Time series / Classification **(11 points)**

Given the following dataset of labelled time series:

| **Time series** | **Label** |
|---|---|
| < 9, 8, 3, 2 > | Y |
| < 9, 1, 5, 2 > | N |
| < 3, 2, 7, 8 > | Y |
| < 1, 1, 2, 1 > | N |

and the following test set of unlabelled time series:

| **Time series** | **Label** |
|---|---|
| < 9, 9, 3, 1 > | |
| < 9, 1, 5, 2 > | |
| < 1, 7, 8, 5 > | |

1) Classify the test set using a 1-Nearest-Neighbor approach, by adopting the Euclidean distance as proximity measure to compare time series.
2) Repeat the same task adopting a Dynamic Time Warping distance.

| Università di Pisa | A.A. 2013- 2014 |
|---|---|

**Data Mining II** July 21th, 2014

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

Given the training dataset below (on the left), apply a K-Nearest-Neighbor Classifier with K=3 to predict the "class" attribute on the test dataset (on the right). Evaluate the accuracy of the classifier.

| X | Y | Z | Class |
|---|---|---|---|
| 46 | 33 | 48 | No |
| 8 | 15 | 25 | No |
| 10 | 11 | 35 | Yes |
| 29 | 15 | 7 | Yes |
| 11 | 32 | 46 | Yes |

**Training set**

| X | Y | Z | Class |
|---|---|---|---|
| 7 | 8 | 45 | Yes |
| 30 | 8 | 40 | No |
| 13 | 23 | 21 | No |
| 47 | 43 | 34 | No |
| 37 | 10 | 29 | Yes |
| 19 | 49 | 31 | No |
| 20 | 13 | 8 | Yes |
| 33 | 44 | 16 | Yes |
| 47 | 12 | 41 | No |
| 49 | 21 | 3 | Yes |

**Test set**

## Exercise 2 - Sequential patterns **(10 points)**

Given the following input sequence

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

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 = 1 (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=1* |
|---|---|---|
| *ex.:* <{F}{D}> | **<2,3> <2,4> <2,7> <6,7>** | **<2,4> <2,7>** |
| $w_1$ = <{A} {C} {D} > | | |
| $w_2$ = <{B}{A}> | | |
| $w_3$ = <{A}{F} {D}> | | |
| $w_4$ = <{A}{E}> | | |

Exercise 3 - Time series / Distances **(11 points)**

Given the following dataset of time series:

| ID | Time series |
|---|---|
| A | < 19, 14, 19, 19, 26, 29, 38, 30 > |
| B | < 15, 10, 0, 2, 4, 7, 1, 9 > |
| C | < 19, 11, 20, 27, 18, 25, 15, 19 > |
| D | < 12, 3, 12, 19, 18, 24, 27, 31 > |

compute the matrix of distances among all pairs of time series adopting a Dynamic Time Warping distance, constrained with a “Sakoe-Chiba Band” of size r=2, i.e. the maximum misalignment allowed in the matching is of 2 positions.

| Università di Pisa | A.A. 2013- 2014 |
|---|---|

**Data Mining II** January 19th, 2014

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

Given the training dataset below (on the left), apply a K-Nearest-Neighbor Classifier with K=1 and then also with K=3 to predict the "class" attribute on the test dataset (on the right). Evaluate the accuracy of the two classifiers. Which one performs better?

| X | Y | Z | Class |
|---|---|---|---|
| 30 | 33 | 48 | No |
| 8 | 15 | 25 | No |
| 15 | 11 | 35 | Yes |
| 29 | 15 | 7 | Yes |
| 35 | 20 | 46 | Yes |

**Training set**

| X | Y | Z | Class |
|---|---|---|---|
| 7 | 8 | 45 | Yes |
| 30 | 8 | 40 | No |
| 13 | 23 | 21 | No |
| 47 | 43 | 34 | No |
| 37 | 10 | 29 | Yes |
| 19 | 49 | 31 | No |
| 20 | 13 | 8 | Yes |
| 33 | 44 | 16 | Yes |
| 47 | 12 | 41 | No |
| 49 | 21 | 3 | Yes |

**Test set**

## Exercise 2 - Sequential patterns **(10 points)**

Given the following input sequence

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

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 (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* |
|---|---|---|
| *ex.:* <{F}{D}> | **<2,3> <2,4> <2,7> <6,7>** | **<2,3> <2,4> <6,7>** |
| $w_1$ = <{A} {C} {D} > | | |
| $w_2$ = <{B}{A}> | | |
| $w_3$ = <{A}{F} {D}> | | |
| $w_4$ = <{A}{E}> | | |

Exercise 3 - Time series / Distances **(11 points)**

Given the following dataset of time series:

| ID | Time series |
|---|---|
| A | < 19, 14, 19, 26, 38 > |
| B | < 15, 10, 0, 2, 4 > |
| C | < 19, 18, 25, 27, 40 > |
| D | < 20, 15, 15, 0, 5 > |

compute the matrix of distances among all pairs of time series adopting a Dynamic Time Warping distance.