University Mohamed Boudiaf M'sila | **M2 SIGL**
MCS Faculty, CS Département | **M'sila, 21/01/2024**

Final Exam

Data Mining & Information Retrieval

Exercice 1 (5 pts) :

The similarity of two attributes with nominal values is evaluated using a generalization of binary variables. We propose to study two characteristics of plants: **Leaf Color** (*yellow*, *green*, *red*) and **Leaf Size** (*small*, *large*). Let's consider the data of plants as follows:

	Leaf Color	**Leaf Size**
A	*red*	*small*
B	*yellow*	*large*
C	*green*	*small*
D	*yellow*	*large*

Convert the plant data into binary values and calculate the distances d(A,B), d(B,C), d(A,C) and d(B,D) using Jaccard index. Comment these distances.

Exercice 2 (7 pts)

Given the dataset D in the following table,

1- Consider the following measurement which calculates the distance between two points a and b in D :

$$d(a, b) = max_i |x_i - y_i|$$

Is this distance a measurement of similarity or dissimilarity?

2- Using the complete link as a measure of distance between 2 clusters, perform a bottom-up hierarchical clustering on D and plot the corresponding dendrogram.

Points	x_i	y_i
A	1	1
B	1.5	0.5
C	0.8	1.2
D	-1	-0.8
E	-0.2	0.5
F	0.2	-1

Exercice 3 (8 pts)

Consider the following data corresponding to 7 observations of three variables X, Y and Z. The target class is the last column.

	X	**Y**	**Z**	**Class**
x1	A	α	1	C1
x2	A	β	1	C1
x3	A	α	1	C1
x4	B	α	3	C1
x5	B	α	1	C2
x6	C	β	2	C2
x7	C	β	2	C2

Apply the naive Bayesian classifier algorithm to this binary classification problem for predicting each of the following 3 new observations:

	X	**Y**	**Z**	**Class**
x8	A	α	2	?
x9	C	β	1	?
x10	B	β	1	?

N.B. *Do not forget to use Laplacian Correction if necessary.*

Dr. Tahar Mehenni

University Mohamed Boudiaf M'sila — **M2 SIGL**
MCS Faculty, CS Department — **M'sila, 21/01/2024**

Solution of Final Exam

Data Mining & Information Retrieval

Exercice 1 (5 pts)

	LeafColorYellow	LeafColorGreen	LeafColorRed	LeafSizeSmall	LeafSizeLarge
A	0	0	1	1	0
B	1	0	0	0	1
C	0	1	0	1	0
D	1	0	0	0	1

<table>
<tr><td></td><td colspan="3">B</td><td></td><td colspan="3">C</td><td></td><td colspan="3">C</td><td></td><td colspan="3">D</td></tr>
<tr><td rowspan="3">A</td><td></td><td>1</td><td>0</td><td rowspan="3">A</td><td></td><td>1</td><td>0</td><td rowspan="3">B</td><td></td><td>1</td><td>0</td><td rowspan="3">B</td><td></td><td>1</td><td>0</td></tr>
<tr><td>1</td><td>0</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td><td>2</td><td>1</td><td>2</td><td>0</td></tr>
<tr><td>0</td><td>2</td><td>1</td><td>0</td><td>1</td><td>2</td><td>0</td><td>2</td><td>1</td><td>0</td><td>0</td><td>3</td></tr>
</table>

d(A,B)= (2+2)/(0+2+2)=1 d(A,C)= (1+1)/(1+1+1)=2/3 d(B,C)= (2+2)/(0+2+2)=1 d(B,D)= (0+0)/(2+0+0)=0
No similarity: A and B, B and C (distance=1)
Partial similarity; A and C (in Leaf Size property!) (distance=0.67)
Complete similarity: B and D (distance=0)
Jaccard Index is a measure of dissimilarity.

Exercice 2 (7 pts)

1- Consider the following measurement which calculates the distance between two points a and b in D :

$$d(a, b) = max_i |x_i - y_i|$$

$$d(A,B)=max_{1,2} (|x_1 - y_1|, |x_2 - y_2|)=max((|1 - 1|, |1.5 - 0.5|)=1$$

Since d(a,b) take the maximum of absolute values, it may be a measurement of dissimilarity.

2- Bottom-up hierarchical clustering of D and corresponding dendrogram.

Points	x_i	y_i
A	1	1
B	1.5	0.5
C	0.8	1.2
D	-1	-0.8
E	-0.2	0.5
F	0.2	-1

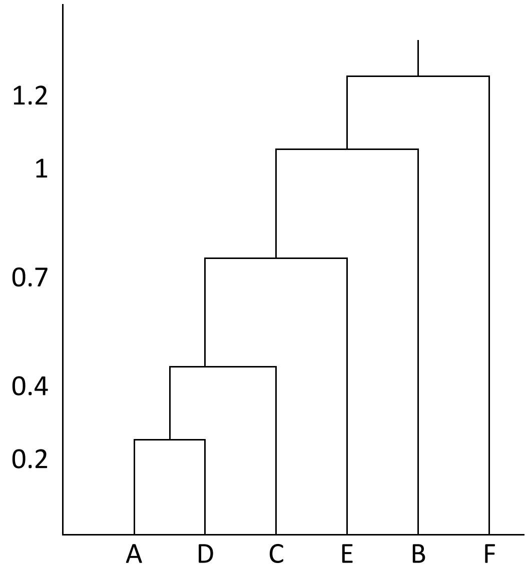
	A	B	C	D	E	F
A	0	1	0.4	**0.2**	0.7	1.2
B		0	1	1	1	1.2
C			0	0.4	0.7	1.2
D				0	0.7	1.2
E					0	1.2
F						0

	AD	B	C	E	F
AD	0	1	**0.4**	0.7	1.2
B		0	1	1	1.2
C			0	0.7	1.2
E				0	1.2
F					0

	ADC	B	E	F
ADC	0	1	**0.7**	1.2
B		0	1	1.2
E			0	1.2
F				0

	ADCE	B	F
ADCE	0	**1**	1.2
B		0	1.2
F			0

	ADCEB	F
ADCEB	0	**1.2**
F		0



Exercice 3 (8 pts)

Consider the following data corresponding to 7 observations of three variables X, Y and Z. The target class is the last column.

	X	Y	Z	Class
x1	A	α	1	C1
x2	A	β	1	C1
x3	A	α	1	C1
x4	B	α	3	C1
x5	B	α	1	C2
x6	C	β	2	C2
x7	C	β	2	C2

Apply the naive Bayesian classifier algorithm to this binary classification problem for predicting each of the following 3 new observations:

	X	Y	Z	Class
x8	A	α	2	?
x9	C	β	1	?
x10	B	β	1	?

$P(C_1)$: = 4/7 = 0.571 $P(C_2)$: = 3/7 = 0.428

1) Classification of x8 (X=A, Y= α, Z=2)

Compute $P(x8|C_i)$ for each class

P(X = A | C1) = 3/4 = 0.75, P(X = A | C2) = 0/3 =0

Laplacian correction , P(X = A | C1) = 4/5 = 0.8, P(X = A | C2) = 1/4 =0.25

P(Y = α | C1) = 3/4 = 0.75, P(Y = α | C2) = 1/3 = 0.333

P(Z = 2 | C1) = 0/4 = 0, P(Z = 2 | C2) = 2/3 = 0.667

Laplacian correction , P(Z = 2 | C1) = 1/5 = 0.2, P(Z = 2 | C2) = 3/4 = 0.75

$P(x8|C_i)$: P(x8|C1) = 0.8 x 0.75 x 0.2 = 0.12

P(x8|C2) = 0.25 x 0.333 x 0.75 = 0.062

$P(x8|C_i)*P(C_i)$: P(x8|C1) x P(C1) = 0.12 x 0.571 = 0.0685

P(x8|C2) x P(C2) = 0.062 x 0.428 = 0.0265

Therefore, x8 belongs to class C1

2) Classification of x9 (X=C, Y= β, Z=1)

Compute $P(x9|C_i)$ for each class

P(X = C | C1) = 0/4 = 0, P(X = C | C2) = 2/3 =0.666

Laplacian correction , P(X = C | C1) = 1/5 = 0.2, P(X = C | C2) = 3/4 =0.75

P(Y = β | C1) = 1/4 = 0.25, P(Y = β | C2) = 2/3 = 0.666

P(Z = 1 | C1) = 3/4 = 0.75, P(Z = 1 | C2) = 1/3 = 0.333

$P(x9|C_i)$: $P(x9|C1) = 0.2 \times 0.25 \times 0.75 = 0.0375$

$P(x9|C2) = 0.75 \times 0.666 \times 0.333 = 0.1663$

$P(x9|C_i)*P(C_i)$: $P(x9|C1) \times P(C1) = 0.0375 \times 0.571 = 0.0214$

$P(x9|C2) \times P(C2) = 0.1663 \times 0.428 = 0.0711$

Therefore, x9 belongs to class C2

3) Classification of x10 (X=B, Y= β, Z=1)

Compute $P(x10|C_i)$ for each class

$P(X = B \mid C1) = 1/4 = 0.25$, $P(X = B \mid C2) = 1/3 = 0.333$

$P(Y = \beta \mid C1) = 1/4 = 0.25$, $P(Y = \beta \mid C2) = 2/3 = 0.666$

$P(Z = 1 \mid C1) = 3/4 = 0.75$, $P(Z = 1 \mid C2) = 1/3 = 0.333$

$P(x10|C_i)$: $P(x10|C1) = 0.25 \times 0.25 \times 0.75 = 0.0468$

$P(x10|C2) = 0.333 \times 0.666 \times 0.333 = 0.0738$

$P(x10|C_i)*P(C_i)$: $P(x10|C1) \times P(C1) = 0.0468 \times 0.571 = 0.0267$

$P(x10|C2) \times P(C2) = 0.0738 \times 0.428 = 0.0315$

Therefore, x10 belongs to class C2