Lebanese University
Faculty of information
Section 1

Final Exam: Data Mining **Time: 90 minutes**
Dr. Abbas RAMMAL

Problem 1: Association Mining Rules (7 points)

Giving the following database with four transactions and a minimum support threshold of 60% and a minimum confidence threshold of 80%.

1. Find all frequent itemsets and association rules using
   a) A priori
   b) FP-Growth.
2. Compare the efficiency of both processes.

Transaction ID	Items
T1	{K, A, D, B}
T2	{D, A, C, E, B}
T3	{C, A, B, E}
T4	{B, A, D}

Problem 2: Decision Tree Classifier (7 points)

Given the training data in the table below:

Outlook (X_1)	Temperature (X_2)	Humidity (X_3)	Play Tennis? (Y)
sunny	hot	high	no
overcast	hot	high	yes
rain	mild	high	yes
rain	cool	normal	yes
sunny	mild	high	no
sunny	mild	normal	yes
rain	mild	normal	yes
overcast	hot	normal	yes

a) Using the dataset above, calculate the mutual information for each feature (X_1, X_2, X_3) to determine the root node for a Decision Tree trained on the above data.
b) Calculate what the next split should be.
c) Draw the resulting tree.

Problem 3: Data preprocessing (6 points)

Use the methods below to normalize the following group of data:

200, 300, 400, 600, 1000

a) min-max normalization by setting min = 0 and max = 1
b) z-score normalization
c) normalization by decimal scaling

2. Given the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
   b) How might you determine outliers in the data?

minSup = 60% = 0.6

TID	Items
T1	K, A, D, B
T2	D, A, C, E, B
T3	C, A, B, E
T4	B, A, D

Item	Support
A	4/4 = 1
B	4/4 = 1
C	2/4 = 0.5
D	3/4 = 0.75
E	2/4 = 0.5
K	1/4 = 0.25

C1

L1 = {A, B, D}

Itemset	Support
A, B	4/4
A, D	3/4
B, D	3/4

L2 = {{A, B}, {A, D}, {B, D}}

Itemset	Support
A, B, D	3/4

L3 = {{A, B, D}}

min Conf = 0.8

Rules

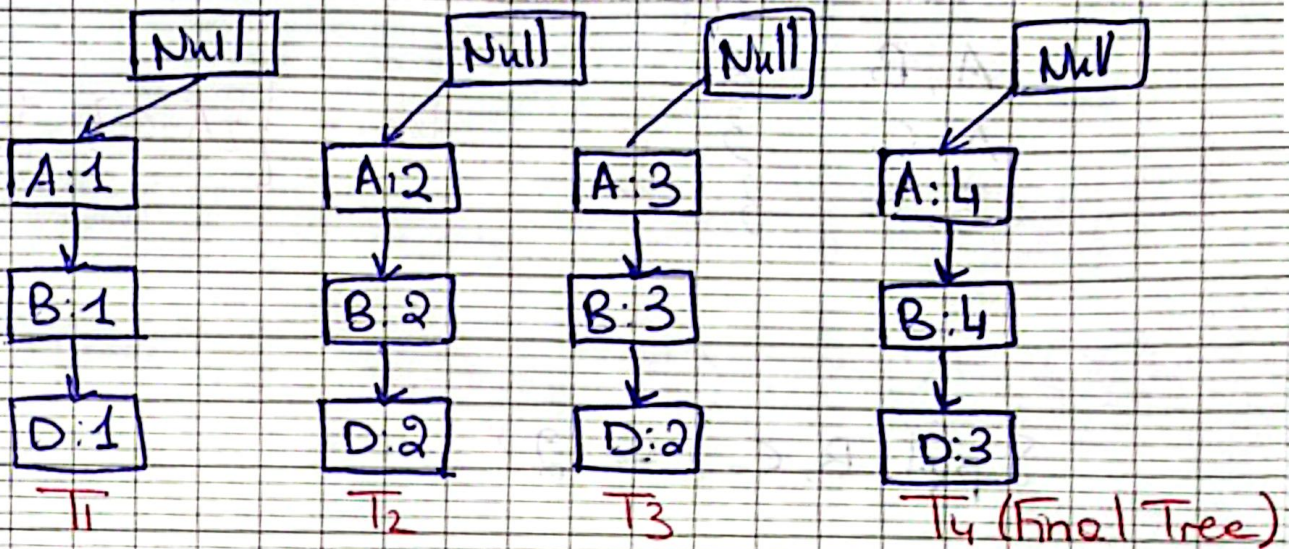
A → B	4/4		A → BD		
B → A	4/4		BD → A	3/3	
A → D	3/4		B → AD	3/4	
D → A	3/3		AD → B	3/3	
B → D	3/4		D → AB	3/3	
D → B	3/3		AB → D	3/4	

Step 1 : Count and Sort

Item	Supp
A	4/4
B	4/4
D	3/4

TID	Items	Sorted items
T_1	K, A, D, B	A, B, D
T_2	D, A, C, E, B	A, B, D
T_3	C, A, B, E	A, B
T_4	B, A, D	A, B, D

Step 2: Build FP_Tree



Step 3: Confidence Pattern Base and Confidence FP-Tree

Item	Conf. Pattern Base	Conf. FP Tree	F.P Generated
D(3)	{A, B : 3}	{A : 3 B : 3}	{D:3} {A,D:3} {B,D:3} {A,B,D:3}
B(4)	{A : 4}	{A : 4}	{B:4} {A,B:4}
A(4)	{ }	{ }	{A:4}

4) Rules

A → D 3/4
D → A 3/3 ✓
B → D 3/4
D → B 3/3 ✓

AB → D 3/4
A → B 4/4 ✓
B → A 4/4 ✓
AB → BD 3/4

BD → A 3/3 ✓
B → AD 3/4
AD → B 3/3 ✓
D → AB 3/3 ✓

Problem 2:

- Before Splitting $\{Y^6, N^2\}$

G.I(B) = $1 - [(6/8)^2 + (2/8)^2] = 0.625$

- G.I for each attrb.

1) Outlook:

Sunny $\{Y^1, N^2\}$ O: $\{Y^2, N^0\}$ R: $\{Y^3, N^0\}$

G.I(S) = $1 - [(1/3)^2 + (2/3)^2] = 0.44$

G.I(O) = 0

G.I(R) = 0

G.I (Outlook) = $3/8\ (0.44) = 0.165$

Gain(S, Outlook) = G.I(B) − G.I(Outlook)

= 0.625 − 0.165 = 0.46

2) Temperature: H: $\{Y^2, N^1\}$ M: $\{Y^3, N^1\}$ C: $\{Y^1, N$

G.I(H) = $1 - [(2/3)^2 + (1/3)^2] = 0.44$

G.I(M) = $1 - [(3/4)^2 + (1/4)^2] = 0.375$

G.I(C) = 0

G.I(Temp) = $3/8\ (0.44) + 4/8\ (0.375)$

= 0.3525

Gain(S, Temp) = 0.625 − 0.3525 = 0.27

3) Humidity: H: {Y², N²} N: {Y⁴, N⁰}

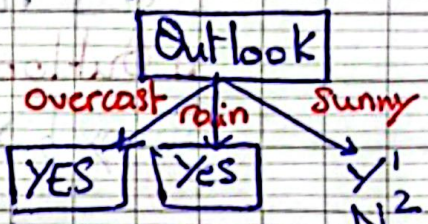
G.I(H) = 1 − [(2/4)² + (2/4)²] = 0.5

G.I(N) = 0

G.I (Humidity) = 4/8 (0.5) = 0.25

Gain (S, Hum.) = 0.625 − 0.25 = 0.37

Highest Gain: Outlook



↳ Filter the dataset on 'overcast' and 'rain'

O	T	H	PY?
Sunny	hot	hight	N
Sunny	mild	hight	N
Sunny	mild	normal	Y

G.I(B) = 1 − [(2/3)² + (1/3)²]
= 0.44

1) Temperature: H: {Y⁰, N¹} M: {Y¹, N¹}

G.I(H) = 0 / G.I (M) = 0.5

G.I(Temp) = 2/3 (0.5) = 0.33

Gain (Temp) = 0.44 − 0.33 = 0.11

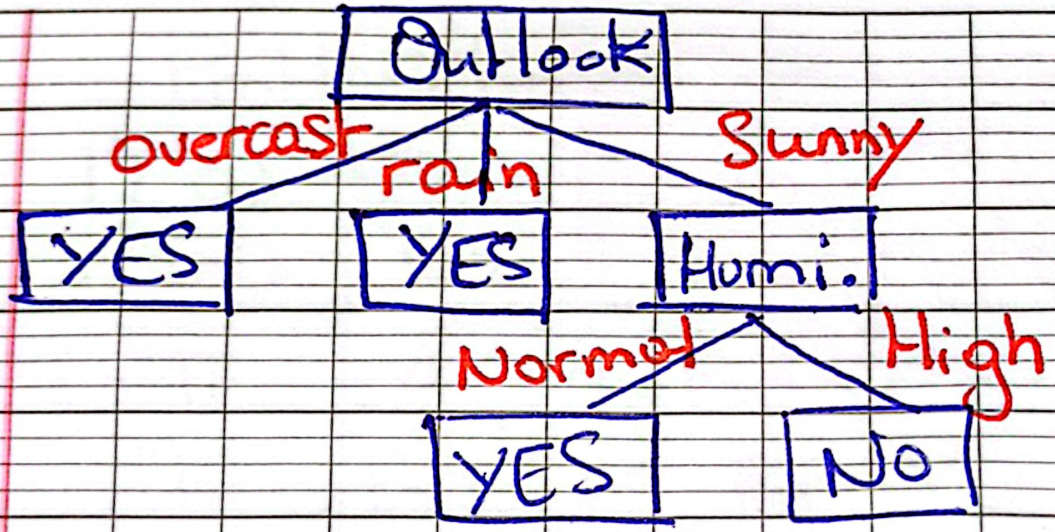
2) Humidity: H: {Y⁰, N²} N {Y¹, N⁰}

G.I(H) = 0

G.I(N) = 0

G.I (Humidity) = 0

Gain (Humidity) = 0.44 (Highest) next split


Outlook
overcast
rain
Sunny
YES
YES
Humi.
Normal
High
YES
No

Problem 3

1) a) Min_Max Normalization

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

min = 200 , max = 1000

$$200' = \frac{200 - 200}{1000 - 200} = 0$$

$$300' = \frac{300 - 200}{1000 - 200} = 0.125$$

$$400' = \frac{400 - 200}{1000 - 200} = 0.25$$

$$600' = \frac{600 - 200}{1000 - 200} = 0.5$$

$$1000' = \frac{1000 - 200}{1000 - 200} = 1$$

0 , 0.125 , 0.25 , 0.5 , 1

b) z-score normalization

$$x = \frac{x - M}{\sigma}$$

Mean (M) = 500
Sd (σ) = 282.84

200' = 200 - 500 / 282.84 = -1.06
300' = 300 - 500 / 282.84 = -0.71
400' = 400 - 500 / 282.84 = -0.35
600' = 600 - 500 / 282.84 = 0.35
1000' = 1000 - 500 / 282.84 = 1.77

c) Decimal Scaling

$x' = x / 10^j$

$j = \log_{10}(\max|x|)$
max = 1000
$j = 3$

200' = $200 / 10^3 = 0.2$
300' = $300 / 10^3 = 0.3$
400' = $400 / 10^3 = 0.4$
600' = $600 / 10^3 = 0.6$
1000' = $1000 / 10^3 = 1$

step1:
2)a) Group into Bins of Depth 3

13 15 16 / 16 19 20 / 20 21 22 / 22 25 25 /
25 25 30 / 33 33 35 / 35 35 35 / 36 40 45
46 52 70 /

Step2: Compute the mean of each Bin

1) 14.66
2) 18.33
3) 21
4) 24
5) 26.6
6) 33.66
7) 35
8) 40.33
9) 56

Step3: Replace each value with the mean of the corresponding bin

14.66 14.66 14.66 18.33 18.33 18.33
21 21 21 24 24 24 26.6 26.6 26.6 ...,

Effect: Reduces noise

b) Z_score Method values with zscore>3 or <-3 are outliers

IQR Method

values less than Q1 - 1.5xIQR or greater than Q3 + 1.5x IQR are considered outliers