الجامعة اللبنانية الأمريكية
Lebanese American University

# Partial Exam
# Introduction to ML
# Spring 2024

**Duration: 70 minutes**
**Dr. Abbas Rammal**

---

## Problem 1: Association Mining Rules (12 points)

We provide the following database of store and customer transactions. Create all association rules with a minimum support of 60% and a minimum confidence of 60%.

1. Find all frequent itemsets using
   a) A priori
   b) FP-Growth.
2. We have the following rule: {Bananas, Bread} => {Milk}. Calculate the Lift, Leverage and conviction coefficients of this rule. Interpret.

| Transaction ID | Items |
|---|---|
| T1 | Apples, Milk, Bread, Cheese |
| T2 | Bananas, Milk, Bread, Yogurt |
| T3 | Apples, Bread, Eggs, Cheese |
| T4 | Bananas, Milk, Eggs, Yogurt |
| T5 | Apples, Milk, Bread, Yogurt |

---

## Problem 2: Decision Tree Classifier (8 points)

We consider a real dataset with three categorical attributes (Weather, Temperature, Day) and a binary target variable (Play: Yes or No). Here are 10 records in the dataset:

| Weather | Day | Temperature | Play |
|---|---|---|---|
| Sunny | Weekday | 30 | Yes |
| Sunny | Weekend | 32 | Yes |
| Rainy | Weekday | 20 | No |
| Sunny | Weekend | 28 | Yes |
| Rainy | Weekday | 18 | No |
| Rainy | Weekend | 22 | No |
| Sunny | Weekday | 26 | Yes |
| Sunny | Weekend | 35 | No |
| Rainy | Weekday | 17 | No |

a) Take the optimal split for the "Temperature" attribute in your decision tree: 27 and >27. Construct a decision tree based on this training data. For splitting, use information gain as measure for impurity. Build a separate branch for each attribute.

b) Split the Temperature attribute in two way using Gini Index.

---

Problem 1:

1) a) Apriori  minSup = 60% = 3/5

| TID | Items |
|---|---|
| T1 | A, M, Br, C |
| T2 | Ba, M, Br, Y |
| T3 | A, Br, E, C |
| T4 | Ba, M, E, Y |
| T5 | A, M, Br, Y |

| Items | Support |
|---|---|
| A | 3/5 ✓ |
| M | 4/5 ✓ |
| Br | 4/5 ✓ |
| C | 2/5 |
| Ba | 2/5 |
| Y | 3/5 ✓ |
| E | 2/5 |

C1

$L_1 = \{A, M, Br, Y\}$

| Itemset | Support |
|---|---|
| A, M | 2/5 |
| A, Br | 3/5 ✓ |
| A, Y | 1/5 |
| M, Br | 3/5 ✓ |
| M, Y | 3/5 ✓ |
| Br, Y | 2/5 |

$L_2 = \{\{A, Br\}, \{M, Br\}, \{M, Y\}\}$

| Itemset | Support |
|---|---|
| A, M, Br | 2/5 x |
| A, M, Y | 1/5 x |
| A, Br, Y | 1/5 x |
| M, Br, Y | 2/5 x |

Frequent Itemsets:

$\{A\}, \{M\}, \{Br\}, \{Y\}$

$\{A, Br\}\ \{M, Br\}, \{M, Y\}$

## Association Rules  minConf: 0.6

Sup(A → Br) = 3/3 ✓
Sup(Br → A) = 3/4 ✓
Sup(M → Br) = 3/4 ✓
Sup(Br → M) = 3/4 ✓
Sup(M → Y) = 3/4 ✓
Sup(Y → M) = 3/3 ✓

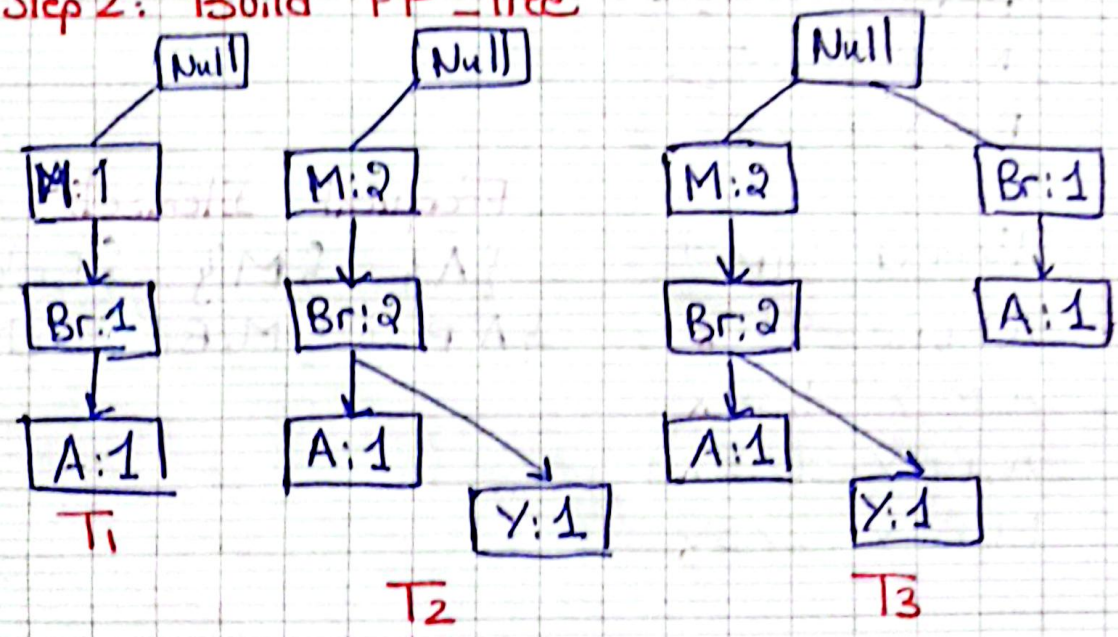
## b) FP_Growth

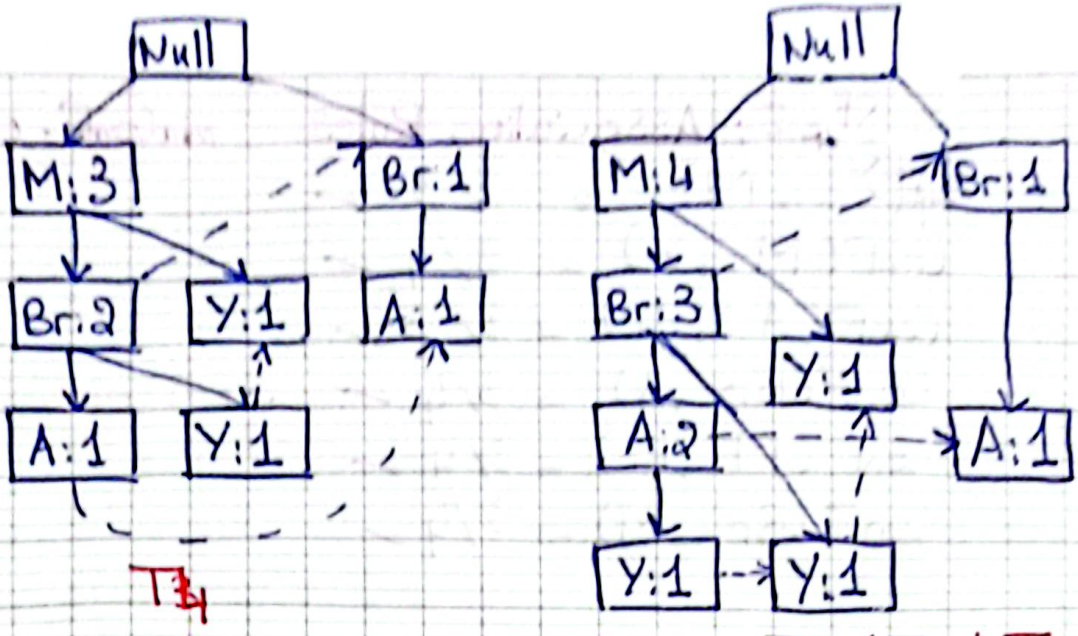
### Step 1: Count and Sort

| Item | Supp |
|---|---|
| M | 4 |
| Br | 4 |
| A | 3 |
| Y | 3 |

| TID | Items | Ordered Items |
|---|---|---|
| T1 | A,M,Br,C | M, Br, A |
| T2 | Ba, M, Br, Y | M, Br, Y |
| T3 | A, Br, E, C | Br, A |
| T4 | Ba, M, E, Y | M, Y |
| T5 | A, M, Br, Y | M, Br, A, Y |

### Step 2: Build FP_Tree






$T_4$

$T_5$ (Final Tree

## Step 3: Conditional Pattern Base and Conditional FP tree

| Items | Cond. Pattern Base | Cond. FP Tree |
|---|---|---|
| Y(3) | {M,Br,A:1} {M,Br:1} {M:1} | {M:3, ~~Br:2~~, ~~A:1~~} |
| A(3) | {M,Br:2} {Br:1} | {~~M:2~~, Br:3} |
| Br(4) | {M:3} | {M:3} |
| M(4) | { } | { } |

## Step 4: Frequent Pattern Generated

| Items | FP Generated |
|---|---|
| Y(3) | {Y:3} {Y,M:3} |
| A(3) | {A:3} {A,Br:3} |
| Br(4) | {Br:4} {M,Br:3} |
| M(4) | {M:4} |

Step 5: Association Rules  minConf = 0.6

Supp(Y → M) = 3/3 ✓
Conf(M → Y) = 3/4 ✓
Conf(A → Br) = 3/3 ✓
Conf(Br → A) = 3/4 ✓
Conf(M → Br) = 3/4 ✓
Conf(Br → M) = 3/4 ✓

2) {Ba, Br} → {M}

→ Lift({Ba, Br} → {M}) = $\frac{\text{Conf}(\{Ba, Br\} \to \{M\})}{\text{Support}\{M\}}$

$= \frac{1/5 \;/\; 1/5}{4/5} = \frac{5}{4} = 1.25$

Interpretation:
Lift = 1.25 > 1 ⇒ {Bananas, Bread} positively correlates with {Milk}, implying the presence of Ba, Br increase the likehood of buying Milk

→ Leverage({Ba, Br} → {M}) = Supp

Supp(Ba, Br, M) − Supp(Ba, Br) × Supp(M)

$= 1/5 - 1/5 \times 4/5 = 0.04$

Interpretation:
The rule occurs more frequent than expected if {Bananas, Bread} and {Milk} were independent

→ $\text{Conviction}(\{Ba, Br\} \rightarrow \{M\}) = \frac{1 - \text{Supp}(M)}{1 - \text{Conf}(Ba, Br \rightarrow M)}$

$= \frac{1 - 4/5}{1 - 1} = \infty$

Interpretation:

The rule is perfectly confident, as Milk always appears when Bananas and Bread are purchased

problem 2:

| | Weather | Day | Temp | Play |
|---|---|---|---|---|
| 1 | Sunny | WeekDay | >27 | Y |
| 2 | Sunny | WeekEnd | >27 | Y |
| 3 | Rainy | weekday | 27 | N |
| 4 | Sunny | weekend | >27 | Y |
| 5 | Rainy | weekday | 27 | N |
| 6 | Rainy | weekend | 27 | N |
| 7 | Sunny | weekday | 27 | Y |
| 8 | Sumy | weekend | >27 | N |
| 9 | Rainy | weekday | 27 | N |

Before Splitting

$E(B) = -\frac{4}{9}\log_2(\frac{4}{9}) - \frac{5}{9}\log_2(\frac{5}{9})$

$= 0.99$

Entropy for each attribute

1) Weather: Sunny: $\{Y^4, N^1\}$ Rainy: $\{Y^0, N^4\}$

$E(S) = -\frac{4}{5}\log_2(\frac{4}{5}) - \frac{1}{5}\log_2(\frac{1}{5})$

$= 0.72$

$E(R) = 0$

$E(\text{Weather}) = \frac{5}{9}(0.72) = 0.4$

$\text{Gain}(\text{Weather}) = 0.99 - 0.4 = 0.59$

2) Day: WD: $\{Y^2, N^3\}$ WE: $\{Y^2, N^2\}$

$E(WD) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5})$

$= 0.97$

$E(WE) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}) = 1$

$E(\text{Day}) = \frac{5}{9}(0.97) + \frac{4}{9} = 0.98$

$\text{Gain}(\text{Day}) = 0.99 - 0.98 = 0.01$

3) Temp: $27: \{Y^1, N^4\}$ $>27\ \{Y^3, N^1\}$

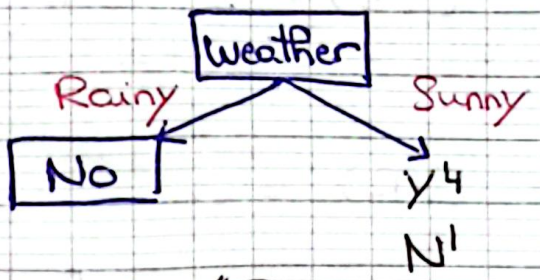
$$E(27) = -\tfrac{1}{5}\log_2(\tfrac{1}{5}) - \tfrac{4}{5}\log_2(\tfrac{4}{5}) = 0.72$$

$$E(>27) = -\tfrac{3}{4}\log_2(\tfrac{3}{4}) - (\tfrac{1}{4})\log_2(\tfrac{1}{4}) = 0.81$$

$$E(\text{Temp}) = \tfrac{5}{9}(0.72) + \tfrac{4}{9}(0.81) = 0.76$$

$$\text{Gain (Temp)} = 0.99 - 0.76 = 0.23$$

Weather have the highest Gain we Split on it first




New dataset filtered on "Sunny"

| Weather | Day | Temp | Play |
|---|---|---|---|
| Sunny | weekday | >27 | Y |
| Sunny | weekend | >27 | Y |
| Sunny | weekend | >27 | Y |
| Sunny | weekday | 27 | Y |
| Sunny | weekend | >27 | N |

$$E(B) = -\tfrac{4}{5}\log_2(\tfrac{4}{5}) - \tfrac{1}{5}\log_2(\tfrac{1}{5}) = 0.72$$

1) Day : WD $\{Y^2, N^0\}$ WE: $\{Y^2, N^1\}$

$E(WD) = 0$

$E(WE) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right)$

$= 0.918$

$E(Day) = \frac{3}{5}(0.918) = 0.5508$

$\therefore$ Gain (S, Day) = 0.72 − 0.5508 = 0.169

2) Temp: 27: $\{Y^1, N^0\}$ >27: $\{Y^3, N^1\}$

$E(27) = 0$

$E(>27) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right)$

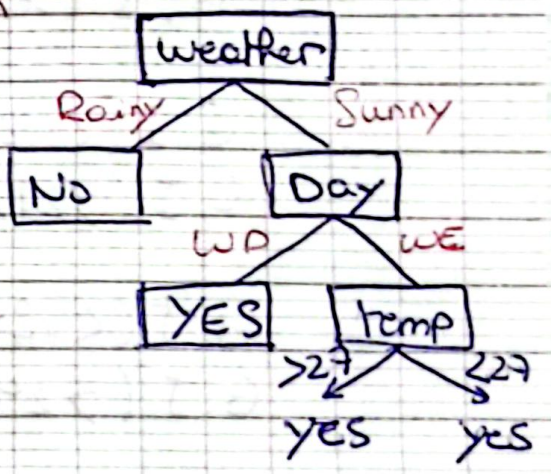
$= 0.81$

$E(Temp) = \frac{4}{5}(0.81) = 0.648$

Gain (Temp) = 0.72 − 0.648 = 0.072

Day have the highest gain

| Weather | Day | Temp | play |
|---|---|---|---|
| Sunny | WE | >27 | Y |
| Sunny | WE | >27 | Y |
| Sunny | WE | >27 | N |



b) 1) Sort Temp in increasing order

| 17 | 18 | 20 | 22 | 26 | 28 | 30 | 32 | 35 |
|---|---|---|---|---|---|---|---|---|
| 17 | 19 | 21 | 24 | 27 | 29 | 31 | 33 | |

| | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > | <= | > |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | 0 | 4 | 0 | 4 | 0 | 4 | 0 | 4 | 1 | 3 | 2 | 2 | 3 | 1 | 4 | 0 |
| N | 1 | 4 | 2 | 3 | 3 | 2 | 4 | 1 | 4 | 1 | 4 | 1 | 4 | 1 | 4 | 1 |
| G.I | ~~0.44~~ | | 0.38 | | 0.29 | | 0.17 | | 0.34 | | 0.44 | | 0.49 | | 0.44 | |

0.44

least G.ni

G.I(<=17) = ~~1/2~~ 0
G.I(>17) = 1 - [(1/8)² + (7/8)²] = 1/2 } avg = ~~3/9~~ 8/9 (1/2)

G.I(<=19) = 0
G.I)(>19) = 1 - [(4/7)² + (3/7)²] = 0.49 } avg = 7/9 (0.49) = 0.38

G.I(<=21) = 0
G.I(>21) = 1 - [(4/6)² + (2/6)²] = 0.44 } avg = 6/9 (0.44) = 0.29

G.I(<=24) = 0
G.I(>24) = 1 - [(4/5)² + (1/5)²] = 0.32 } 5/9 (0.32) = 0.17

G.I(<=27) = 1 - [(1/5)² + (4/5)²] = 0.32 } 5/9 (0.32)
G.I(>27) = 1 - [(3/4)² + (1/4)²] = 0.375 } + 4/9 (0.375) = 0.34

G.I(<=29) = 1 - [(2/6)² + (4/6)²] = 0.44 } 6/9 (0.44)
G.I(>29) = 1 - [(2/3)² + (1/3)²] = 0.44 } + 3/9 (0.44) = 0.44

G.I(<=31) = 1 - [(3/7)² + (4/7)²] = 0.49 } 7/9 (0.49)
G.I(>31) = 1 - [(1/2)² + (1/2)²] = 0.5 } + 2/9 (0.5) = 0.49

$GJ(L=33) = 1 - [(4/8)^2 + (4/8)^2] = 0.5$

$GJ(>33) = 0$

8/9 (0.5)
0.44