

Final Exam

Data Mining and Information Retrieval

Exercise 1 (7 pts)

Consider a dataset consisting of transactions recording the daily sales of products in various locations (Msila, Batna, Setif,) for different days over the year.

Transaction ID	Item	Store location	Date	Price
101123	Smartphone	Msila	06/22/24	45000
101123	Earbuds	Msila	06/22/24	8000
101123	Charger	Msila	06/22/24	3000
101124	Smartphone	Setif	06/23/24	45000
101124	Watch	Setif	06/23/24	12000
101124	Charger	Setif	06/23/24	3000
101125	Charger	Algiers	06/24/24	3000
101126	Battery	Batna	06/25/24	2000
101126	Earbuds	Batna	06/25/24	8000
101126	Charger	Batna	06/25/24	2000
101127	Smartphone	Msila	06/26/24	45000
101127	Watch	Msila	06/26/24	12000
101127	Battery	Msila	06/26/24	2000
101127	Earbuds	Msila	06/26/24	8000
101127	Charger	Msila	06/26/24	3000
101128	Smartphone	Batna	06/27/24	45000
101128	Charger	Batna	06/27/24	3000
101129	Smartphone	Setif	06/29/24	45000
101129	Earbuds	Setif	06/29/24	8000
101129	Charger	Setif	06/29/24	3000
101130	Watch	Algiers	06/30/24	12000
101130	Battery	Algiers	06/30/24	2000

- Transform the dataset into a new table structured as follows: **Transaction ID, List of Purchased Items.**
- Apply the Apriori algorithm to identify all association rules with a minimum support threshold of 50%.

Exercise 2 (6 pts)

Consider the following data :

$$X1 = 0 \quad X2 = 2 \quad X3 = 6 \quad X4 = 11$$

Use the Hierarchical Classification algorithm with the single-linkage method and visualize the results by plotting the dendrogram.

Exercise 3 (07 pts)

Consider a dataset where vertebrates are classified into one of the following categories: mammal, fish or reptile. The attribute set includes properties of a vertebrate such as its body temperature, method of reproduction, ability to fly, ability to live in water, etc.

Name	Body Temperature	Gives Birth	Aquatic Creature	Aeral Creature	Has Legs	Hibernates	Class Label
human	warm-blooded	yes	no	no	yes	no	mammal
python	cold-blooded	no	no	no	no	yes	reptile
salmon	cold-blooded	no	yes	no	no	no	fish
whale	warm-blooded	yes	yes	no	no	no	mammal
dragon	cold-blooded	no	no	no	yes	no	reptile
bat	warm-blooded	yes	no	yes	yes	yes	mammal
cat	warm-blooded	yes	no	no	yes	no	mammal
shark	cold-blooded	yes	yes	no	no	no	fish
eel	cold-blooded	no	yes	no	no	no	fish

Suppose we are given the following characteristics of a creature known as gila monster:

Name	Body Temperature	Gives Birth	Aquatic Creature	Aeral Creature	Has Legs	Hibernates	Class Label
gila monster	cold-blooded	no	no	no	yes	yes	?

Use the Naïve Bayesian classification model built from the dataset to determine the class to which the creature belongs. (Use Laplacian correction to avoid 0-probability case).

Dr. Tahar Mehenni

Final Exam

Data Mining and Information Retrieval

Solution

Exercise 1 (7 pts)

Transform the dataset into a new table structured as follows: **Transaction ID, List of Purchased Items.** (2 pts)

Transaction ID	List of Purchased Items
101213	Smartphone, Earbuds, Charger
101124	Smartphone, Watch, Charger
101125	Charger
101126	Battery, Earbuds, Charger
101127	Smartphone, Watch, Battery, Earbuds, Charger
101128	Smartphone, Charger
101129	Smartphone, Earbuds, Charger
101130	Watch, Battery

- Apply the Apriori algorithm to identify all association rules with a minimum support threshold of 50%.

1- Finding all the frequent itemsets: (3 pts)

We begin by finding all the 1-itemsets and their supports.

{ Smartphone } – 5/8, support = 0.625 { Watch } – 3/8, support = 0.375 { Battery } – 3/8, support = 0.375

{ Earbuds } – 4/8, support = 0.5 { Charger } – 7/8, support = 0.875

Itemsets with minimum support threshold 0.5 are:

{ Smartphone } – 5/8, support = 0.625 { Earbuds } – 4/8, support = 0.5 { Charger } – 7/8, support = 0.875

We generate all the 2-itemsets.

{ Smartphone, Earbuds } – 3/8, support = 0.375 { Smartphone, Charger } – 4/8, support = 0.5 { Earbuds, Charger } – 4/8, support = 0.

Itemsets with minimum support threshold 0.5 (Smartphone, Charger) and (Earbuds, Charger) with support=0.5

We generate all the 3-itemsets.

{ Smartphone, Earbuds, Charger } – 3/8, support = 0.375. **Not accepted.**

Frequent itemsets are : { Smartphone, Earbuds, Charger, (Smartphone, Charger), (Earbuds, Charger) }

2- Generation of rules from frequent 2-itemsets: (2 pts)

Smartphone → Charger confidence= 4/5=0.8 , Charger → Smartphone confidence= 4/7=0.57

Earbuds → Charger confidence= 4/4=1 , Charger → Earbuds confidence= 4/7=0.57

Exercise 2 (6 pts)

We use Euclidean distance (or Manhattan distance).

	x ₁	x ₂	x ₃	x ₄
x ₁	0			
x ₂	<u>2</u>	0		
x ₃	6	4	0	
x ₄	11	9	5	0

group (x₁, x₂) (2 pts)

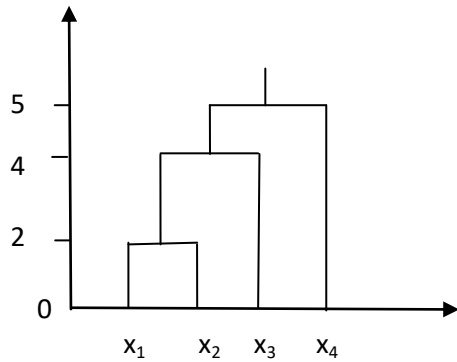
	x ₁ x ₂	x ₃	x ₄
x ₁ x ₂	0		
x ₃	<u>4</u>	0	
x ₄	9	5	0

group(x₃, x₁x₂) (1 pt)

	x ₁ x ₂ x ₃	x ₄
x ₁ x ₂ x ₃	0	
x ₄	<u>5</u>	0

group (x₄, x₁x₂x₃) (1 pt)

The dendrogram (2 pts)



Exercise 1 (7 pts)

Name	Body Temperature	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
salmon	cold-blooded	no	yes	no	no	no	fish
shark	cold-blooded	yes	yes	no	no	no	fish
eel	cold-blooded	no	yes	no	no	no	fish
human	warm-blooded	yes	no	no	yes	no	mammal
whale	warm-blooded	yes	yes	no	no	no	mammal
bat	warm-blooded	yes	no	yes	yes	yes	mammal
cat	warm-blooded	yes	no	no	yes	no	mammal
python	cold-blooded	no	no	no	no	yes	reptile
dragon	cold-blooded	no	no	no	yes	no	reptile
gila monster	cold-blooded	no	no	no	yes	yes	?

C1: mammal	$P(C1)$	0.44
C2: reptile	$p(C2)$	0.22
C3: fish	$P(C3)$	0.33

(1 pt)

gila monster	cold-blooded	no	no	no	yes	yes	$p(\text{gila Monster} C_i)$	$\frac{p(\text{gila Monster} C_i) * p(C_i)}{p(C_i)}$
C1	0.16	0.16	0.75	0.75	0.75	0.25	0.003	0.0013 (1.5 pt)
C2	1	1	1	1	0.5	0.5	0.25	0.0555 (1.5 pt)
C3	1	0.66	0.2	1	0.2	0.2	0.005	0.0017 (1.5 pt)

The creature gila monster belongs to the reptile class (1.5 pt)

Dr. T. Mehenni