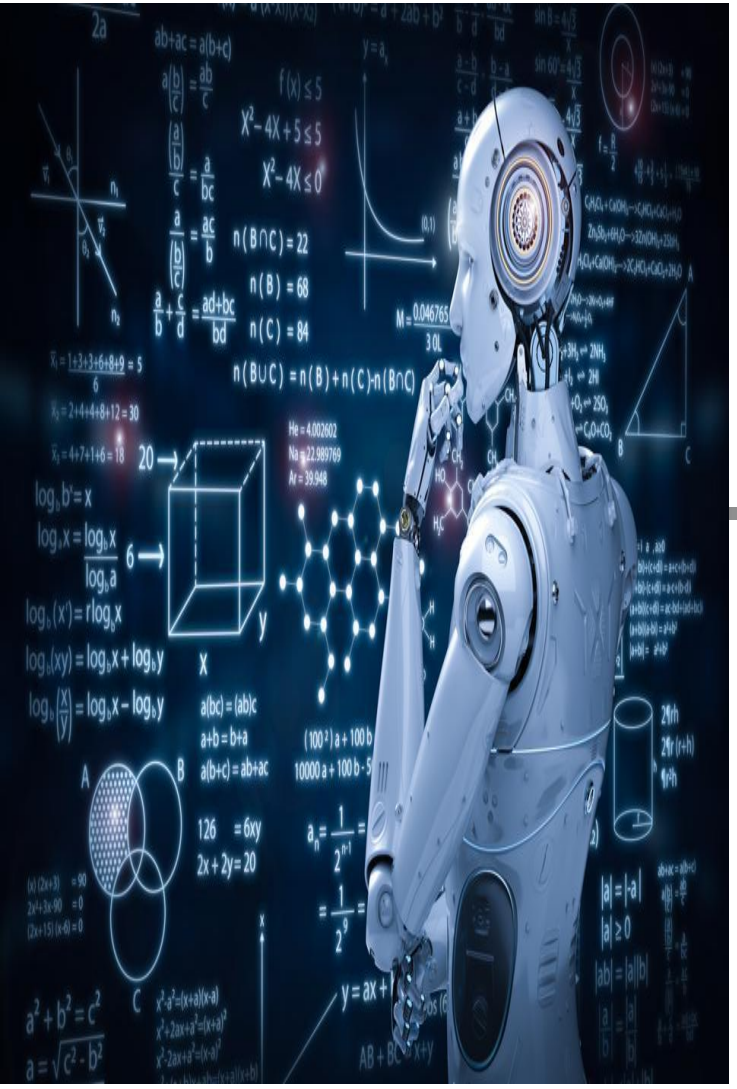


Practical Machine Learning

Day 17: Mar23 DBDA

Kiran Waghmare

Agenda

- Association
  - Apriori
  - Market Basket Analysis

BEST DEALS OF PASONS

PASONS Group

OFFER VALID ONLY ON
2021 JULY 22, 23 & 24

Nido One Plus
Stage-3 1800Gm /pc

Nutella Ferrero 350Gm /pc

Tiffany Wheat/Milk Rusk 2X335Gm

Sadia Chicken
Nuggets 270Gm /pc

Shurooq Sunflower
Oil 1.5Ltr /pc

Nellara Coconut
Oil 1Ltr /pc

Mehran Basmati
Kernel Rice 5Kg

Kwality Ice Cream 1Ltr Asstd /pc

Kwality Ice Cream 1Ltr
Caramel Butterscotch

London Dairy Vanilla/
Double Chocolate
500Ml /pc

Lacnor Milk 4x1Ltr Full Cream Uht

PASONS SUPERMARKET L.L.C
U03, Italy Cluster, International City, Dubai, U.A.E.

LOCATE US

FREE HOME DELIVERY

To get latest Offers & promotions send your 'Name' to +971 56 858 1222 to subscribe

SAVE MORE WOW PRICE

16-22 March 2015

MARINA
Tempura Chicken Nugget
(Assorted)
600g-1kg
Non-Member: RM13.50

AEON MEMBER
RM 12.90 /each
SAVE RM0.60

RM 5.99 /each
SAVE RM1.61

NESTLE
School Pack 140g
Normal Price: RM7.60

AEON MEMBER
RM 7.99 /each
SAVE RM1.91

GOODY
Pudding (Mango/Mix)
12x80g
Non-Member: RM9.90

RM 31.59 /each
SAVE RM1.40

SUSTAGEN
Junior 1+/ Kid 3+
650g
Normal Price: RM32.99

RM 28.99 /each
SAVE RM3.81

NESCAFE
Gold Blend
200g
Normal Price: RM32.80

RM 18.99 /each
SAVE RM3.91

BREEZE
Powder Detergent
(Assorted) 3.6kg - 4kg
Normal Price: RM22.90

RM 19.90 /each
SAVE MORE

AMBI PUR
2 Car Mini Clip 2ml + Gel Fresh 180g
(Lavender) Value Pack

*The promotion above is available at all AEON Stores.

What Is Frequent Pattern Analysis?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining

- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Why Is Freq. Pattern Mining Important?

- Freq. pattern: An intrinsic and important property of datasets
- Foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data

`Basket data'

A very common type of data; often also called *transaction data*.

Each record in a supermarket's transaction DB, for example, corresponds to a basket of specific items.

ID	apples,	beer,	cheese,	dates,	eggs,	fish,	glue,	honey,	ice-cream
1	1	1		1			1	1	
2			1	1	1				
3		1	1			1			
4		1				1			1
5					1		1		
6						1			1
7	1			1				1	
8						1			1
9			1		1				
10		1					1		
11					1		1		
12	1								
13			1			1			
14			1			1			
15								1	1
16				1					
17	1					1			
18	1	1	1	1				1	
19	1	1		1			1	1	
20					1				

Data

Transactions	Items
1	A B C D
2	A C D
3	A B C
4	C D E
5	A B C E

Matrix representation

Transactions \ Items	A	B	C	D	E
1	1	1	1	1	0
2	1	0	1	1	0
3	1	1	1	0	0
4	0	0	1	1	1
5	1	1	1	0	1

Execution of Apriori algorithm, $\varepsilon = 1$

Iteration 1

Candidates of size 1	*Support*
A	4
B	3
C	5
D	3
~~E~~	~~1~~

Iteration 2

Candidates of size 2	*Support*
A B	3
A C	4
A D	2
B C	3
~~B D~~	~~1~~
C D	3

Iteration 3

Candidates of size 3	*Support*
A B C	3
~~A B D~~	~~1~~
A C D	2

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \longrightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Discovering Rules

A common and useful application of data mining

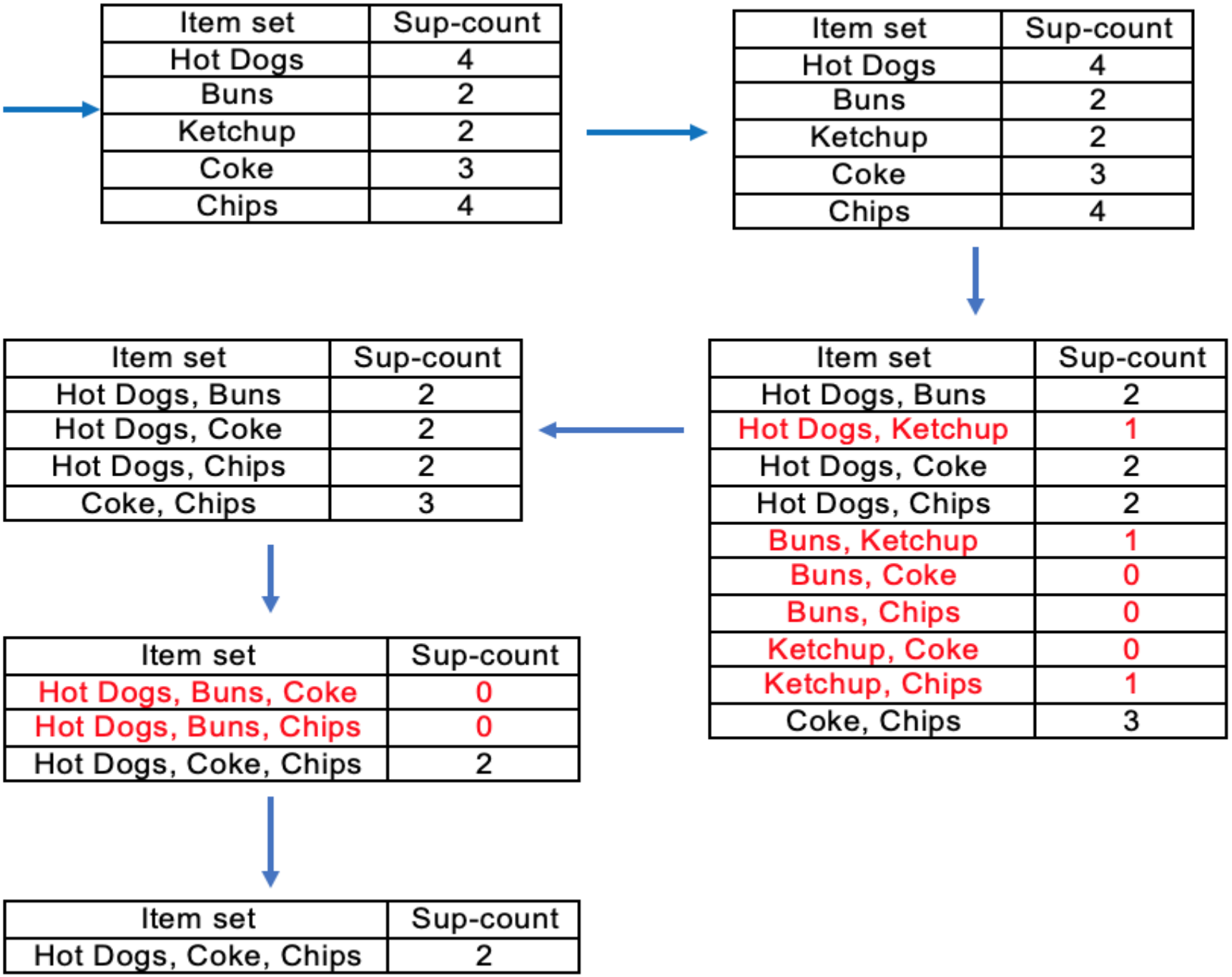
A `rule' is something like this:

If a basket contains apples and cheese, then it also contains beer

Any such rule has two associated measures:

1. *confidence* – when the `if' part is true, how often is the `then' bit true? This is the same as *accuracy*.
2. *coverage* or *support* – how much of the database contains the `if' part?

- If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.



Item set	Sup-count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

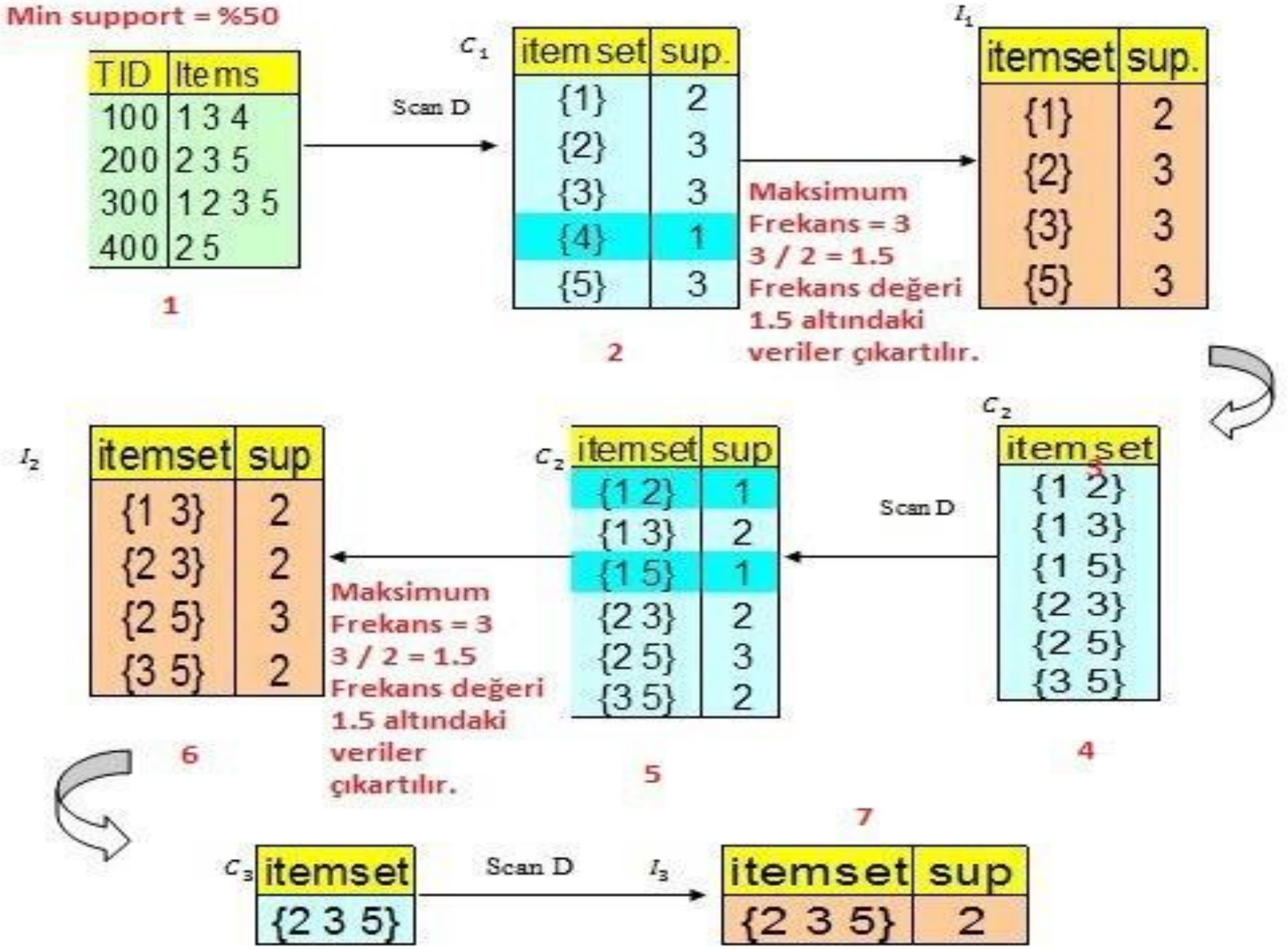
Item set	Sup-count
Hot Dogs	4
Buns	2
Ketchup	2
Coke	3
Chips	4

Item set	Sup-count
Hot Dogs, Buns	2
Hot Dogs, Ketchup	1
Hot Dogs, Coke	2
Hot Dogs, Chips	2
Buns, Ketchup	1
Buns, Coke	0
Buns, Chips	0
Ketchup, Coke	0
Ketchup, Chips	1
Coke, Chips	3

Item set	Sup-count
Hot Dogs, Buns	2
Hot Dogs, Coke	2
Hot Dogs, Chips	2
Coke, Chips	3

Item set	Sup-count
Hot Dogs, Buns, Coke	0
Hot Dogs, Buns, Chips	0
Hot Dogs, Coke, Chips	2

Item set	Sup-count
Hot Dogs, Coke, Chips	2


Min support = %50
TID
Items
100
1 3 4
200
2 3 5
300
1 2 3 5
400
2 5
1
Scan D
C_1
itemset
sup.
{1}
2
{2}
3
{3}
3
{4}
1
{5}
3
2
Maksimum
Frekans = 3
3 / 2 = 1.5
Frekans değeri
1.5 altındaki
veriler çıkartılır.
I_1
itemset
sup.
{1}
2
{2}
3
{3}
3
{5}
3
C_2
itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}
4
Scan D
C_2
itemset
sup
{1 2}
1
{1 3}
2
{1 5}
1
{2 3}
2
{2 5}
3
{3 5}
2
5
Maksimum
Frekans = 3
3 / 2 = 1.5
Frekans değeri
1.5 altındaki
veriler
çıkartılır.
I_2
itemset
sup
{1 3}
2
{2 3}
2
{2 5}
3
{3 5}
2
6
7
C_3
itemset
{2 3 5}
Scan D
I_3
itemset
sup
{2 3 5}
2