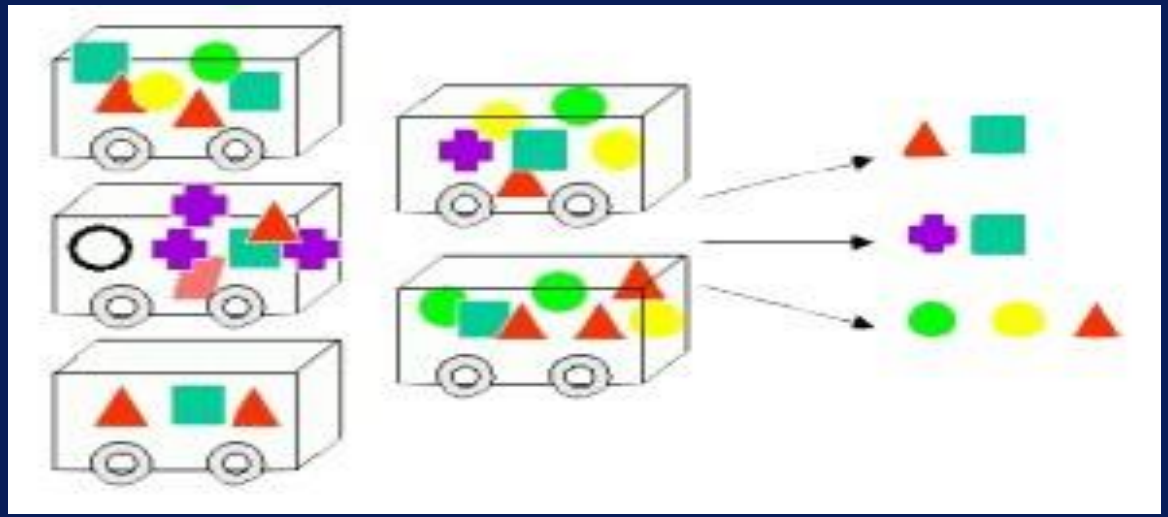
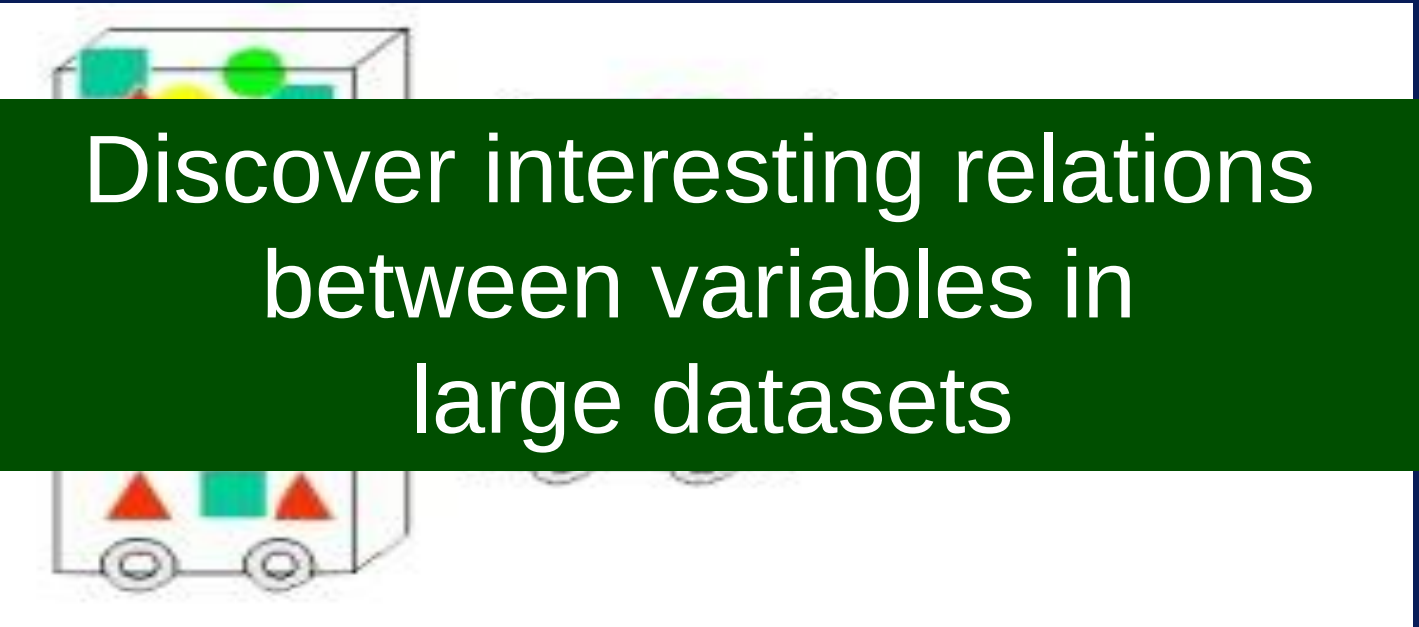
Mining Association Rules

Standard separate-and-conquer method



Mining Association Rules

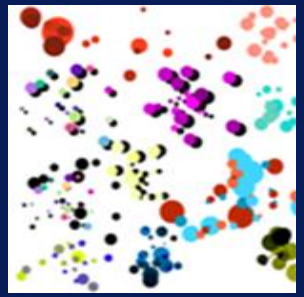
- **Standard separate-and-conquer method**



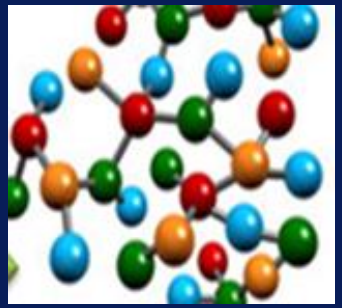
Discover interesting relations between variables in large datasets

Is there any insight in these baskets?

Why Association Rules?



1111100001000011110000011
0101010110000011111111111
0101010000111110101010101
0101010100111111110000001
0101100000001010110101010
1010101010101010110101010



From Transaction Data to Discovering Hidden Associations and Discerning Insight

Market Basket Analysis

71%

43%

26%

Market Basket Analysis

71%

43%

26%

What % of customers that buy milk buy eggs?

Market Basket Analysis

71%

43%

26%

Out of the Customers who bought milk
71% bought bread
43% bought eggs
26% included Coffee

Market Basket Analysis

What % of customers that buy milk buy eggs?

26%

What % of customers that buy milk and eggs bought cake mix?

26% included Coffee

What kinds of questions can we answer?

Is cereal typically purchased with bananas?

Does the brand/type of cereal matters?

Where should cookies be placed in the store to maximize the sales?

Huggies and Chuggies

?

If a customer is looking to buy a black purse
– what should we offer her?

Lesson #2

- **Mining Association Rules**

Mining Association Rules

- **Standard separate-and-conquer method**
- **Looking at every possible combination of attributes, every combination of values on right-hand side**
- **Problems:**
  - Computational complexity
  - Resulting in enormous number of rules
    - pruned based on support and confidence

What are Item sets?

Item

One attribute-value pair

Item set

All items occurring in a rule

Attribute=purse
Value=black color

Creating the Item Sets for Association Rule learning

Coverage = Support

Number of instances rule predicts correctly

Item sets

Accuracy = Confidence

proportion of the number of instances that the rule applies to

Goal for Rule Generation

Produce only rules that exceed pre-defined support

Find all item sets with the given minimum support

Enormous possible number of Rules
Need to restricts

sets

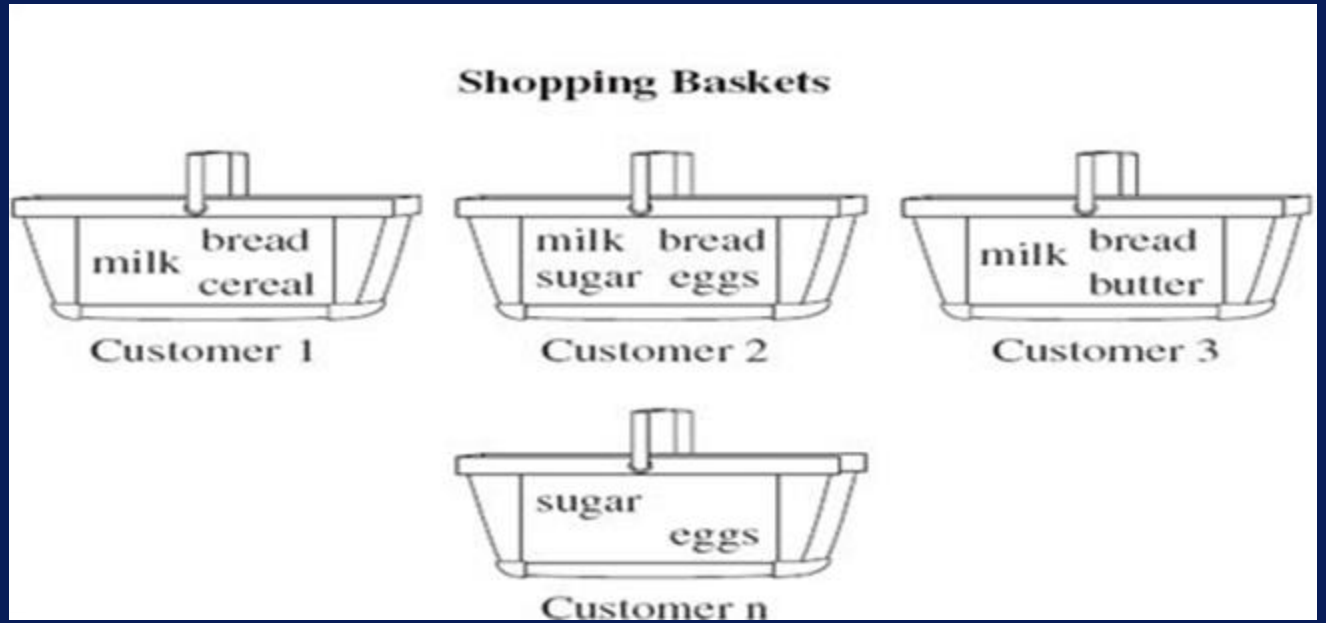
For Specified Minimum Support

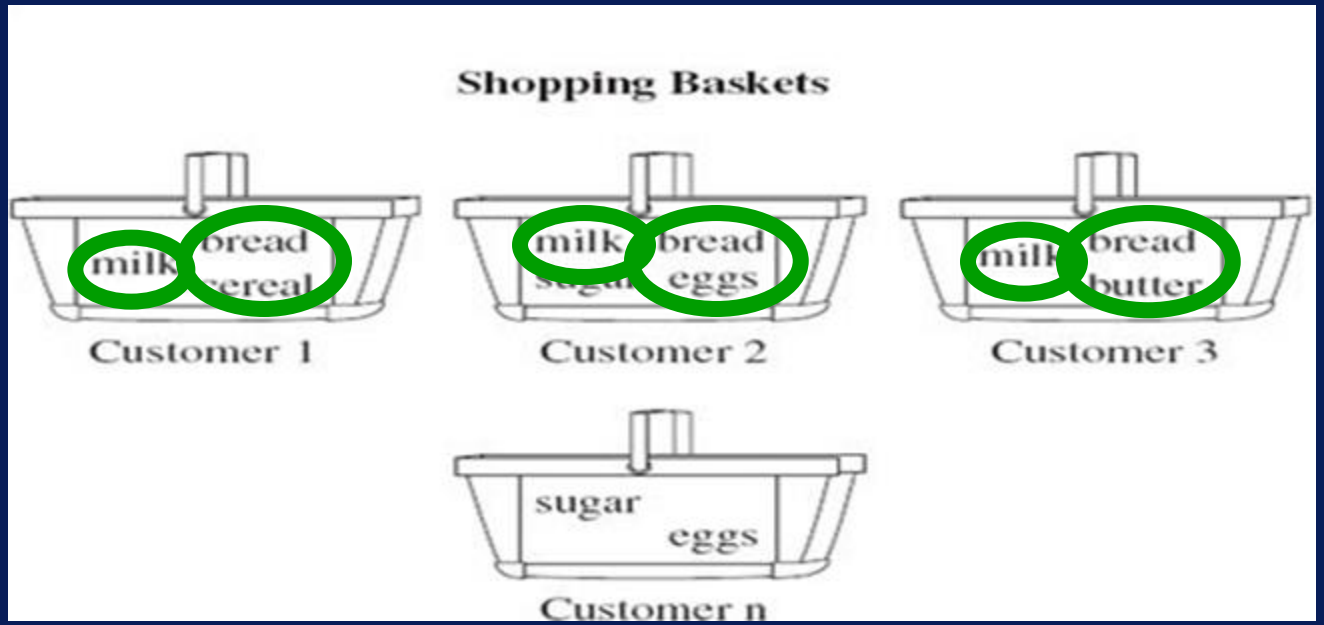
Generate one item sets

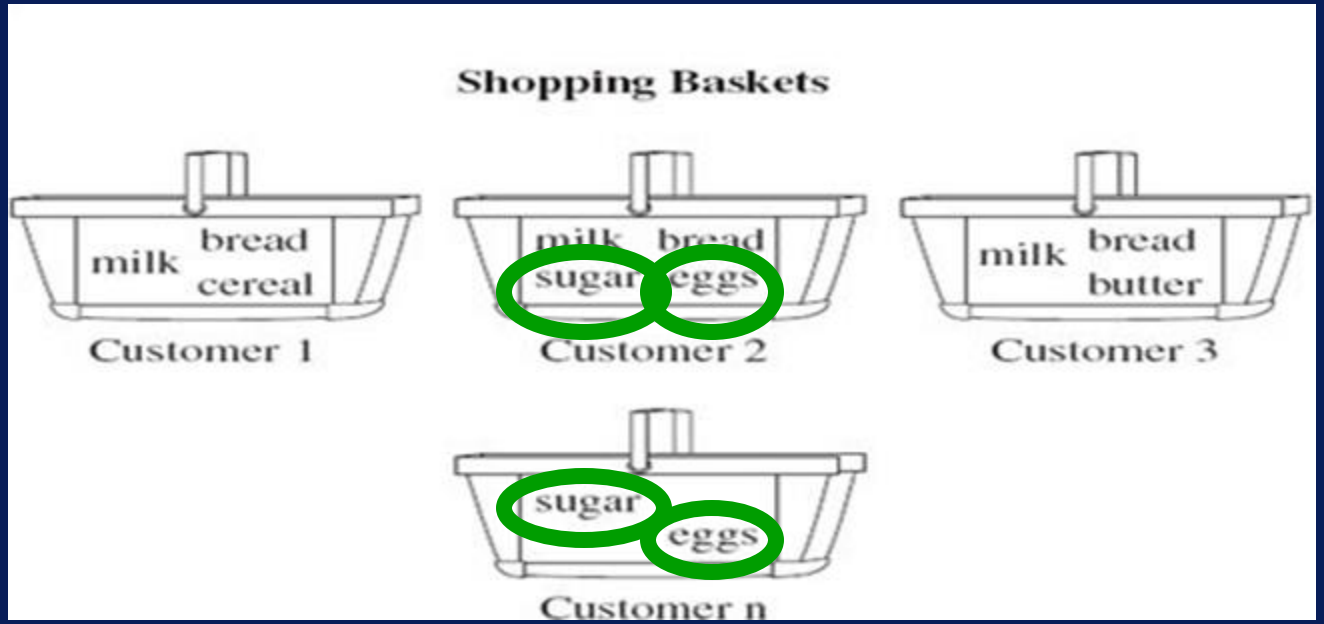
-> two item sets

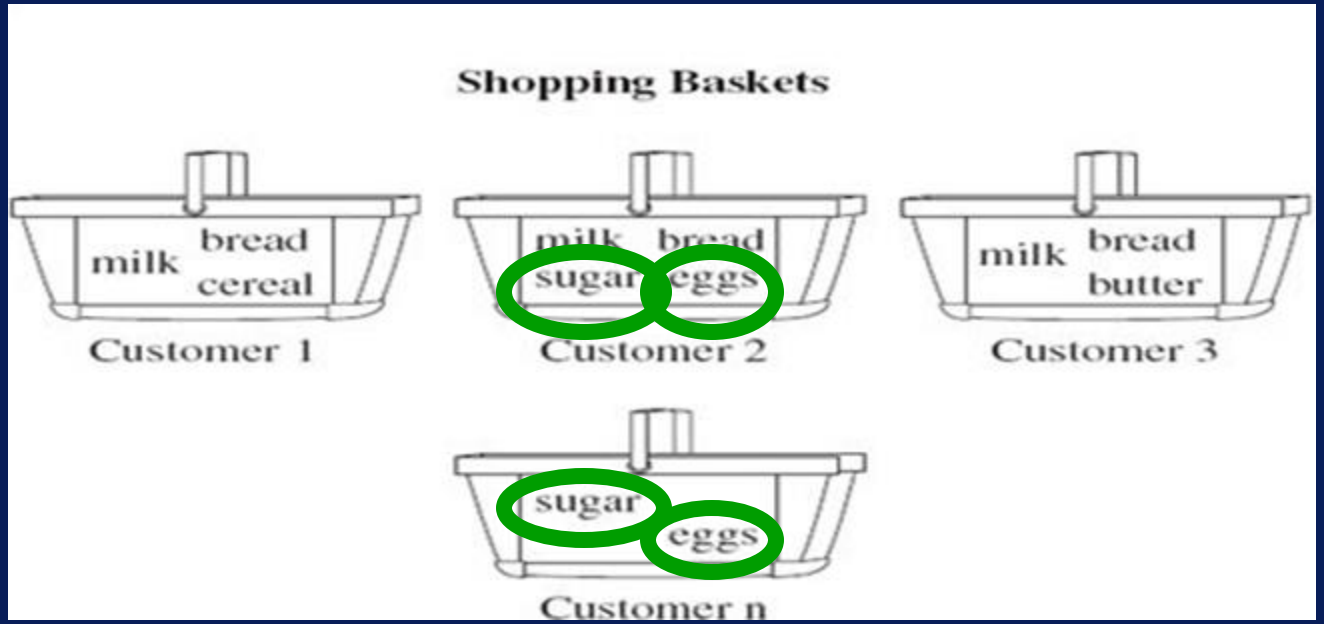
-> three item sets, etc.

We only generate rules
with specified min Support


Shopping Baskets
milk
bread
cereal
Customer 1
milk
bread
sugar
eggs
Customer 2
milk
bread
butter
Customer 3
sugar
eggs
Customer n


Shopping Baskets
milk
bread
cereal
Customer 1
milk
bread
eggs
Customer 2
milk
bread
butter
Customer 3
sugar
eggs
Customer n


Shopping Baskets
milk
bread
cereal
Customer 1
milk
bread
sugar
eggs
Customer 2
milk
bread
butter
Customer 3
sugar
eggs
Customer n



What is the Support and Accuracy for sugar and eggs?

Basket	Milk	Bread	Cereal	Sugar	Eggs	Butter
Customer 1	1	1	1			
Customer 2	1	1		1	1	
Customer 3	1	1				1
Customer 4				1	1	

Rule Examples:

Sugar->Eggs

Sugar, Eggs -> Milk

Milk - >Bread, Butter

Milk - >Bread, Cereal

Lesson #3

Mining Association Rules on the Weather data example

Weather Data Set

5 Attributes

Class Attribute

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

14 Instances

Utilizing the Weather data Example

Let’s create item sets with support =2

Weather Data Example

One-item sets	Two-item sets	Three-item sets	Four-item sets
Outlook = Sunny(5)	Outlook = Sunny Temp = Mild (2)	Outlook = Sunny Temp = Hot Humidity= High (2)	Outlook = Sunny Temp = Hot Humidity= High Play = No(2)
Temp=Cool(4)	Outlook = Sunny Humidity = High (3)	Outlook = Sunny Humidity=High Windy = False(2)	Outlook=Rainy Temp=Mild Windy-False Play=Yes(2)

Item sets
with support =2

Weather Data Set

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Weather Data Example

One-item sets	Two-item sets	Three-item sets	Four-item sets
Outlook = Sunny(5)	Outlook = Sunny Temp = Mild (2)	Outlook = Sunny Temp = Hot Humidity= High (2)	Outlook = Sunny Temp = Hot Humidity= High Play = No(2)
Temp=Cool(4)	Outlook = Sunny Humidity = High (3)	Outlook = Sunny Humidity=High Windy = False(2)	Outlook=Rainy Temp=Mild Windy-False Play=Yes(2)

Item sets
with support =2

Weather Data Set

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Weather Data Example

One-item sets	Two-item sets	Three-item sets	Four-item sets
Outlook = Sunny(5)	Outlook = Sunny Temp = Mild (2)	Outlook = Sunny Temp = Hot Humidity= High (2)	Outlook = Sunny Temp = Hot Humidity= High Play = No(2)
Temp=Cool(4)	Outlook = Sunny Humidity = High (3)	Outlook = Sunny Humidity=High Windy = Weak (2)	Outlook=Rainy Temp=Mild Windy=Weak Play=Yes(2)

Item sets
with support =2

Weather Data Set

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Total Number of Item Sets For Weather Data Set

With minimum support = 2

12 one-item sets

47 two-item sets

39 three-item sets

6 four-item sets

0 five-item sets

Total number of item sets

With minimum support = 2

12 one-item sets

47 two-item sets

39 three-item sets

6 four-item sets

0 five-item sets

Once all item sets with minimum support have been generated they are turned into association rules

Association Rules

Example: 3 item set with coverage=4

Humidity = Normal, Windy = False, Play = Yes (4)

How to make rules
From Item Sets?

Association Rules

Example: 3 item set with coverage=4

Humidity = Normal, Windy = False, Play = Yes (4)

Produces seven (2N-1) potential rules!

(2N-1) Potential Rules for

Humidity = Normal, Windy = False, Play = Yes (4)

Produces Seven Rules:

If Humidity=Normal and Windy=False then Play=Yes	4/4
If Humidity=Normal and Play=Yes then Windy=False	4/6
If Windy=False and Play=Yes then Humidity=Normal	4/6
If Humidity=Normal then Windy=False and Play=Yes	4/7
If Windy=False then Humidity=Normal and Play=Yes	4/8
If Play=Yes then Humidity=Normal and Windy=False	4/9
If True then Humidity=Normal and Windy=False and Play=Yes	4/14

Lesson #4

Specifying Support and Coverage

We can specify

Rules with support > 1 and confidence = 100%

Rule #	Association Rule	Support	Confidence
1	Humidity=Normal and Windy=False =>Yes	4	100%
2	Temp=Cool =>Humidity=Normal	4	100%
3	Outlook=Overcast =>Play=Yes	4	100%
4	Temp=Cold and Play = Yes => Humidity=Normal	3	100%
…..	. . . . .		
50	Outlook=Sunny and Temp=Hot =>Humidity=High	2	100%

We can specify

Rules with support > 1 and confidence = 100%

Rule #	Association Rule	Support	Confidence
1	Humidity=Normal and Windy=False =>Yes	4	100%
2	Temp=Cool =>Humidity=Normal	4	100%
3	Outlook=Overcast =>Play=Yes	4	100%
	Normal	3	100%
	y=High	2	100%

Support in decreasing Order

We can specify

Rules with support > 1 and confidence = 100%

Rule #	Association Rule	Support	Confidence
1	Humidity=Normal and Windy=False =>Yes	4	100%
2	Temp=Cool =>Humidity=Normal	4	100%
3	Outlook=Overcast =>Play=Yes	4	100%
			100%
			100%

58 Rules Total with confidence=1:

3 Rules with coverage = 4

5 Rules with coverage = 3

50 Rules with coverage = 2

How to efficiently find all frequent item sets?

First find one-item sets

- Use them to generate two-item sets
- Use two-item sets to generate three-item sets
- Use three-item sets to generate 4-item sets….

How to efficiently find all frequent item sets?

- **If (A B) is frequent item set then**
  - (A) and (B) have to be frequent item sets as well

- **if X is frequent k-item set than**
  - all (k-1)- item subsets of X are also frequent
  - compute k-item set by merging (k-1)-item sets

Evaluating Association Rules

For each Rule: X =>Y and N = number of Item Sets for the rule can calculate the:

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

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

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

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

Basket	Milk	Bread	Cereal	Sugar	Eggs	Butter
Customer 1	1	1	1			
Customer 2	1	1		1	1	
Customer 3	1	1				1
Customer 4				1	1	

Item set {Milk,Bread} has a support of 75%

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

Basket	Milk	Bread	Cereal	Sugar	Eggs	Butter
Customer 1	1	1	1			
Customer 2	1	1		1	1	
Customer 3	1	1				1
Customer 4	1			1	1	

Item set {Milk,Bread} has a support of 75%
Rule: Milk->Bread has accuracy of 0.8

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

Association Rule Mining Challenges

Computational complexity

Pruning based on support and confidence

Generating a pre-specified number of rules

Data format inefficiency