

# Market Basket Analysis



## INTRODUCTION

- Small service businesses create one-to-one relationships with their customers in order to compete and succeed.
- Small businesses build customer relationships by remembering their names, needs, preferences, and learning from past interactions.

- Large business organizations must deploy a system to emulate this ability of a good small business to create relationships with customers.
- How can a large business achieve this, when customers may never directly interact with the organization, or if they do interact, it is with different employees each time?

- The answer is – by using data mining tools!
- DATA MINING (DM) is the exploration of large amounts of data to find useful and actionable patterns, relationships and rules.
- The typical tasks in DM applications are – summarization, classification or supervised learning, clustering or unsupervised learning, prediction or forecasting, and link analysis.

## LINK ANALYSIS

Association discovery  
find rules about items  
appearing together in  
a list.

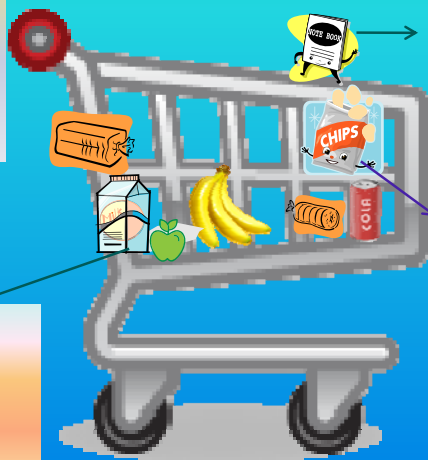
Market Basket  
Analysis –  
Association discovery  
from customer  
transactions data.

Sequence discovery  
Temporal association  
discovery.

## Market Basket Analysis (MBA)

How is customer  
buying affected by  
neighborhood  
demographics?

Are bananas  
purchased  
with apples,  
bread, milk?



Where should  
stationery be  
placed to  
maximize sales?

Is cola  
typically  
purchased  
with chips and  
bagels?

## WHAT IS MBA?



Bread  
Milk  
Apple  
Bagels  
Chips  
Notebook



Bread  
Milk  
Apple  
Bananas  
Cola  
Chips



Milk  
Bananas  
Bagels  
Cola  
Chips  
Notebook

MBA is a tool for determining customer buying habits by finding associations between different items in customers' shopping carts.

## WHAT IS MBA?

MBA is an 'undirected' or 'unsupervised' method, i.e., it has no response variable and no predictor variables.

## WHAT DOES MBA DO?

Customer	Bread	Milk	Apple	Banana	Bagels	Chips	Cola	Notebook
1	1	1	1	0	1	1	0	1
2	1	1	1	1	0	0	1	0
3	0	1	0	1	1	1	0	1

Given a database of customer transactions, MBA finds groups of items that are frequently purchased together.

## THE MBA PROCESS

- INPUT to MBA = large database of customer transactions
- OUTPUT of MBA = useful and actionable information on customer purchasing behavior
- USES OF MBA = improve store arrangement, determine which products to use for promotions, which products to put next to promotional items, ...

## WHO CAN USE MBA?

- Credit card companies (card offering)
- Banking services (marketing various services)
- DVD rental businesses (marketing, inventory)
- Cell phone companies (marketing plans)
- Medical diagnosis/research

## WHO CAN USE MBA?

- Website navigation analysis
- Homeland security
- Casinos (slot placement on floor, marketing)
- Slot manufacturers (type of features to have in new slot games)

## Association Rules - DEFINITION

•AR is an “if-then” statement : If a customer buys Product A, then the customer also buys product C.

This AR is mathematically expressed as:

$$A \Rightarrow C$$

A = antecedent, C = consequent\*

\* Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining Association Rules in Very Large Databases. Proc. of the ACM SIGMOD Conference on Management of Data, pp. 207-216, Washington, D.C.

## AR - terminology

I = Itemset

Support(I) = #(transactions containing I)/(Total # of transactions)

$\sigma$  = minimum support set by user (say 20%)

Frequent Itemset = any I with support(I)  $\geq \sigma$

## Support and Confidence of an AR

$$\begin{aligned}\text{Support}(A \Rightarrow C) &= \frac{\text{\# of transactions with both A and C}}{\text{total \# of transactions}} \\ &= \text{Probability that A and C occur together} \\ &= P(A \cap C)\end{aligned}$$

$$\begin{aligned}\text{Confidence}(A \Rightarrow C) &= \frac{\text{\# of transactions with both A and C}}{\text{\# of transactions with A}} \\ &= \frac{P(A \cap C)}{P(A)} \\ &= \text{conditional probability of C, given A} \\ &= P(C|A)\end{aligned}$$

## Lift of an AR

$$\begin{aligned}\text{Lift}(A \Rightarrow C) &= \frac{P(A \cap C)}{P(A)P(C)} \\ &= \text{Probability that A and C occur together,} \\ &\quad \text{divided by same probability, assuming A} \\ &\quad \text{and C to be independent}\end{aligned}$$

### NOTE:

Look for AR with lift > 1.

If  $\text{Lift}(A \Rightarrow C) < 1$ , then  $\text{Lift}(A \Rightarrow \text{Not } C) > 1$

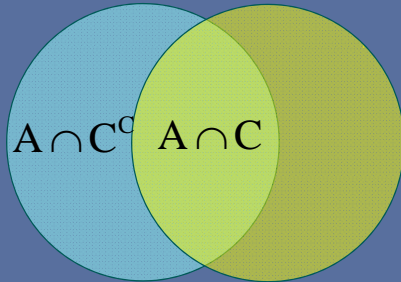
If  $Lift(A \Rightarrow C) < 1$ , then  $Lift(A \Rightarrow \text{Not } C) > 1$

$$Lift(A \Rightarrow C) = \frac{P(A \cap C)}{P(A)P(C)} < 1$$

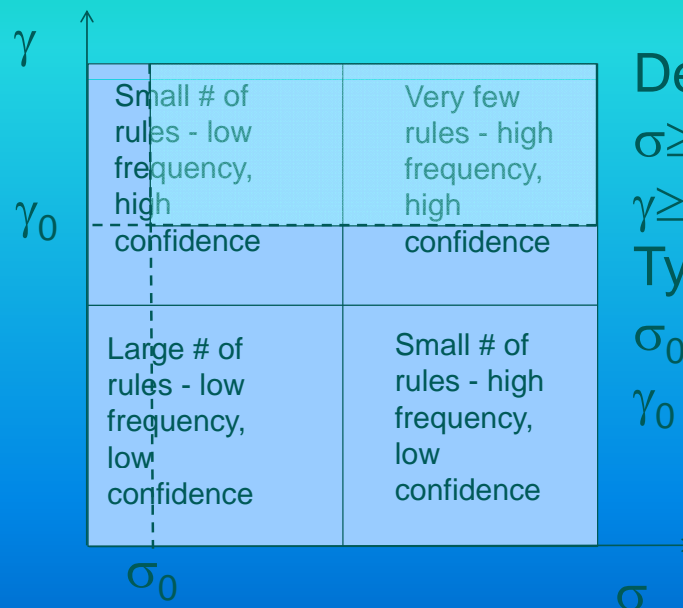
$$P(A \cap C) < P(A)P(C)$$

$$\begin{aligned} P(A \cap C^c) &= P(A) - P(A \cap C) \\ &> P(A) - P(A)P(C) \\ &= P(A)P(C^c) \end{aligned}$$

$$\frac{P(A \cap C^c)}{P(A)P(C^c)} = Lift(A \Rightarrow C^c) > 1$$



## Desired ARs



Desired rules:

$$\sigma \geq \sigma_0$$

$$\gamma \geq \gamma_0$$

Typically

$$\sigma_0 = 5 - 10\%$$

$$\gamma_0 = 75 - 90\%$$

## Remarks on Association Rules

- Association Rules can be actionable (useful rules that suggest product placement, e.g., in a DVD rental place, put action next to Comedy)
- Association Rules can be trivial – e.g., may reflect marketing promotions or product bundling.
- Association Rules may have no explanation and may not be actionable.

## Types of Association Rules

- Dimensionality 1 or more than 1 (associations on single items purchased vs. associations between 2 or more items)
- Associations between categorical vs. quantitative data
- Single level vs. multi-level (e.g., what cola brands go with what chips?)

## Data File Formats for AR Mining

Relational - <Tid,item>

<Customer 1, Item1>

<Customer 1, Item2>

<Customer 2, Item2>

<Customer 2, Item3>

Compact - <Tid,itemset>

<Customer 1, Item1,Item2>

<Customer 2, Item2,Item3>

Item = single item, Itemset = set of items

## Algorithms for Computing AR

### Basic Apriori Algorithm

1. Find the **frequent itemsets**, i.e., find all  $I$  with **support**  $\geq \sigma$ . [Note: if  $I=(A,B)$  is frequent, then both A and B are also frequent].
2. Find frequent itemsets with cardinality ranging from 1 to  $k$ .
3. Generate ARs from frequent itemsets.

Let us look at a simple example of sales at a roadside stand selling yams, cranberry sauce, and turkey.

## Example 1

<b>TID</b>	<b>Basket of Items Purchased</b>
1	yams, cranberry sauce, bread
2	yams, turkey
3	yams, cranberry sauce, turkey
4	yams, turkey
5	cranberry sauce, turkey
6	yams, cranberry sauce, turkey
7	yams, turkey, bread
8	cranberry sauce, turkey

## Association Rules for Example 1

Let T = Turkey, Y = Yams, C = Cranberry sauce  
B = bread,  $\sigma_0 = 60\%$

Frequent Itemsets	Frequency	Support
T	7	87.5
Y	6	75
C	5	62.5
B	2	25
T,Y	5	62.5
T,C	4	50
Y,C	3	37.5
T,Y,C	2	25

STATS24x7.com© 2010 ADI-NV, INC

25

## Association Rules for Example 1

There are only two Association Rules with the minimum support of 60%; since the item B is not frequent, we do not need to look at (B,T), (B,Y), (B,C) etc.

AR	Support	Confidence
$T \Rightarrow Y$	62.5	71.4
$Y \Rightarrow T$	62.5	83.3

STATS24x7.com© 2010 ADI-NV, INC

26

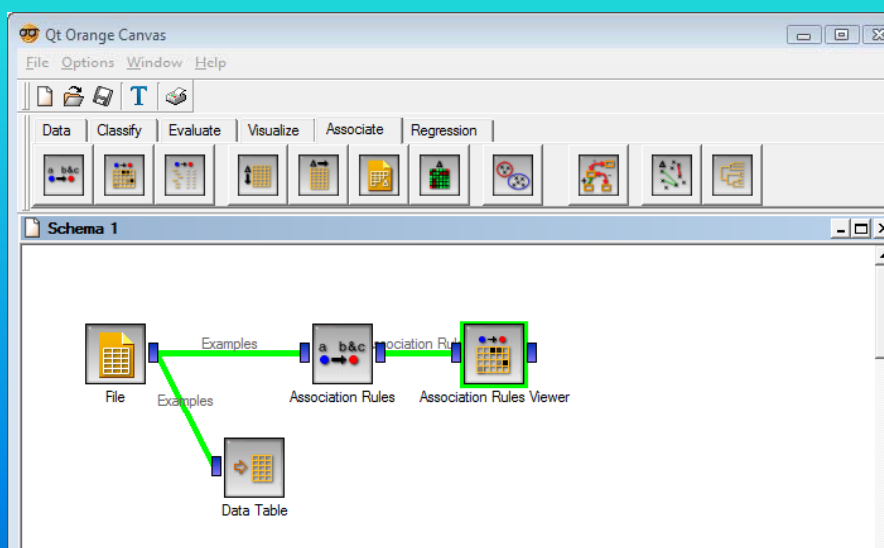
AR's for Example 1 can be found/evaluated by software such as ORANGE on data file shown below:

Customer	Turkey	Yams	Cranberry Sauce
1	0	1	1
2	1	1	0
3	1	1	1
4	1	1	0
5	1	0	1
6	1	1	1
7	1	1	0
8	1	0	1

STATS24x7.com© 2010 ADI-NV, INC

27

AR's for Example 1 using software package ORANGE



STATS24x7.com© 2010 ADI-NV, INC

28

## AR's for Example 1 using software package ORANGE

Supp	Conf	Lift	Rule
0.250	1.000	1.600	Turkey=1 Yams=0 -> Cranberry Sauce=1
0.250	1.000	1.600	Yams=0 -> Cranberry Sauce=1
0.375	0.600	1.600	Turkey=1 Yams=1 -> Cranberry Sauce=0
0.500	0.571	0.914	Turkey=1 -> Cranberry Sauce=1
0.375	0.500	1.333	Yams=1 -> Cranberry Sauce=0
0.375	0.500	0.800	Yams=1 -> Cranberry Sauce=1

## Example 2 – Pizza sales

A small pizzeria has following cheese pizza sales in a month: O = onions, S = sausage, P = pepperoni (total pizzas sold = 5000)

300 pizzas with only O

300 pizzas with only S

300 pizzas with only P

500 pizzas with only O and S

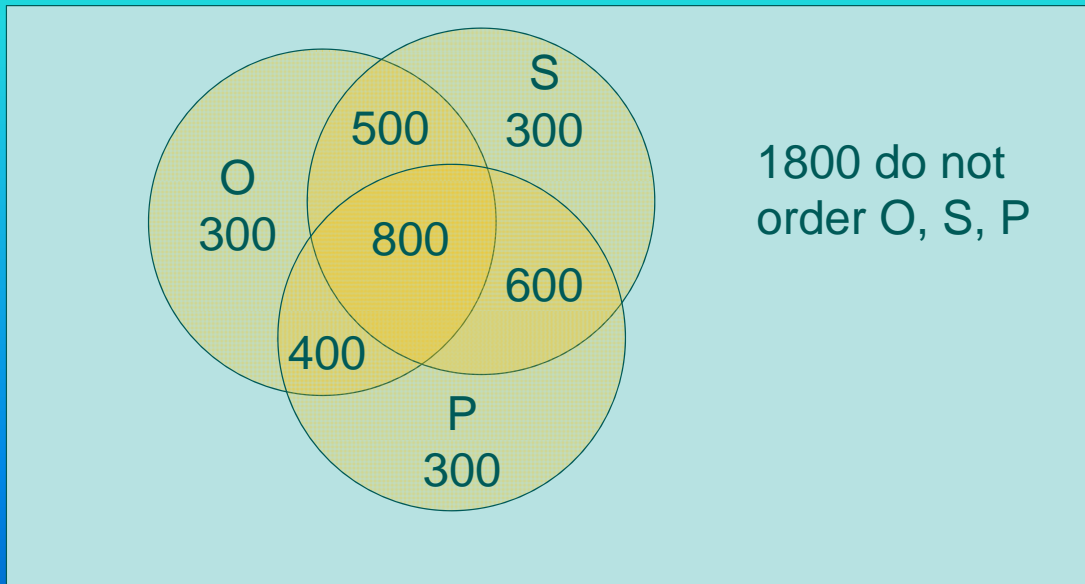
400 pizzas with only O and P

600 pizzas with only S and P

800 pizzas with O, S, and P

1800 with neither of the three (O, S, P)

## Example 2 – Pizza sales



STATS24x7.com© 2010 ADI-NV, INC

31

## Association Rules for Example 2

AR	#	Support P(AC)	P(A)	Confidence $P(C A)=P(AC)/P(A)$	P[C]	Lift $P(AC)/(P(A)P[C])$
$O \Rightarrow S$	1300	0.26	0.40	0.65	0.44	1.48
$S \Rightarrow O$	1300	0.26	0.44	0.59	0.40	1.48
$O \Rightarrow P$	1200	0.24	0.40	0.60	0.57	1.05
$P \Rightarrow O$	1200	0.24	0.42	0.57	0.40	1.43
$S \Rightarrow P$	1400	0.28	0.44	0.64	0.57	1.11
$P \Rightarrow S$	1400	0.28	0.42	0.67	0.44	1.52
$O,S \Rightarrow P$	800	0.16	0.26	0.62	0.42	1.47
$O,P \Rightarrow S$	800	0.16	0.24	0.67	0.40	1.67
$S,P \Rightarrow O$	800	0.16	0.28	0.57	0.65	0.88

STATS24x7.com© 2010 ADI-NV, INC

32

## Association Rules for Example 2

From the above table, we see that –

The best AR for this pizzeria is

$O \Rightarrow S$

(if a customer orders Onions, then customer also orders Sausage on pizza)

with

Support = 26%

Confidence = 65%

Lift = 1.48

## Statistical Analysis of Association Rules

We next look at the relationship between the standard measure of dependence (Chi-square statistic) and (support, confidence, lift) of Association Rules.

# Chi-square statistic for a 2x2 table

Observed Table

	B	Not B	
A	$NP(A \cap B)$	$NP(A \cap B^c)$	$NP(A)$
Not A	$NP(A^c \cap B)$	$NP(A^c \cap B^c)$	$NP(A^c)$
Total	$NP(B)$	$NP(B^c)$	$N$

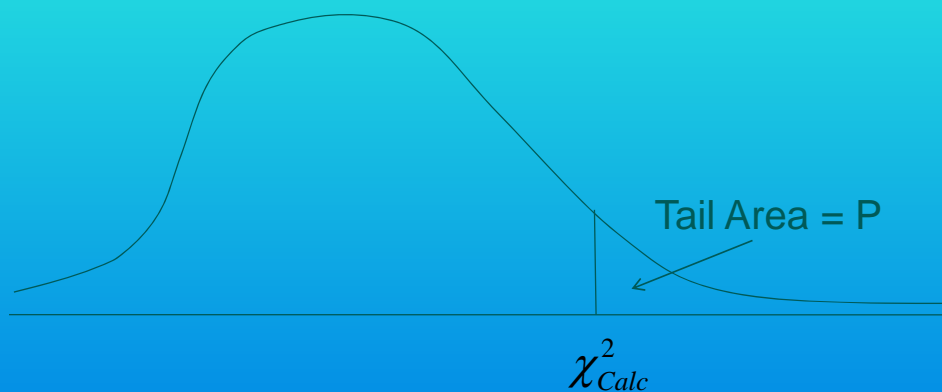
Expected Table Under Independence

	B	Not B	Total
A	$NP(A)P(B)$	$NP(A)(1-P(B))$	$NP(A)$
Not A	$N(1-P(A))P(B)$	$N(1-P(A))(1-P(B))$	$NP(A^c)$
Total	$NP(B)$	$NP(B^c)$	$N$

$$\chi_{Calc}^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$O_{ij}$  = Observed frequency,  $E_{ij}$  = Expected frequency

## Are A and B independent?



$P < .05 \Rightarrow A, B$  are dependent at 95% confidence

## Example 3: Lift and Chi-square

Observed Table

Drink Gatorade	Play Tennis		Total
	Yes	No	
Yes	5000	2500	7500
No	5000	2500	7500
Total	10000	5000	15000

Expected Table Under Independence

Drink Gatorade	Play Tennis		Total
	Yes	No	
Yes	5000	2500	7500
No	5000	2500	7500
Total	10000	5000	15000

	Support	Confidence	Lift
G ⇒ T	0.3333	0.6667	1.0000
No G ⇒ T	0.3333	0.6667	1.0000
G ⇒ No T	0.1667	0.3333	1.0000
No G ⇒ No T	0.1667	0.3333	1.0000

	Support	Confidence	Lift
T ⇒ G	0.3333	0.5000	1.0000
No T ⇒ G	0.3333	0.5000	1.0000
T ⇒ No G	0.1667	0.5000	1.0000
No T ⇒ No G	0.1667	0.5000	1.0000

$$\chi^2_{Calc} = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 0 \Rightarrow A \text{ and } B \text{ independent } (P=1)$$

$O_{ij}$  = Observed frequency,  $E_{ij}$  = Expected frequency

## Example 3: Lift and Chi-square

Observed Table

Drink Gatorade	Play Tennis		Total
	Yes	No	
Yes	7400	100	7500
No	2600	4900	7500
Total	10000	5000	15000

Expected Table Under Independence

Drink Gatorade	Play Tennis		Total
	Yes	No	
Yes	5000	2500	7500
No	5000	2500	7500
Total	10000	5000	15000

	Support	Confidence	Lift
G ⇒ T	0.4933	0.9867	1.4800
No G ⇒ T	0.1733	0.3467	0.5200
G ⇒ No T	0.0067	0.0133	0.0400
No G ⇒ No T	0.3267	0.6533	1.9600

	Support	Confidence	Lift
T ⇒ G	0.4933	0.7400	1.4800
No T ⇒ G	0.1733	0.2600	0.5200
T ⇒ No G	0.0067	0.0200	0.0400
No T ⇒ No G	0.3267	0.9800	1.9600

$$\chi^2_{Calc} = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 6912 \Rightarrow A \text{ and } B \text{ dependent } (P=0.000)$$

$O_{ij}$  = Observed frequency,  $E_{ij}$  = Expected frequency

## Mosaic Plots of Observed and Expected Frequencies for Example 3

### TENNIS

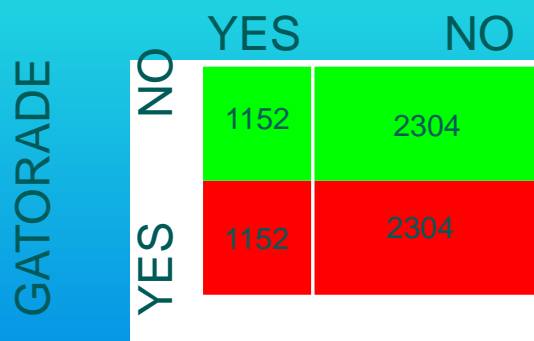


STATS24x7.com© 2010 ADI-NV, INC

39

## Contribution of each cell to $\chi^2$ for Example 3

### TENNIS



$$\chi^2_{\text{CALC}} = 6912, \quad df = 1, \quad P = 0.000$$

Tennis and Gatorade are **DEPENDENT**

STATS24x7.com© 2010 ADI-NV, INC

40

## Odds and Odds ratios for Example 3

Drink Gatorade	Play Tennis		Total	EVENTS	ODDS
	Yes	No			
Yes	7400	100	7500	T=Y vs T = N G=Y	74:1
No	2600	4900	7500	T=Y vs T = N G=N	26:49
Total	10000	5000	15000	T=Y vs T = N	2:1

ODDS also indicate strong association between TENNIS and GATORADE.

## Example 4: Preference for location of training camp for WWII recruits<sup>1</sup>

RACE	Region of Origin	Location of Present Camp	# of Soldiers Preferring Camp in		
			N	S	TOTAL
BLACK	N	N	387	36	423
BLACK	N	S	876	250	1126
BLACK	S	N	383	270	653
BLACK	S	S	381	1712	2093
WHITE	N	N	955	162	1117
WHITE	N	S	874	510	1384
WHITE	S	N	104	176	280
WHITE	S	S	91	869	960
TOTAL			4051	3985	8036

<sup>1</sup> Discrete Multivariate Analysis: Theory and Practice By Yvonne M. M. Bishop, Stephen E Fienberg, SpringerLink (page 138)

## Interesting Rules for Example 4

RACE	ANTECEDENT		CONSEQUENT	SUPPORT	CONFIDENCE	LIFT
	Region of Origin	Location of Present Camp	Camp Preference			
BLACK	N	N	PREFER N	0.0482	0.9149	1.8149
BLACK	N	S	PREFER N	0.1090	0.7780	1.5433
BLACK	S	N	PREFER N	0.0477	0.5865	1.1635
BLACK	S	S	PREFER N	0.0474	0.1820	0.3611
WHITE	N	N	PREFER N	0.1188	0.8550	1.6960
WHITE	N	S	PREFER N	0.1088	0.6315	1.2527
WHITE	S	N	PREFER N	0.0129	0.3714	0.7368
WHITE	S	S	PREFER N	0.0113	0.0948	0.1880

STATS24x7.com© 2010 ADI-NV, INC

43

## Interesting Rules for Example 4

RACE	ANTECEDENT		CONSEQUENT	SUPPORT	CONFIDENCE	LIFT
	Region of Origin	Location of Present Camp	Camp Preference			
BLACK	N	N	PREFER S	0.0045	0.0851	0.1716
BLACK	N	S	PREFER S	0.0311	0.2220	0.4477
BLACK	S	N	PREFER S	0.0336	0.4135	0.8338
BLACK	S	S	PREFER S	0.2130	0.8180	1.6495
WHITE	N	N	PREFER S	0.0202	0.1450	0.2925
WHITE	N	S	PREFER S	0.0635	0.3685	0.7431
WHITE	S	N	PREFER S	0.0219	0.6286	1.2676
WHITE	S	S	PREFER S	0.1081	0.9052	1.8254

STATS24x7.com© 2010 ADI-NV, INC

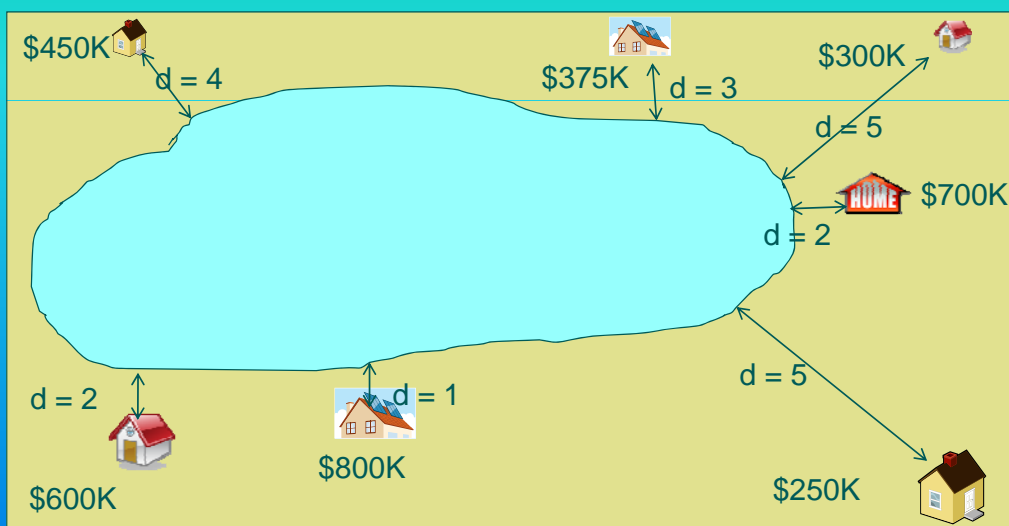
44

# SPATIAL ASSOCIATION RULES

- In many data-mining applications, data is spatial (related to a location on earth's surface) or spatio-temporal – e.g., geomarketing, crime hot-spot analysis, slot placement on casino floor.
- A Spatial Association Rule (sAR) is an AR involving a spatial variable\*.

P. Laube, M. de Berg, M. van Kreveld (2008). Headway in Spatial Data Handling (Eds. Anne Ruas, Christopher Gold), Lecture Notes in Geoinformation and Cartography Series, pp. 575 – 593.

## Example 5: A Spatial Association Rule



sAR: home is close to the lake  $\Rightarrow$  home is expensive.

## SUPPORT/CONFIDENCE USING BOOLEAN ALGEBRA

Define CLOSE/EXPENSIVE:

CLOSE =  $d \leq 2$ , EXPENSIVE : Value  $\geq 400$

d	Value	Close	Expensive
4	450	0	1
3	375	0	0
5	300	0	0
2	700	1	1
5	250	0	0
1	800	1	1
2	600	1	1

STATS24x7.com© 2010 ADI-NV, INC

47

## SUPPORT/CONFIDENCE OF sAR USING BOOLEAN ALGEBRA

$P(\text{Close}) = 3/7$ ,  $P(\text{Expensive}) = 4/7$ ,

$P(\text{Close} \cap \text{Expensive}) = 3/7$

Support( $\text{Close} \Rightarrow \text{Expensive}$ ) =  $3/7 = 0.43$

Confidence( $\text{Close} \Rightarrow \text{Expensive}$ ) =  $(3/7)/(3/7) = 1$

Lift( $\text{Close} \Rightarrow \text{Expensive}$ ) =

$P(\text{Close} \cap \text{Expensive}) / [P(\text{Close}) \times P(\text{Expensive})] =$

$(3/7) / [(3/7) \times (4/7)] = 1.75$

STATS24x7.com© 2010 ADI-NV, INC

48

## SUPPORT/CONFIDENCE OF sAR USING FUZZY LOGIC\*

Since the number of possible distances from the lakeshore is unlimited, a threshold and a cut-off point are needed to obtain meaningful support and confidence values for an sAR.

\* Dubois, Didier, Hüllermeir, Eyke and Prade, Henri (2006). A systematic approach to the assessment of fuzzy association rules. Data Mining and Knowledge Discovery. Vol. 13(2), pp. 167 – 192.

STATS24x7.com© 2010 ADI-NV, INC

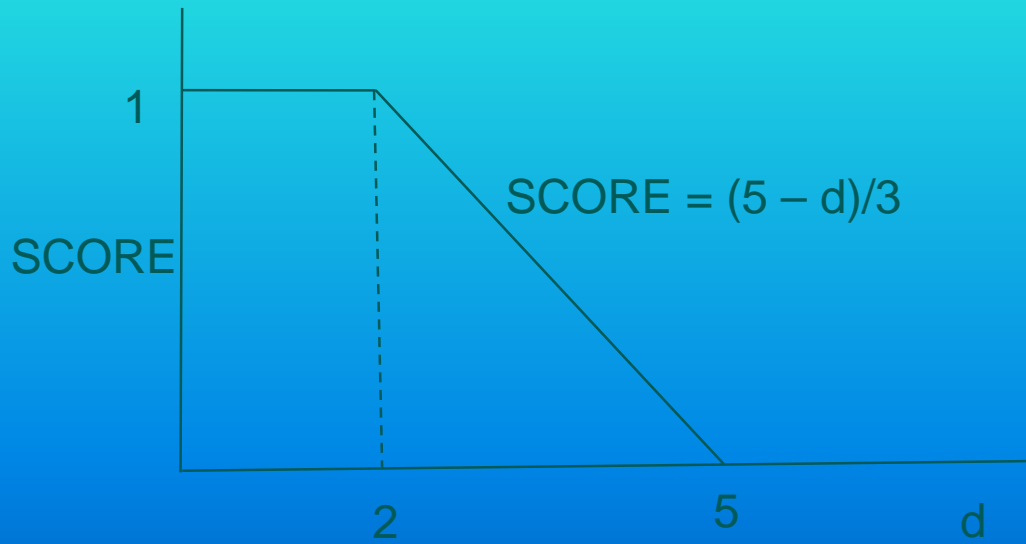
## SUPPORT/CONFIDENCE OF sAR USING FUZZY LOGIC\*

Convert each continuous variable (e.g., distance  $d$  or value  $V$ ) into a score in the range  $[0,1]$ , and calculate SUPPORT and CONFIDENCE of sAR  $A \Rightarrow C$  from:

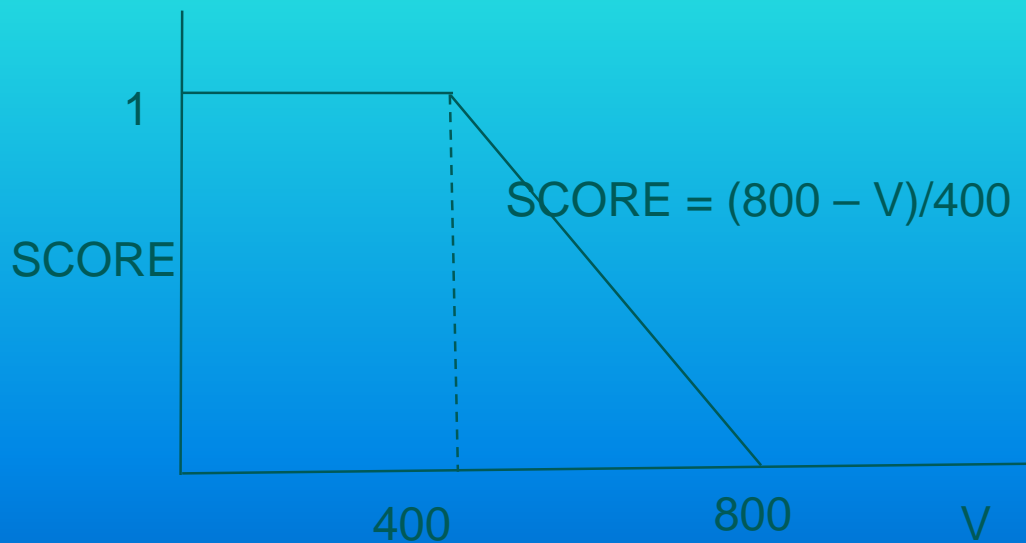
$$\text{Spatial support}(A \Rightarrow C) = \frac{\sum_x \text{score}_A(x) \times \text{score}_C(x)}{n}$$

$$\text{Spatial confidence}(A \Rightarrow C) = \frac{\sum_x \text{score}_A(x) \times \text{score}_C(x)}{\sum_x \text{score}_A(x)}$$

## SPATIAL VARIABLE d TO SCORE CONVERSION



## VALUE V TO SCORE CONVERSION



# CALCULATING SUPPORT AND CONFIDENCE OF sAR

d	Value	d_score	V_score	Support
4	450	0.33	0.875	0.29
3	375	0.67	1	0.67
5	300	0	1	0
2	700	1	0.25	0.25
5	250	0	1	0
1	800	1	0	0
2	600	1	0.5	0.5
<b>TOTAL</b>		<b>4</b>	<b>4.63</b>	<b>1.71</b>

CONFIDENCE =  $1.71/4 = 0.43$

## Example 6: sAR IN GAMING (Lewin, Singh, Cardno, Feb 2009, CEM)

Coin-in per day online for 28 slot games on a casino floor

distance from entrance	Coinin	Dist_score	Coinin_score	Sp_support	distance from buffet	Dist_score	Sp_support
6	74327.7	1	0.83	0.83	1	0.5	0.41
6	73494.7	1	0.72	0.72	1	0.5	0.36
6	75168.6	1	0.93	0.93	1	0.5	0.47
6	75722.9	1	1.00	1.00	1	0.5	0.50
6	73807.5	1	0.76	0.76	1	0.5	0.38
6	74755.7	1	0.88	0.88	1	0.5	0.44
7	72183.9	0.8	0.56	0.45	0	1	0.56
7	72620.8	0.8	0.61	0.49	0	1	0.61
7	72325.5	0.8	0.58	0.46	0	1	0.58
7	72341	0.8	0.58	0.46	0	1	0.58
7	72612.3	0.8	0.61	0.49	0	1	0.61
7	70419.1	0.8	0.34	0.27	0	1	0.34
7	74664.6	0.8	0.87	0.69	0	1	0.87
7	71671.5	0.8	0.49	0.40	0	1	0.49
7	75574.7	0.8	0.98	0.79	0	1	0.98
7	74075.6	0.8	0.79	0.64	0	1	0.79
7	72602	0.8	0.61	0.49	0	1	0.61
7	72348.5	0.8	0.58	0.46	0	1	0.58
6	73694	1	0.75	0.75	1	0.5	0.37
6	73575.6	1	0.73	0.73	1	0.5	0.37
6	74207.6	1	0.81	0.81	1	0.5	0.41
6	73868.6	1	0.77	0.77	1	0.5	0.38
11	68988.7	0	0.16	0.00	2	0	0.00
11	68462.9	0	0.09	0.00	2	0	0.00
11	69014.4	0	0.16	0.00	2	0	0.00
11	68411	0	0.09	0.00	2	0	0.00
11	67705	0	0.00	0.00	2	0	0.00
11	69001.4	0	0.16	0.00	2	0	0.00

## Two sAR's for Example 6

sAR1: slot machine is close to the casino entrance  $\Rightarrow$  high coin-in

sAR2: slot machine is close to the casino buffet  $\Rightarrow$  high coin-in

	Spatial Support	Spatial Confidence
SAR 1	0.51	0.73
SAR 2	0.42	0.69

## COMPUTING SCORE\* FOR ANTECEDENT "A<sub>1</sub> AND A<sub>2</sub>"

sAR: Home close to lake (A<sub>1</sub>) AND close to club house (A<sub>2</sub>) , then home is expensive

$$s_1 = \text{score}(A_1)$$

$$s_2 = \text{score}(A_2)$$

Then

$$\text{Score}(A_1 \text{ AND } A_2) = s_1 \times s_2$$

\* P. Laube, M. de Berg, M. van Kreveld (2008). Headway in Spatial Data Handling (Eds. Anne Ruas, Christopher Gold) Lecture Notes in Geoinformation and Cartography Series, pp. 575 – 593.

# COMPUTING SCORE\* FOR ANTECEDENT "A<sub>1</sub> OR A<sub>2</sub>"

sAR: Home close to lake (A<sub>1</sub>) OR close to club house (A<sub>2</sub>) , then home is expensive

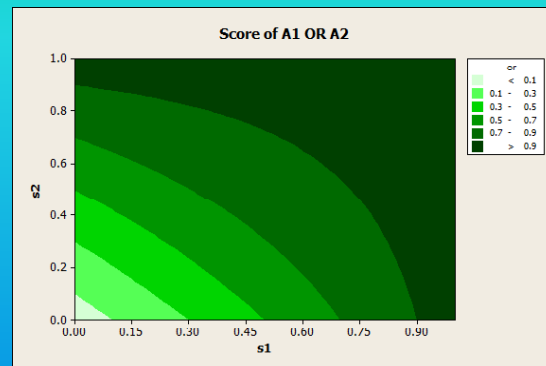
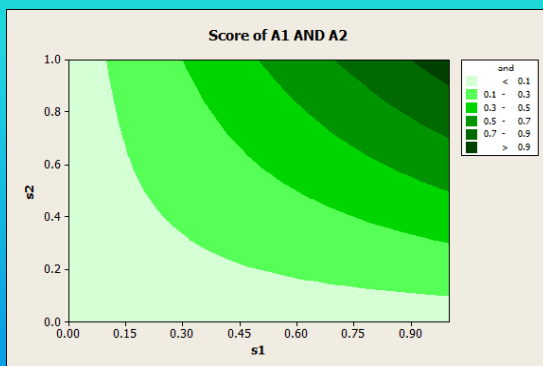
$$s_1 = \text{score}(A_1)$$

$$s_2 = \text{score}(A_2)$$

$$\text{Then } \text{Score}(A_1 \text{ OR } A_2) = \frac{s_1 + s_2}{1 + s_1 s_2}$$

\* P. Laube, M. de Berg, M. van Kreveld (2008). Headway in Spatial Data Handling (Eds. Anne Ruas, Christopher Gold), Lecture Notes in Geoinformation and Cartography Series, pp. 575 – 593.

# PLOT OF SCORES FOR COMPOSITE ANTECEDENTS



$$\text{Score}(A_1 \text{ AND } A_2) = \begin{cases} \text{low, both scores low} \\ \text{high, both scores high} \end{cases}$$

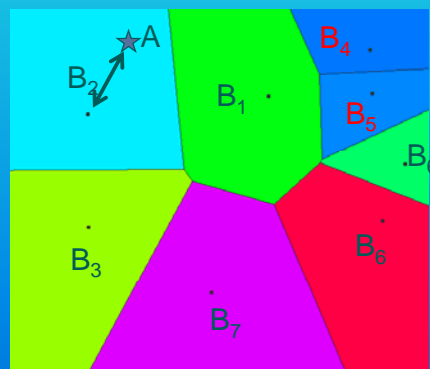
$$\text{Score}(A_1 \text{ OR } A_2) = \begin{cases} \text{low, either score low} \\ \text{high, either score high} \end{cases}$$

## DISTANCE MEASURES FOR SPATIAL ASSOCIATION RULES

In this section, we briefly discuss the notion of distance between various types of geographic objects.

## NEAREST NEIGHBOR DISTANCE

Crime location A close to a bank  $\Rightarrow$  crime involves money



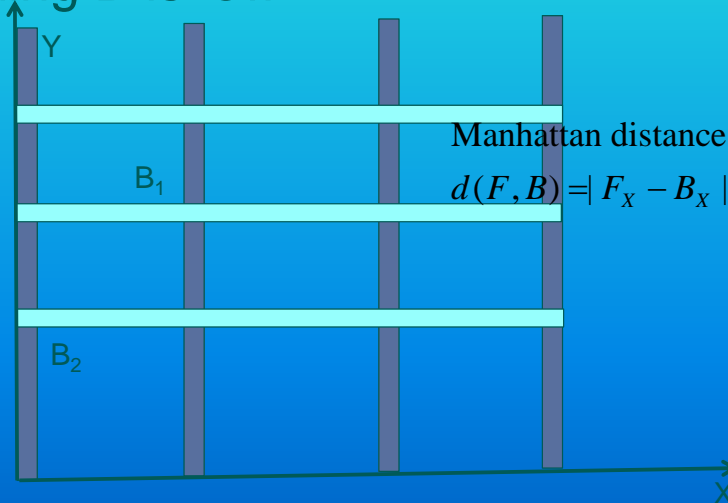
Voronoi diagram is used to define distance of A to closest bank (B<sub>j</sub>)

# MANHATTAN DISTANCE

Building B close to a fire-station  $\Rightarrow$  risk to building B is low



F



Manhattan distance between  $F$  and  $B$ :

$$d(F, B) = |F_x - B_x| + |F_y - B_y|$$