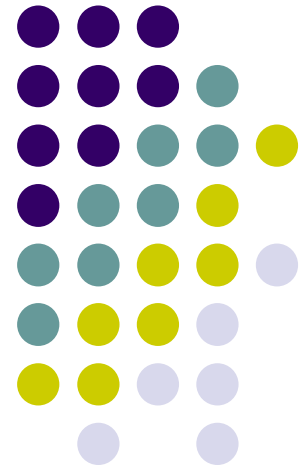


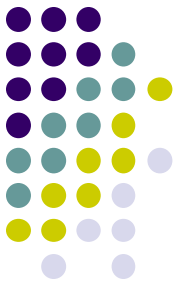
# Data Mining

## Association Rules

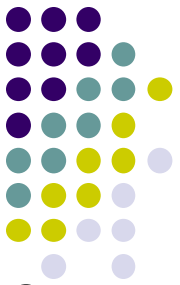
Prof.Dr. Songül Varlı  
Yıldız Technical University  
Computer Engineering Department  
[svarli@yildiz.edu.tr](mailto:svarli@yildiz.edu.tr)



# Association Rules



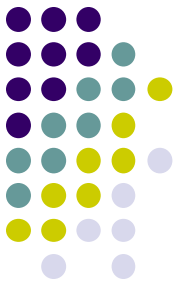
- Association rules are one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems



- Many business enterprises accumulate large quantities of data from their day-to-day operations
- For example, huge amounts of customer purchase data are collected daily at the checkout counters of grocery stores.

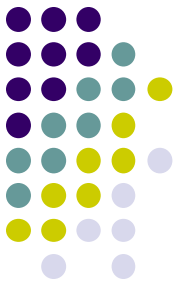
TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Egg}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

An example of market basket transaction



- It is a form of data mining that most closely resembles the process that most people think about when they try to understand the data mining process; namely, “mining” for gold through a vast database.
- The gold in this case would be a rule that is interesting, that tells you something about your database that you didn’t already know and probably weren’t explicitly articulate.

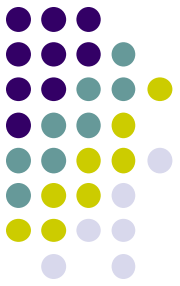




# Market Basket Analysis

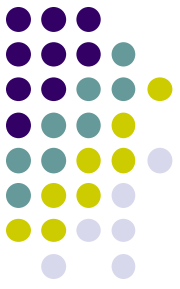
- A **market basket** is a collection of items purchased by a customer in a single transaction, which is a well defined business activity.
- One common analysis run against a transactions database is to find **sets of items**, or **itemsets**, that appear together in many transactions.

# Market Basket Analysis continuing

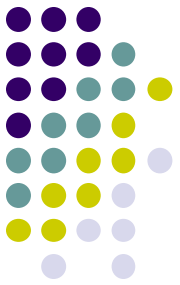


- A business can use knowledge of these patterns **to improve the placement** of these items in the store or the layout of mail-order catalog pages and web pages.

# Basic Concepts: Frequent Patterns



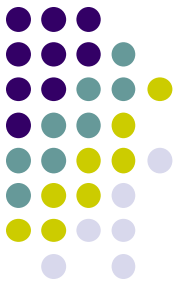
- An itemset containing  $i$  items is called an  $i$ -itemset.
- The percentage of transactions that contain an itemset is called the itemset's support.
- For an itemset to be interesting, its support must be higher than a user-specified minimum. Such itemsets are said to be frequent.



# Basic Concepts: Frequent Patterns

- Let  $I = \{i_1, i_2, \dots, i_m\}$  be a set of literals, called **items**.
- Let DB be a set of transactions, where **each transaction T is a set of items** such that  $T \subseteq I$ .
- Note that the **quantities** of the items bought in a transaction are **not considered**, meaning that each item is a binary variable indicating whether an item was bought or not.

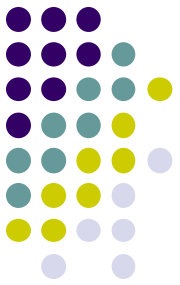




# Basic Concepts: Frequent Patterns

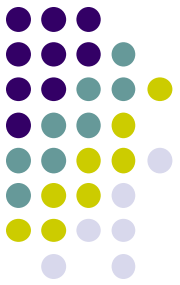
- An example of the model for such a transaction database is given as follow;

Transaction-id	Items bought
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E



# Basic Concepts: Frequent Patterns

- Let  $X$  be a set of items. A transaction  $T$  is said to contain  $X$  if and only if  $X \subseteq T$
- An **association rule** implies the form  $X \Rightarrow Y$ , where  $X \subseteq I$ ,  $Y \subseteq I$  and  $X \cap Y = \emptyset$ .
- The rule  $X \Rightarrow Y$  holds in the transaction set DB with **confidence  $c$**  if  $c\%$  of the transaction in  $D$  that contain  $X$  also contain  $Y$ .

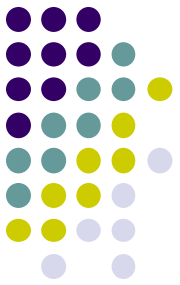


# Basic Concepts: Frequent Patterns

- The rule  $X \Rightarrow Y$  holds in the transaction set DB with **confidence c** if  $c\%$  of the transaction in D that contain X also contain Y.
- The rule  $X \Rightarrow Y$  has **support s** in the transaction set D if  $s\%$  of the transaction in DB that contain  $X \cup Y$ .

$$\text{Support } (X \Rightarrow Y) = P(X \cup Y)$$

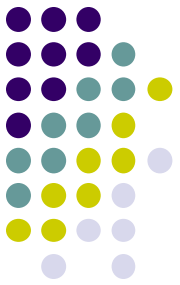
$$\text{Confidence}(X \Rightarrow Y) = P(Y|X)$$



Transaction ID	Items Bought
1	{Shoes, Trousers, Shirt, Belt}
2	{Shoes, Trousers, Shirt, Hat, Belt, Scarf}
3	{Shoes, Shirt}
4	{Shoes, Trousers, Belt}

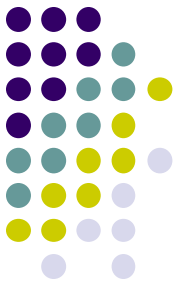
Transaction ID	Shoes	Trousers	Shirt	Belt	Hat	Scarf
1	1	1	1	1	0	0
2	1	1	1	1	1	1
3	1	0	1	0	0	0
4	1	1	0	1	0	0

Support(Trousers  $\Rightarrow$  Shirt)=2/4  
Confidence(Trousers  $\Rightarrow$  Shirt)=2/3



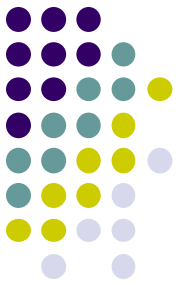
# Basic Concepts: Strong Rules

- **Confidence** denote **strength of implication** and **support** indicates the **frequency of the patterns** occurring in the rule.
- It is often desirable to pay attention to only those rules that may have a reasonable large support.
- Such rules with **high confidence** and **strong support** are referred to as **strong rules**.
- The task of **mining association rules** is essentially to **discover strong association rules** in large databases.



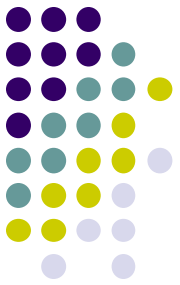
# Mining Association Rules

- The problem of mining association rules may be decomposed into **two phases**:
  - **Discover the large itemsets**, i.e. the sets of items that have transaction **supports above predetermined minimum threshold**
  - **Use the large itemsets** to generate the association rules for the database that have **confidence  $c$  above a predetermined mining threshold**



# ALGORITHM APRIORI

- The algorithm Apriori computes the frequent itemsets in the database through several iterations.
- Iteration  $i$  computes all frequent  $i$ -itemsets (itemsets with  $i$  elements)
- Each iteration has two steps:
  - Candidate generation
  - Candidate counting and selection



# ALGORITHM APRIORI

- In the **first phase** of the first iteration, the **generated set of candidate itemsets contains all 1-itemsets** (i.e. All items in the database)
- In the **counting phase**, the algorithm **counts their support** searching again through the whole database. Finally, only 1-itemsets (items) with  $s$  above required threshold will be selected as frequent.
- Thus, **after the first iteration, all frequent 1-itemsets will be known.**



# The Apriori Algorithm—An Example



$\text{Sup}_{\min} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1<sup>st</sup> scan

$C_1$

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

$L_1$

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

$C_2$

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

2<sup>nd</sup> scan

$C_2$

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

$L_2$

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

$C_3$

Itemset
{B, C, E}

3<sup>rd</sup> scan

$L_3$

Itemset	sup
{B, C, E}	2

# An itemset lattice diagram

