

Data Mining:

Association and Correlation

Lecturer: Mohammed L. Mahmood



Outline

- Correlation
- Pearson Correlation Coefficient
- Association
- Apriori Algorithm

Example of Correlation

Is there an association between:

- Children's IQ and Parents' IQ
- Degree of social trust and number of membership in voluntary association ?
- Urban growth and air quality violations?
- Number of police patrol and number of crime
- Grade on exam and time on exam

Correlation

- Correlation is the degree of inter-relatedness among the two or more variables. Correlation analysis is a process to find out the degree of relationship between two or more variables by applying various statistical tools and techniques.
- It is used in deriving the degree and direction of relationship within the variables.
- It is used in reducing the range of uncertainty in matter of prediction.
- It is used in presenting the average relationship between any two variables through a single value of coefficient of correlation.

Correlation coefficient

- Correlation coefficient: statistical index of the degree to which two variables are associated, or related.
- We can determine whether one variable is related to another by seeing whether scores on the two variables *covary* (*change together*)---whether they vary (change) together.

Types of correlation

```
graph TD; A[Types of correlation] --> B[On the basis of degree of correlation]; A --> C[On the basis of number of variables]; A --> D[On the basis of linearity]; B --> B1[•Positive correlation]; B --> B2[•Negative correlation]; C --> C1[•Simple correlation]; C --> C2[•Partial correlation]; C --> C3[•Multiple correlation]; D --> D1[•Linear correlation]; D --> D2[•Non – linear correlation];
```

On the basis of
degree of
correlation

- Positive correlation
- Negative correlation

On the basis of
number of variables

- Simple correlation
- Partial correlation
- Multiple correlation

On the basis of
linearity

- Linear correlation
- Non – linear correlation

Degree of Correlation

- Positive Correlation: It happens when one variable is increasing and with its impact on average other variable is also increasing.
- For example :
 - Income (Rs.) : 350, 360, 370, 380
 - Weight (Kg.) : 30, 40, 50, 60
- Negative correlation: It happens when a variable is increasing, with its impact on average, other variable will be decreasing.
- For example :
 - Income (Rs.) : 350, 360, 370, 380
 - Weight (Kg.) : 80, 70, 60, 50

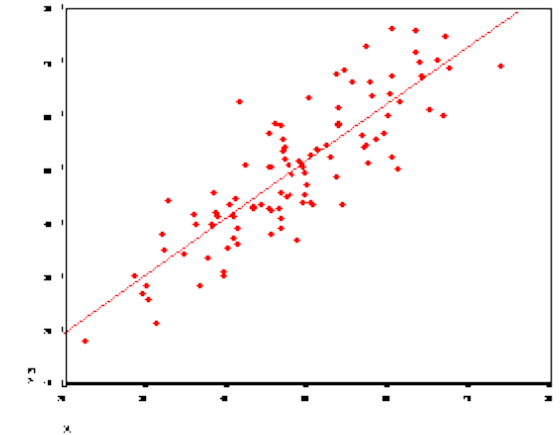
Correlation Based on Number of Variables

- Simple correlation: Correlation is simple when only two variables are used in the process of the analysis.
- Partial correlation : When three or more variables are considered for analysis but only two influencing variables are studied and rest influencing variables are kept constant.
- In case of multiple correlation three or more variables are studied simultaneously.

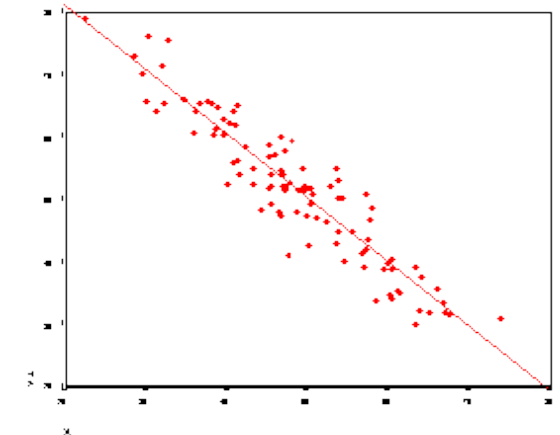
Correlation Based on Linearity

$r = .85$

- Linear correlation : If the change in amount of one variable tends to make changes in amount of other variable bearing constant changing ratio it is said to be linear correlation.
- Non - Linear correlation : If the change in amount of one variable tends to make changes in amount of other variable but not bearing constant changing ratio it is said to be non - linear correlation.



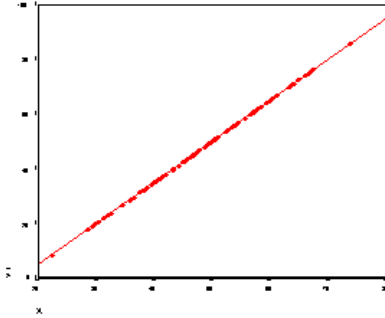
$r = -.94$



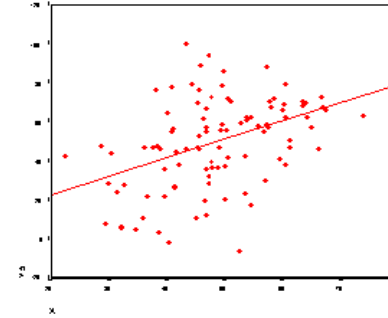
$r = .42$

Strong and weak correlation examples

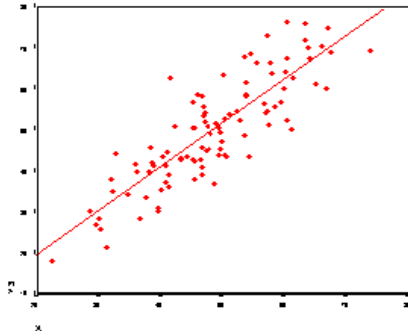
$r = 1.00$



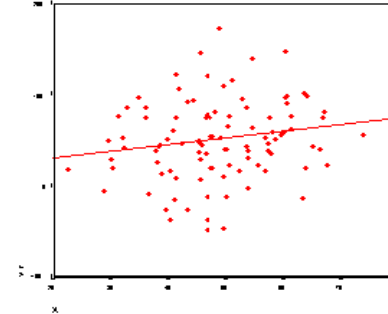
$r = .42$



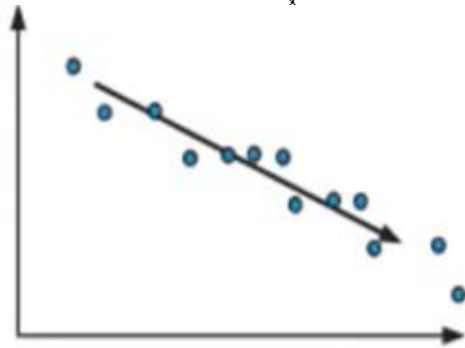
$r = .85$



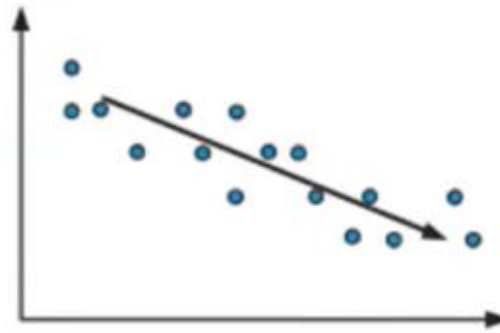
$r = .17$



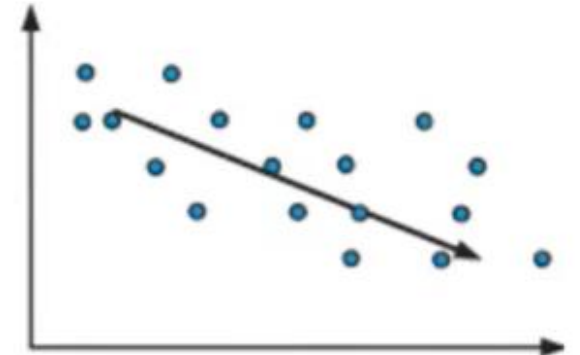
strong



moderate



weak



Pearson's correlation coefficient

- There are many kinds of correlation coefficients but the most commonly used measure of correlation is the Pearson's correlation coefficient. (r)
 - In order to test the linear association between two variables x and y we can use the Pearson correlation coefficient
- The Pearson r range between -1 to +1.
 - Sign indicate the direction.
 - The numerical value indicates the strength.
 - Perfect correlation : -1 or 1
 - No correlation: 0
 - A correlation of zero indicates the value are not linearly related.
 - However, it is possible they are related in **curvilinear** fashion.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Association

- Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.
- Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction.

What Is Frequent Pattern Analysis in Association?























- **Frequent pattern**: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- Motivation:
 - What products were often purchased together?— Pepsi and chips?!
 - What are the subsequent purchases after buying a PC?
 - What kinds of DNA are sensitive to this new drug?
 - Can we automatically classify web documents?
- Applications
 - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Apriori Algorithm

- **Apriori algorithm** is given by R. Agrawal and R. Srikant in 1994 for finding frequent itemsets in a dataset for boolean association rule.
- Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties.
- We apply an iterative approach or level-wise search where k -frequent itemsets are used to find $k+1$ itemsets.

Apriori Algorithm

- Assume the following figure as an example:

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

Apriori

- Support : how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. In the previous slides table, the support of {apple} is 4 out of 8 (4/8), or 50%. Itemsets can also contain multiple items. For instance, the support of {apple, Pepsi, rice} is 2 out of 8 (2/8), or 25%.
- Confidence: how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears. In the previous table, the confidence of {apple -> Pepsi} is 3 out of 4 (3/4), or 75%.

$$\text{Confidence } \{\text{🍎} \rightarrow \text{🍺}\} = \frac{\text{Support } \{\text{🍎}, \text{🍺}\}}{\text{Support } \{\text{🍎}\}}$$

- Lift: How likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. In the previous table, the lift of {apple - > Pepsi} is 1, which implies no association between items. A lift value greater than 1 means that item Y is likely to be bought if item X is bought, while a value less than 1 means that item Y is unlikely to be bought if item X is bought.

$$\text{Lift} \{ \text{🍎} \rightarrow \text{🍺} \} = \frac{\text{Support} \{ \text{🍎}, \text{🍺} \}}{\text{Support} \{ \text{🍎} \} \times \text{Support} \{ \text{🍺} \}}$$

The Apriori Algorithm—An Example

