Association Rule

* Association rule learning can be divided into 3 kinds of algorithm
  * Apriori
  * Eclat
  * T-P Growth

* Association rule learning works on the concept of if and else statement

  eg: If A then B.

  where if A is an incident

  then B is consequent.

Let us discuss about Apriori algorithm

* A priori algorithm is generally considered as unsupervised learning algorithm

~~problem~~ formula's

~~For~~

$$\text{support} = \frac{\text{No of items x \& Y bought together}}{\text{Total no of transcations}}$$

$$\boxed{\text{Support} = \frac{fre(x,y)}{N.}}$$

$$\text{confidence} = \frac{\text{Total no. of items x \& Y bought together}}{\text{Total no. of items 'x' bought}}$$

$$\text{confidence} = \frac{\text{Freq}(x,y)}{\text{Freq}(x)}.$$

$$\text{lift} = \frac{\text{support}(x,y)}{\text{support}(x) + \text{support}(y)}.$$

Problem

Construct the following transcations apply the association rule mining to get the association rule with minimum support two and confidence of 50%.

⇒ Association rule mining is explained using Apriori Algorithm

Transcational data for All electronics branch.

T_{ID}	List of item IDs
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

Step 1 :

Now scan each itemset and count how many candidate each has.

1-itemset	supcount
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

Step 2

Check whether the candidate support count is with minimum of "two".

1-itemset	supcount
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

step3

Now Generate two candidate keys.

2-itemset

{I1,I2}
{I1,I3}
{I1,I4}
{I1,I5}
{I2,I3}
{I2,I4}
{I2,I5}
{I3,I4}
{I3,I5}
{I4,I5}.

step4:

Now Scan the support count of candidate keys.

2-itemset	Supcount.
{I1,I2}	4
{I1,I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

5:

Check whether the candidate support count is with minimum of 'two', if not remove the candidate key. (from step 4)

2-itemset	supcount
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

Step 6:

Now Generate three candidate keys. with sup count

3 itemset	supcount
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2
$\{I_1, I_2, I_4\}$	1
$\{I_1, I_3, I_4\}$	0
I_1, I_3, I_5	1
I_1, I_4, I_5	0
I_2, I_3, I_4	0
I_2, I_3, I_5	1
I_2, I_4, I_5	0

Step 7:

Check whether the candidate support count is with minimum of 'two'

3-itemset	supcount
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2

Step 8:

Generate four candidate keys.

itemset	supCount
$\{I_1\ I_2\ I_3 I_5\}$	1

So not possible, because it has only 1 supCount

* Non-empty subsets of frequency sets are three item sets

$\{I_1\ I_2\ I_3\}$ $\{I_1\ I_2\ I_5\}$.

→ $\{\{I_1),(I_2),(I_3),(I_1,I_2),(I_1,I_3),(I_2,I_3)\}$

→ $\{(I_1),(I_2),(I_5),(I_1,I_2),(I_1,I_5),(I_2,I_5)\}$.

* Association rule for every non-empty subsets

$$S \Rightarrow (I - S)$$

* Let us consider, Association rule mining subset creation between three Items sets.

$I \Rightarrow \{I_1\ I_2\ I_3\}$.

Non-empty subsets.

$\{(I_1)\ (I_2)\ (I_3)\ (I_1\ I_2)\ (I_1, I_3)\ (I_2 I_5)\}$.

le 1 :

$\{I_1\} \rightarrow \{I_2, I_3\}$.

$\downarrow$ S $\downarrow$ I-S

support $= \frac{2}{9} = 22.22\%$

Confidence $=$ support $\frac{(1,2,3)}{\text{support}(1)} = \frac{2/9}{6/9} = 1/3 = 33.33\%$.

R ∴ The condition is invalid ($< 50\%$).

Rule 2

$\{2\} = \{1,3\}$

Support $= 2/9 = 22.22\%$.

Confidence $= \frac{\text{support } (1,2,3)}{\text{support } (2)} = \frac{2/9}{7/9} = 2/7 = 28.57\%$.

∴ The condition is invalid ($< 50\%$)

Rule 3

$\{3\} = \{1,2\}$

support $= 2/9 = 22.22\%$.

Confidence $= \frac{\text{support } (1,2,3)}{\text{support } (3)} = \frac{2/9}{6/9} = 2/6 = 1/3$ 33.34

∴ The condition is invalid ($< 50\%$

Rule 4

$\{1,2\} = \{3\}$

support $= 2/9 = 22.22\%$.

Confidence = $\frac{\text{support}(1,2,3)}{\text{support}(1,1)} = \frac{2/9}{4/9} = 2/4 = 50\%$

∴ The above condition is valid (>=50%).

le 5

$\{1,3\} = \{2,3\}$

support = $2/9 = 22.22\%$

confidence = $\frac{\text{support}(1,2,3)}{\text{support}(1,3)} = \frac{2/9}{4/9} = 2/4 = 50\%$

∴ The above condition is valid (>=50%)

le 6

$\{2,3\} = \{1,3\}$

support = $2/9 = 22.22\%$

confidence = $\frac{\text{support}(1,2,3)}{\text{support}(2,3)} = \frac{2/9}{7/9} = 2/7 = 28.57$

∴ The above condition is invalid (< 50%)

6/2/23

UNIT-II

CLASSIFICATION:

* Classification is a process of categorizing a given set of data into classes.
* The classes are often referred to as target, label or categories.
* Classification is defined as the process of recognition, understanding and grouping of objects and ideas to proceed categories.

In classification algorithms are used to predict the output of the categorical data.

CLASSIFIER:

The algorithm which implement the classification on a dataset is known as "classifier."

NAIVE BAYESIAN CLASSIFICATION:

* Invented by Reverend Thomas Bayes in 1761.
* Naive Bayes assumes that all attributes are equally important, independent of one another in a given class.

BAYES THEOREM:

Bayes Theorem is one of the most popular machine learning concept that helps to calculate the probability of occuring one event with uncertain knowledge while other are already occured.

$$P(class|data) = \frac{(P(data|class) * P(class))}{P(data)}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P(A|B) - The probability of 'A' being true given that 'B' is True.

P(A) - The probability of 'A' being true.

7/2/23

Dataset: Weather condition.

Problem: If the weather is sunny, then the player should play or not?

	outlook	play
0	Rainy	yes
1	Sunny	yes
2	overcast	yes
3	overcast	yes
4	sunny	yes
5	Rainy	yes
6	sunny	yes
7	overcast	yes
8	Rainy	NO
9	sunny	NO
10	sunny	NO
11	Rainy	NO
12	overcast	yes
13	overcast	yes

FREQUENCY TABLE:

Weather	yes	NO
overcast	5	0
Rainy	2	2
sunny	3	2

Likelyhood Table:

weather	yes	NO	
overcast	5	0	$5/14 = 0.35$
Rainy	2	2	$4/14 = 0.29$
sunny	3	2	$5/14 = 0.3$

All 10/14 4/14 = 0.29

= 0.71

Applying Bayes Theorem: —

$$P(yes|sunny) = \frac{P(sunny|yes) * P(yes)}{P(sunny)}$$

P(sunny|yes) = 3/10 = 0.3

P(sunny) = 0.35

P(yes) = 0.71

so, $P(yes|sunny) = \frac{0.3 \times 0.71}{0.35} = 0.60$

$$P(NO|sunny) = \frac{P(sunny|NO) * P(NO)}{P(sunny)}$$

P(sunny|NO) = 2/4 = 0.5

P(NO) = 0.29

P(sunny) = 0.35

so $P(NO|sunny) = \frac{0.5 \times 0.29}{0.35} = 0.41$

so P(yes|sunny) > P(NO|sunny)

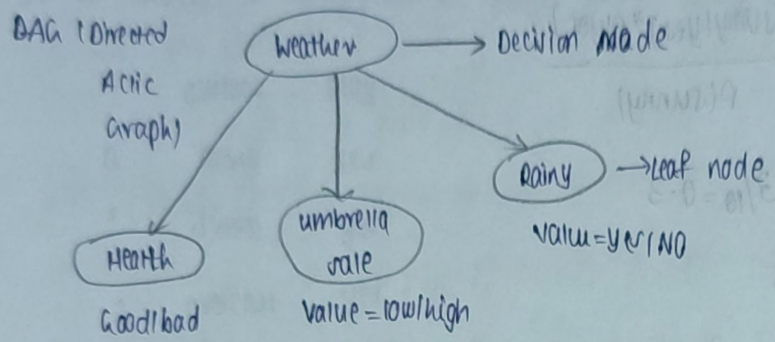
Hence on sunny days player can play the game.

13/2/23

BAYERIAN NETWORK:

* Bayerian networks are the type of probabilistic graphical model that can be used to build models from data.
* They are also commonly reffered as Bayes Network, Belief N/w.

Bayerian new are probabilistic because they are built from probability distribution.



- probability of Good Health based on weather.
- probability of umbrella sale based on weather.
- Random variables (weather, health) can give a hypothesis.

TYPES OF PROBABILITIES:

1. Join probability
2. conditional probability.

DECISION TREE:

* Decision Tree are some of the most used machine learning algorithm in supervised learning.
* Decision Trees are used in both classification & Regression.
* They can be used for both linear and non-linear data but they are mostly used in non-linear data.
* It is a tree structured classifier.

14/2/23

Predict the probability of a class variable using Bayes classification.

Training Dataset: All electronics customer database

Variable (X) = (age = youth, income = Medium, student = yes, credit = rating = fair)

RID	Age	Income	student	credit rating	buys-computer
1.	youth	high	NO	fair	NO
2	youth	high	no	Excellent	NO
3	Middle age	high	no	fair	yes
4.	senior	medium	no	fair	yes
5.	senior	low	yes	fair	yes
6.	senior	low	yes	Excellent	NO
7.	Middle age	low	yes	Excellent	yes
8.	youth	Medium	NO	fair	NO
9.	youth	Low	yes	fair	yes
10.	senior	Medium	yes	fair	yes
11.	youth	Medium	yes	Excellent	yes
12.	Medium age	Medium	NO	Excellent	yes
13.	MA	High	yes	fair	yes
14.	senior	Medium	NO	Excellent	NO

Class: c_1 : buys-computer = yes

C_2 : buys-computer = no

probability:

$$P(c_1) = p(C_1) = P(\text{buys\_computer} = yes) = \frac{9}{14} = 0.643$$

$$P(c_2) = P(\text{buys-computer} = NO) = \frac{5}{14} = 0.357$$

compute P(X|Ci) for each class:

P(age = youth | buys_computer = yes) = $\frac{2}{9}$ = 0.222

P(age = medium | buys_computer = yes) = $\frac{4}{9}$ = 0.444

P(student = yes | buys_computer = yes) = $\frac{6}{9}$ = 0.667

P(credit_rating = fair | buys_computer = yes) = $\frac{6}{9}$ = 0.667

my

P(age = youth | buys_computer = NO) = $\frac{3}{5}$ = 0.6

P(income = medium | buys_computer = NO) = $\frac{2}{5}$ = 0.4

P(student = yes | buys_computer = NO) = $\frac{1}{5}$ = 0.2

P(credit_rating = fair | buys_computer = NO) = $\frac{2}{5}$ = 0.4

SO P(X|Ci) : P(X | buys_computer = yes)

= 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(X | buys_computer = NO) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

P(X|Ci) x (P(Ci)

P(X | buys_computer) x P(buys_computer = yes)

= 0.044 x 0.643 = 0.028

P(X | buys_computer = NO) x P(buys_computer = NO)

= 0.029 x 0.357 = 0.007

RESULT:

P(X|buys_comp = yes) P(buys_comp = yes) > (greater than) P(X | buys_computer = yes)

P(buys_comp = no)

The variable X (age = youth, income = medium, student = yes, credit - ranking = fair), have the probability of buying a computer.

15/2/23

ASSOCIATION RULE:

* Association rule learning is a type of supervised learning in machine learning.
* It checks the dependency of one data item on another data item and maps accordingly.
* It is mostly used in questions, decision making and predict behaviours.
* It is a rule based machine learning method for discovering interesting relatives blw variables in large database.

ASSOCIATION RULE - APRIORI ALGORITHM:

* Apriori algorithm is used to find the frequent itemset in a dataset for boolean association rule.
* Apriori - means prior knowledge of frequent item set properties.
* Apriori is an algorithm used for association rule mining.
* Types of Association are multi-relational, quantative and generalized.

Apriori ALGORITHM MEASUREMENTS:

There are 2 measures mostly used in association rule.

1. support.

$$\text{support} = \frac{\text{frequency}(X,Y)}{N}$$

2. confidence

$$\text{confidence} = \frac{\text{frequency}(X,Y)}{\text{frequency}(X)}$$

Example: Apriori Algorithm.

The following are the set of items transacted in a market. Find the support and confidence?

Trans ID	
101	milk, bread, eggs
102	milk, juice
103	Juice, Butter.
104	milk, Bread, eggs
105	coffee, eggs
106	coffee
107	coffee, Juice
108	milk, bread, cookies, eggs
109	cookies, butter
110	milk, bread.

Index the Item sets

Item	Numbers
Milk	1
bread	2
Eggs	3
Juice	4
Butter	5
coffee	6
cookies	7

Find the overall support (Total item = 10 = 100%)

Itemset	support	%
Milk	5	50%,
Bread	4	40%,
Eggs	4	40%,
Juice	3	30%,
Butter	2	20%,
coffee	3	30%,
cookies	2	20%,

Find the support for 2 item set from Index.

Itemset	support	%
{1,2}	4	40%
{1,3}	3	30%
{1,4}	1	10%
{1,5}	0	-
{1,6}	0	-
{1,7}	1	10%
{2,3}	3	30%
{2,4}	0	-
{2,5}	0	-
{2,6}	0	-
{2,7}	1	10%
{3,4}	0	-
{3,5}	0	-
{3,6}	1	10%
{3,7}	1	10%
{4,5}	1	10%
{4,6}	1	10%
{4,7}	0	-
{5,6}	0	10%
{5,7}	1	10%
{6,7}	0	-

Itemset	support
{1,2}	4
{1,3}	3
{2,3}	3

Itemset	support
{1,2,3}	3

use 20% for the minimum support

CONFIDENCE: 70% or greater in the item set

$$\frac{\text{support}(1,2,3)}{\text{support}(1)} = \frac{3}{5} = 60\% < 70\% \quad \times$$

$$\frac{\text{support}(1,2,3)}{\text{support}(2)} = \frac{3}{4} = 75\% > 70\% \quad \checkmark$$

$$\frac{\text{support}(1,2,3)}{\text{support}(3)} = \frac{3}{4} = 75\% > 70\% \quad \checkmark$$

confidence R:

	Confidence
Bread → {milk, eggs}	75%
Eggs → {milk, bread}	75%

$$\frac{\text{support}(1,2,3)}{\text{support}(1,2)} = 75$$

$$\frac{\text{support}(1,2,3)}{S(1,3)} = 100\%$$

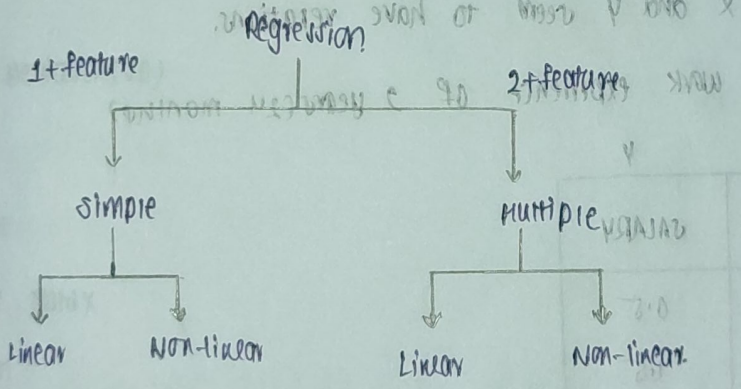
$$\frac{S(1,2,3)}{S(2,3)} = 100\%$$

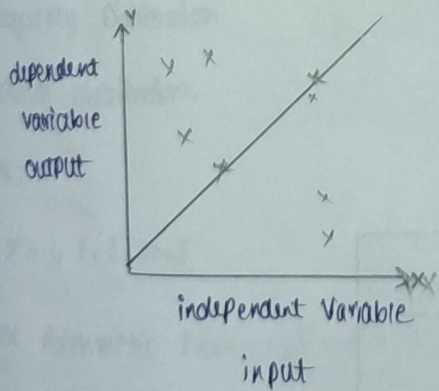
16/2/23

REGRESSION:

* Regression is a statistical method to determine the strength and character of the relationship b/w one dependent variable (usually denoted by Y) and a series of the other variables (known as dependent variable X).
* regression is a supervised Machine learning algorithm.
* Regression analysis is the process of estimating the relationship b/w a dependent variables and to independent variables.
* Regression will predict a continuous outcome (Y) based on the value of one or more predictor variables (X).

REGRESSION MODELS:





SIMPLE LINEAR REGRESSION:

simple linear Regression is a technique which plots a straight line within the data points. It is one of the most simple and basic type of machine learning Regression.

MULTIPLE LINE REGRESSION:

It is a technique used when more than one independent variable is used. polynomial is an example of a multilinear regression.

SIMPLE LINEAR REGRESSION:

* One variable (x) is called independent variable or predictor
* The other variable (y) is known as dependent variable or outcome, and the simple linear regression equation is

$$Y = b_0 + b_1 x$$

What is linear regression? The following table shows the work experience and salary obtained from employees

a) plot the data do x and y seem to have relations.

b) predict the salary for work experience of 2 years (24 months)

X WORK EXPERIENCE months	Y SALARY
3	0.5
6	1.5
9	1.5
12	4
15	6

X	Y	X^2	Y^2	XY
$\Sigma X = 45$	13.5	162	495	56.75

$$y = b_0 + b_1 x$$

$$b_0 = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n \cdot \Sigma x^2 - (\Sigma x)^2}$$

$$b_1 = \frac{n \cdot \Sigma xy - \Sigma x \Sigma y}{n \cdot \Sigma x^2 - (\Sigma x)^2}$$

$$y = -1.35 + 0.45 \times 24$$

$$= 9.45$$

20/9/23

STATISTICS:

1. Simple Arithmetic mean.
2. Median
3. Mode
4. Standard Deviation (SD)
5. variance
6. co-variance
7. co-variance Matrix
8. Distribution
9. Frequency Distribution
10. Normal Distribution.

MEAN:

$x = 1, 2, 3, 4, 5$

Simple Arithmetic Mean $= \frac{1+2+3+4+5}{5}$

$= \frac{15}{5} = 3$

MEDIAN :

50th percentile position

Position = N+1/2

V = 15, 10, 8, 12, 14

Arrange in highest to lowest 15, 14, 12, 10, 8

N=5 $\frac{N+1}{2} = \frac{6}{2} = 3^{rd}$ position

= 12

MODE: No. of occuring of data.

3, 5, 7, 10 - No mode

3, 5, 3, 7, 3, 10, 3 Mode = 3

POPULATION VARIANCE:

$$\sigma = \frac{(X-\bar{X})^2}{n}$$

(X − X̄)² → variable, X̄ → Mean, n → No. of score

SAMPLE VARIANCE:

$$\sigma = \frac{X-\bar{X}}{n-1}$$

mean x = 3

$$PV = (1-3) + (2-3) + (3-3) + (4-3) + (5-3)/5$$

$$= (-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2/5$$

$$= \frac{4+1+0+1+4}{5}$$

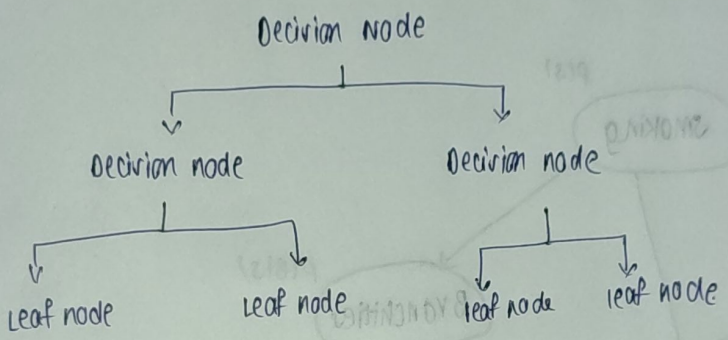
$$P.C.V = \frac{|X-\bar{X}|^2 + |Y-\bar{Y}|^2}{n}$$

Standard deviation: Square root of variance.

2/12/23

DECISION TREE:

* Decision Trees are some of the most used machine learning algorithm.
* Decision Tree is a supervised learning algorithm.
* Decision Tree can be used in both classification and regression techniques.
* Decision Tree can be used for both linear (continuous) and non-linear (discrete) data but mostly used in non-linear data.



INTERNAL NODE: Represent features of dataset.

BRANCHES: Represents the decision rule.

LEAF NODE: Represents the outcome.

BAYESIAN NETWORK:

Bayesian Nlw are a type of probabilistic graphical model that can be used to build models from data.

They are also commonly reffered to as Bayes Network, Belief Nlw, caried as Nlw model.

* Bayesian Nlw are probabilistic because they are built from probability distribution.

2 types of probability

1. Join probability
2. conditional

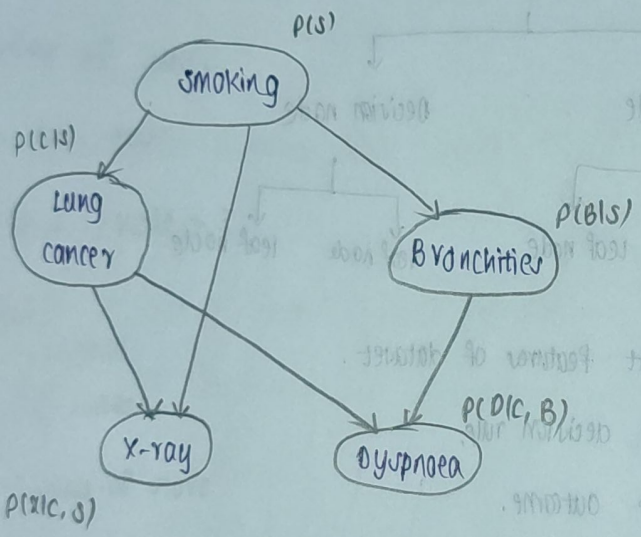
28/2/23

BAYESIAN NETWORK:

Structured Graphical representation of probabilistic relationship btw several random variables.

PROBABILITIES:

1. Joint probability
2. Conditional probability distribution (cpd)



BAYESIAN NETWORK (BN = G, θ)

G - Directed Acyclic Graph (DAG) nodes - random variables edges - direct dependencies

θ - set of parameters in all conditional probability Distribution (cpd)

cpd:-

C	B	D=0	D=1
0	0	0.1	0.9
0	1	0.7	0.3
1	0	0.8	0.2
1	1	0.9	0.1

0 - False, 1 - True

cpd of node X = P(X / parents(X))

X - random variable (smoking, lung cancer, Bronchities, X-ray, Dyspnoea)

compact representation of Joint Distribution in a product form

$P(S, C, B, X, D) = P(S)\, P(C/S)\, P(B/S)\, P(X/C,S)\, P(D/C,B)$

$= 1 + 2 + 2 + 4 + 4 = 13$ parameters instead of $2^5 = 32$