Friday
5-06-2020

DWDM Assignment-3

UNIT-3

N.V.S.k.kalyani
17UN1A0584
3rd B.Tech CSE-C

2 Marks

1Ans **Frequent pattern Mining : A Road map:**

Frequent pattern mining can be classified in various ways, based on the following criteria

i, Based on the completeness of patterns to be min
ii, Based on the levels of abstraction involved in th rule set.
iii, Based on the no.of data dimensions involved in the rules.
iv, Based on the types of values handled in the rul
v, Based on the kinds of rules to be mined.
vi, Based on the kinds of patterns to be mined.

2Ans **Apriori steps :-**

1. It is an influential algorithm for mining freque itemset for Boolean association rules, this uses pr knowledge of itemset properties.
2. It follows iterative approach known as "level-wise" search where 'k' itemsets are used to expl (k+1) itemset.
3. First, the set of frequent itemset are found denoted by 'L_1', which is used to find 'L_2' which in turn use to find 'L_3' and so on.
4. Apriori property says that "All non-empty subs of a frequent itemset must also be frequent", is based on the observations that. if an itemset doesnot satisfy

Step 5: If an item 'A' is added to the itemset 'I' then resulting itemset (I U A) cannot occur more frequently than 'I' i.e., (I U A) is not found frequent.

Therefore probability of (I U A) < min_sup

Step 6: A two step process is followed in generating candidate and frequent itemsets such as join and prune actions.

Algorithm

C_k: candidate itemset of size k and L_k: Frequent itemset of size k

L_1 = {frequent items};

for ($k=1$; $L_k != \phi$; $k++$) do begin

C_{k+1} = candidate generated from L_k;

for each transaction t in database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support.

end

return $L = \cup_k L_k$;

3Ans FP-growth :-

It is an Algorithm is an Efficient and Scalable method for mining the complete set of frequent pattern by pattern fragment growth, using an Extend prefix Tree structure for storing compressed and crucial Information about frequent pattern named frequent pattern tree.

17UN1A0584

4Au a) correlation is a term that is a measure of the strength of a linear relationship b/w two quantita-tive variables.

Ex:- positive correlation may be that the more exercise. The more calories you will burn.

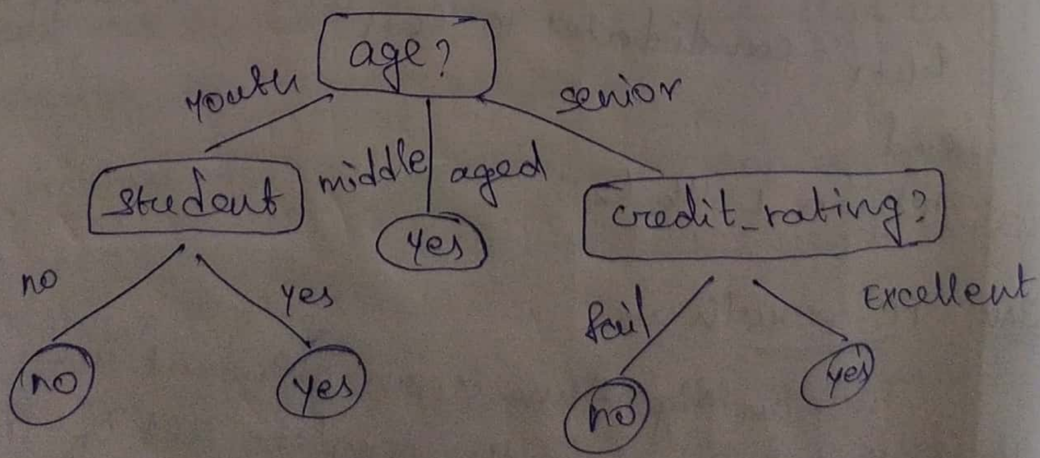
b) outline chi-square test with example

The chi-square test of association evaluates relationships b/w categorial variables.

5Au Decision Tree Induction is the learing of decision T from class-labeled training tuples. A Decision Tre

* flow-chart like Tree structure.
* Each Internal node denotes a test on an attribu
* each branch represents an outcome of the test
* Each leaf node holds a class label.
* The topmost node in a tree is the root node

Ex:-



17YN1A0584

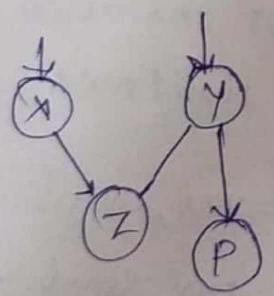
10 Marks

1A4 **Bayesian Network:-**

This can be deal with dependencies. This specifies joint conditional probability distribution. Bayesian belief Network are also known as belief network, Bayesian Network, and probabilistic networks.
A belief network is defined by two components

- A directed acyclic graph and
- A set of conditional probability tables.

A graphical model of causual relationship.
Represents dependency among the variables and gives a specification of joint probability distribution

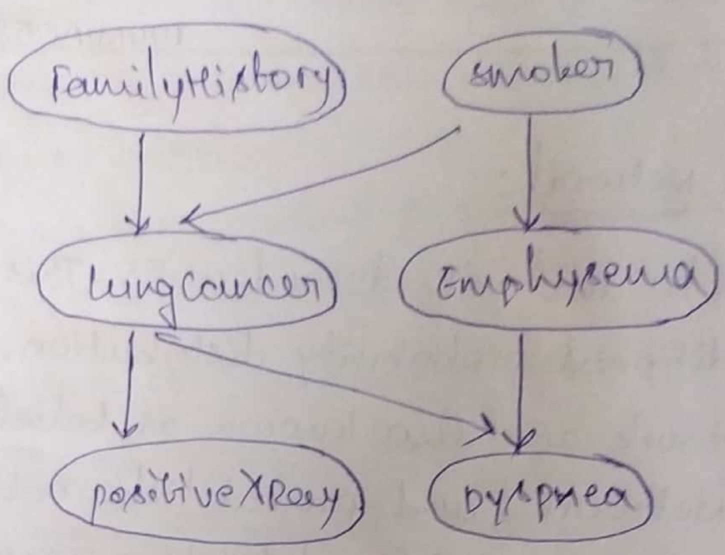


- Nodes: random variables
- Links: dependency
- X & Y are the parents of Z, and Y is the parent of P.
- No dependency b/w Z & P.
- Has no loops or cycles.

Ex:- A Simple Bayesian Network:
a) A proposed causal model represented by a directed acyclic graph.
b) The conditional probability table for the values of the variable, LungCancer(LC) showing each possible combination of the values of its parents node, FamilyHistory (FH and smoker (S).

17UNIA0584

a)



b)

	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

The conditional probability for each known valu of LungCancer is given for each possible combina-tion of values of its parents. For instance, from the upper leftmost and bottom rightmost entries

P(LungCancer = yes | FamilyHistory = yes, Smoker = yes) = 0

P(LungCancer = no | FamilyHistory = no, Smoker = no) = 0

let $\mathbf{X} = (x_1, \ldots x_n)$ be a data tuple described by the variables or attributes $Y_1, \ldots Y_n$ respectively.

$P(x_1, \ldots, x_n)$ is the probability of a particula combination of values of X, and the values for $P(x_i | parents(Y_i))$ correspond to the entries in the CPT for Y_i.

2 a) classification by Backpropogation

Ans :- • Iteratively process a set of training tuples and compare the network prediction with the actual known target value.

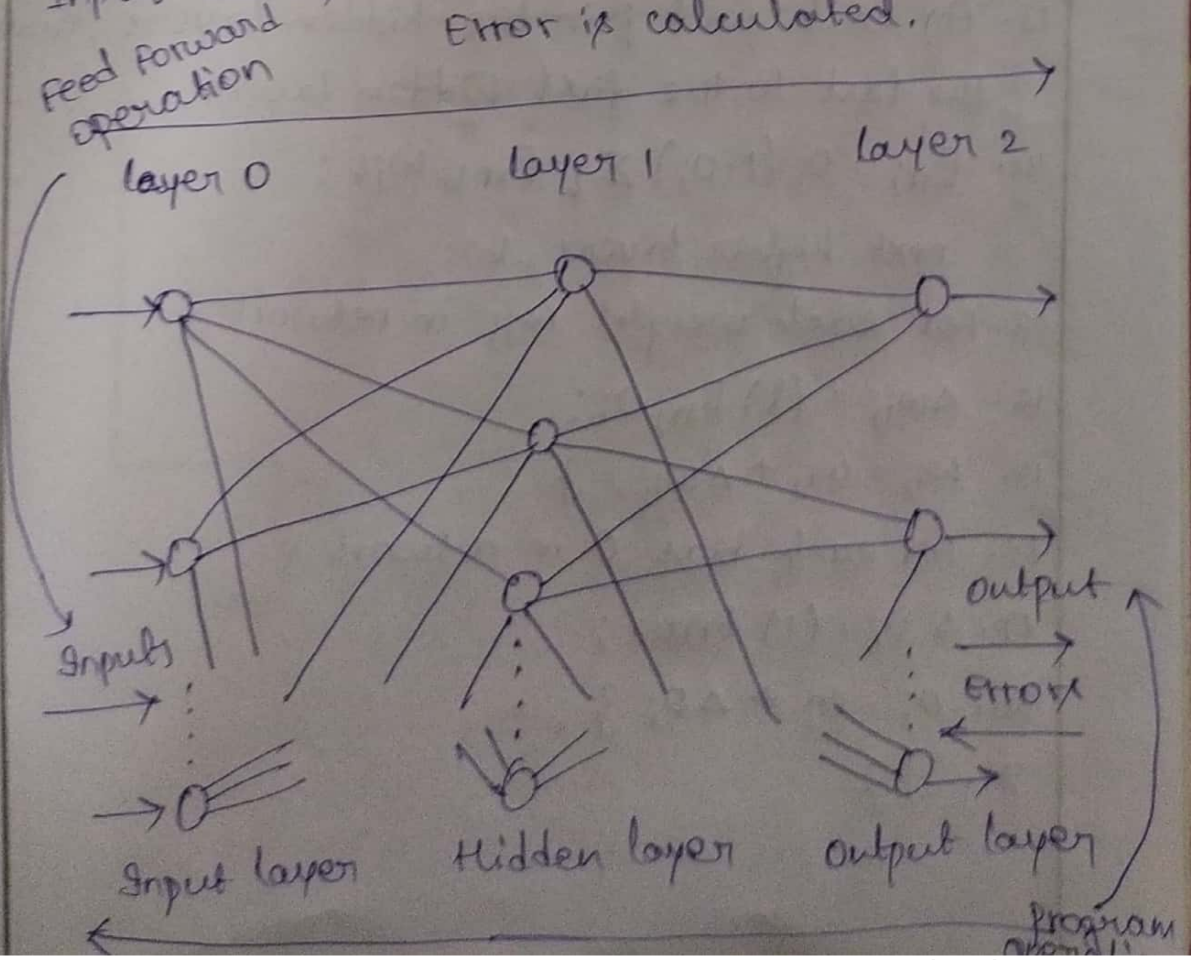
• For each training tuple, the weights are modified to minimize the mean squared error b/w the networks prediction

• Modifications are made in the backward direction

Steps :

* Initialize weight and biases in the network
* propogate the inputs forward.
* Back propogate the error.
* Terminating condition

Input vectors → actual value is calculated then Error is calculated.



17UNIA0584

Method :-

1. Initialize all weights and bases in network ;
2. while terminating condition is not satisfied {
3. For each training tuple x in D {
4. // propogate the Inputs forward ;
5. For each Input layer until j {
6. $O_j = I_j$;
7. For each hidden or output layer unit j {
8. $I_j = \sum_i w_{ij} O_j + \theta_j$;
9. $O_j = \dfrac{1}{1+e^{-I_j}}$ } ;
10. // Backpropogate the errors ;
11. for each unit j in the output layer
12. $Err_j = O_j (1-O_j)(T_j - O_j)$;
13. For each unit j in the hidden layers, from the last to the first hidden layers.
14. $Err_j = O_j (1-O_j) \sum_k Err_k w_{jk}$;
next higher layers, k
15. For each weight w_{ij} in network {
16. $\Delta w_{ij} = (l) E_{rrj} O_i$;
17. $w_{ij} = w_{ij} + \Delta w_{ij}$; }
18. For each bias θ in network {
19. $\Delta \theta_j = (l)$ Error ;
20. $\theta_j = \theta_j + \Delta \theta_j$ } ;
21. } }

Scanned with CamScanner

174N1A0584

2b) Support Vector Machine:-

It is an algorithm for classification of both linear and nonlinear data. SVM is an algorithm that works as follows

- It uses a non linear mapping to transform the original training data into a higher dimension.
- within this new dimensions, it searches for the linear optional sperating hyperplane.
- with an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be seperated by a hyperplane.
- The SVM finds this hyperplane using support vectors ("essential training tuples) and margins.

SVM general philosophy will be as follows: