Question 2: Explain the difference between star and snowflake schema. How do you organize the indizes in each case?

Answer:

STAR	SNOWFLAKE
• A schema where a single fact table is connected to multiple dimension tables	A schema where dimension tables are normalized and organized into multiple related tables.
• Denormalized tables with redundant data.	Normalized tables with less redundata Kob data
• Simple join structure with fewr tables	Complex join structure with more tables
• Faster query performance	Relatively slower query performance
• More storage requirement	less storage requirement

• Easier to maintain	More complex to maintain
• Example: Sales table connected to dimension tables such as date, product, etc	Example: Sales fact table connected to normalized tables such as date product

- In Star Schema multiple indexes need to be created as the data is in a denormalized form.
- In ~~star~~ snowflake schema, indexes can be created on the primary keys of the relational tables as they uniquely identify the entries.

Question 3: What are outliers? How would you detect outliers in high dimensional datasets?

Answer:

An outlier is a data object that deviates significantly

from the rest of the objects, as if it were generated from a different mechanism, compared to other data objects.

Outlier Detection in High-Dimensional Data :

The detection of outliers in high-dimensional data uses the following techniques,

i) Extending Conventional Outlier Detection :

The conventional outlier detection methods like proximity-based models can be extended with some changes to account for the subspaces and sparsity.

HilOut is one such algorithm which ranks objects in the descending order of their sum of distances to the k-closest neighbors and labels the first 'l' data objects as outliers.

ii) Finding outliers in subspaces :

Applying conventional methods on all the dimensions is computationally infeasible and hence subspaces are constructed and the outliers of those subspaces are detected.

The subspaces are constructed using equal-depth ranges and the subspaces densities are measured and compared with each other. The subspace with density much lower than the average density is determined to contain outliers.

iii) Modeling High-Dimensional Outliers :

An alternative method is to develop new models to detect outliers directly. Such models usually avoid proximity measures and use different heuristics for the identification of outliers.

Question 4 : What is classification? Explain classification with logistic regression.

Answer :

Classification is the process of assigning a data object to a particular group based on previous data objects similar to the present data object. The process of assigning classes to data objects is carried out by statistical models or machine learning models.

Logistic Regression :

- It is a classification algorithm that is mostly used for binary classification.
- It uses gradient descent to find the optimal weights.

The steps followed by logistic regression are given below.

- Step 1 : Initialize weight matrix W of dimensions [1, features+1], extra 1 is for the bias

Step2 : Iterate through the data objects.

Step3 : For each data object perform the following,

$$o = x_i * W[:, 1:].T + w[0, 0]$$

$$y' = s \cdot \frac{1}{1+e^{-o}}$$

$$error = y[i] - y'$$

$$w[:, 1:] += l * x_i * error$$

$$w[0, 0] += l * error$$

Step 4 : Repeat step 2 and 3 until the error becomes negligible or the number of iterations reaches the maximum limit.

Code implementation on the Iris data is shown in the following pages.

Qn 1 : Identify frequent itemsets

Answer :

The given table consists of 35 tuples. The minimum support is chosen as 20 %.

∴ minimum support count = $\frac{20}{100} \times 35 = 7$

Frequent itemsets of size 1 :

France : 10	England : 10	US : 12
Italy : 12	Spain : 20	
China : 16	India : 25	
UAE : 12	Australia : 15	

Frequent itemsets of size 2 :

{Italy, Spain} : 7	{India, Italy} : 8
{Australia, India} : 10	{Australia, UAE} : 7
{India, US} : 7	{India, UAE} : 7
{China, India} : 10	{England, Spain} : 7

and compare its compatibility

{Spain, US} : 8 {Australia, China} : 7 {India, Spain}: 15

{China, Spain} : 9 {England, India} : 9

Frequent Itemsets of Size 3:

{England, India, Spain} : 7

There are no frequent itemsets of size greater than 3 for the chosen minimum support count.

The code used to verify the answer is attached in the following pages.

Question 2 : Explain the underlying concept of Bayesian Belief Networks.

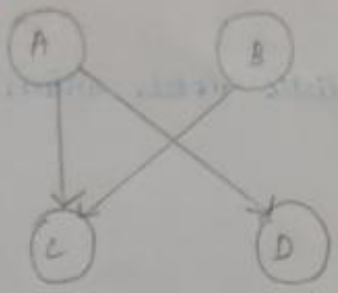
- Bayesian Belief Networks are probabilistic models which are primarily used for classification.
- In contrast to Naive Bayes Classifier, they do not assume Class Conditional Independence.
- A belief network is made up of 2 components

  DAG - Directed Acyclic Graph which represents the dependencies between the attributes through nodes and edges

  CPT - Conditional Probability Table, which is created for each attribute based (conditioned) on its parents.

Sample belief Network :

DAG :



CPT :

	$A\bar{B}$	$\bar{A}B$	AB	$\bar{A}\bar{B}$
C				
$\bar{C}$				

- From the dag, we can construct the CPT for all the attributes using the given tuples.
- The probability is the calculated using the following formula,

$$P(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(x_i)),$$

where Parents (x_i) is the attributes on which the attribute x_i is dependent on.

Training the networks:

When no value is missing and the conditional dependencies are known, the network topology can be easily constructed from the given tuples.

But when, some values are missing or we dont know the conditional dependencies, the network has to be trained using some techniques like gradient descent, whose steps are illustrated below.

The weights or the conditional probabilities are maintained in a 3-dimensional matrix called W. An entry 'Wijk' represents the conditional probability for an attribute Yi at state 'j', whose parent's state is 'ik'.

The task is to maximize the probability by adjusting the weights, $P_w(D) = \prod_{d \in I}^{|D|} P_w(X_d)$.

2nd ..

Step 1: Computing the gradients.

Calculating the gradients for the product of probabilities will lead to very small values and very small updates to the weights. Hence the gradients are calculated after applying 'log' to the product.

$$\frac{\partial \ln P_w(D)}{\partial w_{ijk}} = \sum_{d=1}^{|D|} \frac{P(Y_i = y_{ij}, U_i = u_{ik})}{w_{ijk}}$$

Step 2: Update the weights

The weights are updated in the direction of the gradient as we want to maximize the probability.

$$w_{ijk} \leftarrow w_{ijk} + (l) \frac{\partial \ln P_w(D)}{\partial w_{ijk}}$$

where l is the learning rate.

3rd step: Renormalizing the weights

Since the weights w_{ijk} represent probabilities, they must lie between 0 and 1 i.e. $\sum_j w_{ijk} = 1$ for all i and k.

This is how a belief network is trained using gradient descent.

Question 3: Explain a) Entropy b) Information Gain

c) Gini Index.

Answer:

a) Entropy:

It is the expected amount of information needed to classify the tuples at a node N in a decision tree. It is calculated using the formula,

$$Ent(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where m is the number of classes.

For example, suppose there are 14 tuples out of 9 belong to class 1 and remaining belong to class 2. Then entropy is

$$Ent(D) = \frac{-9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= 0.940$$

b) Information Gain :

Information Gain is the reduction in amount of information needed to classify tuple at node N for the attribute chosen for splitting the tuples.

Suppose an attribute A has v classes, which is chosen as the splitting attribute at a node N, then the information still required to classify the tuples is given as,

$$Ent_A(D) = \sum_{i=1}^{v} \frac{|D_i|}{|D|} Info(D_i)$$

The gain in information is given as,

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

c) Gini Index :

The gini index measures the impurity of the tuples at a node N. The impurity of a node will be high when there's an even distribution of the classes present at node N.

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Gini index for the given example is,

$$\text{Gini}(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

Question 4: What is a 'Pivot table'? How is the phrase 'Data Cube' related with pivot tables?

Answer:

A pivot table is a type of report or analysis that allows you to summarize and analyze data that has been organized into a multidimensional data structure, known as a data cube.

Pivot tables allow you to quickly and easily summarize data based on different criteria, such as time, geography or product category. It can also be used to slice the data to create multiple views of the data cube and analyse specific subsets to identify trends and patterns.

Example:

Suppose we have a data cube with the following dimensions:

- Time: Years and Quarters

The gain in information is given as,

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

c) Gini Index :

The gini index measures the impurity of the tuples at a node N. The impurity of a node will be high when there's an even distribution of the classes present at node N.

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Gini index for the given example is,

$$\text{Gini}(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- Geography : Regions and Countries
- Products : Categories and Subcategories.

A pivot table can be formed using any of these dimensions and the summarized data can be analyzed.

Question 5 : Differentiate probability and likelihood

Answer :

Probability	likelihood
• The measure of the likelihood of an event occuring	The measure of the compatibility of a hypothesis with the observed data
• Ranges from 0 to 1	It has no upper limit
• Calculated based on the total number of outcome and number of favourable outcome	Calculated based on the observed data and hypothesis
• Used to predict the chances of an event occuring	Used to test a hypothesis and compare its compatibility

Example:

When rolling a die, probability is used to predict the chance of an event occuring, while likelihood is used to test if the dice is biased towards any particular event.

Question 1 : Find the similar and dissimilar statements from the given set of statements.

Answer :

The given statements have to be first converted into a numerical format.

The statements are converted into a numerical format using a process called Term Frequency - Inverse Document Frequency.

$$tf\text{-}idf(t) = tf(t) * idf(t)$$

$$idf(t) = \log\left(\frac{1+d}{1+df(d,t)}\right)$$

$tf(t) \rightarrow$ Term frequency, number of times 't' occurs in a document.

$df(d,t) \rightarrow$ Document frequency, number of documents that contain term 't'

After vectorizing each statement, their similarity is measured using cosine-similarity.

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$

Measuring the cosine similarity, we get -

Most similar documents : 1 and 5 with a score of 0.4999

Most dissimilar documents : 1 and 3 with a score of 0.3193

Code implementation is given in the following pages.

Question 2: Construct decision trees on the given data using entropy and gini index.

Answer:

Using Entropy:

The decision tree constructed using entropy is shown in the next page.

At root node, the attribute which gives the maximum information gain is 'PO6' for the condition PO6 <= 1.5.

The entropy of root node is $-\left[\frac{16}{40}\log_2\left(\frac{16}{40}\right) + \frac{12}{40}\log_2\left(\frac{12}{40}\right) + \frac{12}{40}\log_2\left(\frac{12}{40}\right)\right]$, which is 1.5741.

Splitting using 'PO6' reduces the entropy required to 1.309 and 1.474, in it's respective branches. Repeated iterations based on gain entropy is carried out, but the maximum depth is cut down to 3 to reduce overfitting.

Using Gini Index :

The Decision Tree constructed using Gini Index is shown in the next page.

Similar to entropy, the root node chooses 'POS' to split the tuples. The impurity at each node and the gain is calculated at each node using following formula,

$$\text{Gini Index}(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$\text{Gain}(D) = \text{Gini Index}(D) - \min\left\{ \frac{|D_{s1}|}{|D|} \text{Gini Index}(D_{s1}) + \frac{|D_{s2}|}{|D|} \text{Gini Index}(D_{s2}), \right.$$

for A in $(a_1, a_2, \ldots a_n)$

The maximum depth is again restricted to 3 to prevent overfitting.

Question 3 : Complete the table and answer the questions

Answer :

Boys in class = 40

Girls in class = 20

Total Students = 60

OBSERVATION	Boys %	No. of Boys	Girls %	No. of Girls
Newspapers	60	24	40	8
Games	40	16	60	12
Movie	80	32	80	16
Shopping	25	10	75	15
Foreign Country	45	18	85	17
Partying	50	20	60	12
Travelling	60	24	20	4
Whatsapp	100	40	100	20
1st class CGPA	40	16	55	11
Vehicle	35	14	20	4

1) $P(B/G', Th, C', V')$

$\propto P(G'/B) * P(Th/B) \ldots$

$= P(B/G') * P(B/Th) * P(B/C') * P(B/V')$

$= [P(G'/B) * P(B)] * [P(Th/B) * P(B)] * [P(C'/B) * P(B)] *$

$[P(V'/B) * P(B)]$

$= [0.6 * \frac{2}{3}] * [0.6 * \frac{2}{3}] * [0.6 * \frac{2}{3}] * [0.65 * \frac{2}{3}]$

$=$

2) $P(G/P, V, F', N') = P(G/P) * P(G/V) * P(G/F') * P(G/N')$

$= [P(P/G) * P(G)] * [P(V/G) * P(G)] * [P(F'/G) * P(G)]$

$* [P(N'/G) * P(G)]$

$= [0.6 * 0.33] * [0.2 * \frac{1}{3}] * [0.15 * \frac{1}{3}] *$

$[0.6 * \frac{1}{3}]$

$=$

3) $P(G_1 / GA', Tr, \overline{CGPA}, V')$

$= P(G_1/GA') * P(G_1/Tr) * P(G_1/\overline{CGPA}) * P(G_1/V')$

$= \left[P(GA'/G_1) * P(G_1)\right] * \left[P(Tr/G_1) * P(G_1)\right] * \left[P(\overline{CGPA}/G_1) * P(G_1)\right] *$

$\left[P(V'/G_1) * P(G_1)\right]$

$= \left[0.4 * \frac{1}{3}\right] * \left[0.2 * \frac{1}{3}\right] * \left[0.45 * \frac{1}{3}\right] * \left[0.8 * \frac{1}{3}\right]$

$=$

4) $P(B/P, V, F', N') = P(B/P) * P(B/V) * P(B/F') * P(B/N')$

$= \left[P(P/B) * P(B)\right] * \left[P(V/B) * P(B)\right] * \left[P(F'/B) * P(B)\right] *$

$\left[P(N'/B) * P(B)\right]$

$= \left[0.5 * \frac{2}{3}\right] * \left[0.35 * \frac{2}{3}\right] * \left[0.55 * \frac{2}{3}\right] * \left[0.4 * \frac{2}{3}\right]$

$=$

5) Influence of whatsapp and watching movies in deciding gender:

Whatsapp's influence is same on both boys and girls as it is 100%, while watching movies' influence differs.

$$P(\text{Boy}/\text{Watching Movie}) = \frac{P(\text{Watching Movie}/\text{Boy}) * P(\text{Boy})}{P(\text{Watching Movie})}$$

$$= \frac{0.8 * 2/3}{48/60} = \frac{2}{3}$$

$$P(\text{Girl}/\text{Watching Movie}) = \frac{P(\text{Watching Movie}/\text{Girl}) * P(\text{Girl})}{P(\text{Watching movie})}$$

$$= \frac{0.8 * 1/3}{48/60} = \frac{1}{3}$$

Influence on deciding gender using 'Watching Movies' is