Krithika Devi Chandran

CDB22DW0222,Data

1. ***What is data normalization? How is it different from database normalization (1st/2nd/3rd)?***

* Data in database is stored in terms of enormous quantity. Retrieving certain data will be a tedious task if the data is not organized correctly.
* So, with the help of **normalization** we can organize this data and also reduce the redundant data.
* Normalization split a large table into smaller tables and define relationships between them to increase the clarity in organizing data in databases.
* **First Normal Form (1NF):**
* Using **1NF**, data redundancy increases, as there will be many columns with dame data in multiple rows but each row as a whole will be unique. That is, ***each column is unique in 1NF***.

***Sample employee table***

|  |  |  |
| --- | --- | --- |
| **Employee** | **Age** | **Department** |
| Melvin | 32 | Marketing, Sales |
| Edward | 45 | Quality Assurance |
| Alex | 36 | HR |

***Employee table following 1NF***

|  |  |  |
| --- | --- | --- |
| **Employee** | **Age** | **Department** |
| Melvin | 32 | Marketing |
| Melvin | 32 | Sales |
| Edward | 45 | Quality Assurance |
| Alex | 36 | HR |

* **Second Normal Form (2NF):**
* The entity should be considered already in 1NF, and all attributes within the entity should depend solely on the unique identifier of the entity.
* For a table to be in the 2NF, it must satisfy two conditions:
* The table should be in 1NF.
* There should be **no** partial dependency. (Partial dependency exists when any attribute in the table depends only on a part of the primary key & not on the complete primary key).

***Sample products table***

|  |  |  |
| --- | --- | --- |
| **ProductID** | **Product** | **Brand** |
| 1 | Monitor | Apple |
| 2 | Monitor | Samsung |
| 3 | Scanner | HP |
| 4 | Headphone | JBL |

***Products table following 2NF***

***Products table***

|  |  |
| --- | --- |
| **ProductID** | **Product** |
| 1 | Monitor |
| 2 | Scanner |
| 3 | Headphone |

***Brand table***

|  |  |
| --- | --- |
| **BrandID** | **Brand** |
| 1 | Apple |
| 2 | Samsung |
| 3 | HP |
| 4 | JBL |

***Products Brand table***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PbID** | **ProductID** | **BrandID** | **Product** | **Brand** |
| 1 | 1 | 1 | Monitor | Apple |
| 2 | 1 | 2 | Monitor | Samsung |
| 3 | 2 | 3 | Scanner | HP |
| 4 | 3 | 4 | Headphone | JBL |

* **Third Normal Form (2NF):**
* For a table to be in the 3NF, it must satisfy two conditions:
* The table should be in 2NF.
* And it should **not** have transitive dependency. (Transitive dependency is when column depends on non-prime attributes but on the prime attributes).

**Example:-** If there are score\_id, student\_id, subject\_id, marks, exam\_name and total\_marks in a **score table,** then total\_marks depends on exam\_name and not depends on id’s.

Thus we say here the total\_marks is a transitive dependency. Whereas, exam\_name depends on the primary key student\_id and subject\_id, we say these id’s as composite key.

To remove this transitive dependency, get exam\_name and total\_marks in a separate table.

1. ***What is a distribution? What are the uses for frequency and probability distribution?***

* The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur.
* A distribution is simply a collection of data or scores on a variable. Usually, these scores are arranged in order from ascending to descending and then they can be presented graphically.
* The distribution provides a parameterized mathematical function which will calculate the probability of any individual observation from the sample space.

**Frequency Distribution:**

1. Frequency distribution is the organization of raw data in table form, using classes and frequencies.
2. Two type of frequency distribution that are most often used are the categorical frequency distribution and the grouped frequency distribution.
3. It is an organized tabulation/graphical representation of the number of individuals in each category on the scale of measurement. It shows whether the observations are high or low and also whether they are concentrated in one area or spread out across the entire scale.
4. The important use case is, it presents a picture of how the individual observations are distributed in the measurement scale.

**Probability Distribution:**

1. Probability distribution yields the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences.
2. It is used widely in simulation studies where to stimulate a real phenomenon or process as closely as possible and to study problems where actual experimentation is impossible.
3. There are discrete and continuous type of probability distribution. It can be used in either of the one way.

* By using a discrete distribution, you can define the probability of occurrence of each value of a discrete random variable. It has only countable values.
* By using a continuous distribution, you can define the probability of the possible values of a continuous random variable. It has a set of possible infinite and uncountable values.

1. ***What is a decision? How's it different from inference?***

* Decisions are decisions made on the basis of observations of a phenomenon that obeys probabilistic laws that are not completely known.

Decision problems are capable of being tightly formulated in terms of initial conditions and choices or course of action, with their consequences. That is, such consequences are not known with certainty but are expressed as a set of probabilistic outcomes.

* Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability Inferential statistical analysis infer properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

1. ***Google- what is Gini in probability, and explain in your own terms***

* Gini is a statistical measure that determines the likelihood that a certain characteristic would be categorized wrongly when a random sample is chosen. It is commonly used measurement of inequality.
* It is a metric to measure how often a randomly chosen element would be incorrectly identified.
* Decision tree algorithm CART(Classification and Regression Trees) uses the Gini method to create split points.
* Gini (Gini Impurity) considers a binary split for each attribute. You can compute a weighted sum of the impurity of each partition.
* In case of a discrete-valued attribute, the subset gives the minimum gini index for that is selected as a splitting attribute.
* In case of a continuous-valued attribute, the strategy is to select each pair of adjacent values as possible split-point and point with smaller gini index chosen as the splitting point. The attribute with minimum Gini index is chosen as the splitting attribute.

1. ***What is entropy?***

* Entropy is defined as a measure of randomness or disorder of a system.
* Entropy allows us to make precise statements and perform computations with regard to one of life’s most pressing issues: not knowing how things will turn out.
* Entropy, in other words, is a **measure of uncertainty**.
* In Decision Tree algorithm, entropy helps us to build an appropriate decision tree for selecting the best splitter. Entropy can be defined as a measure of purity of the sub split. Entropy always lies between 0 and 1.

1. ***What is euclidean distance?***

* The Euclidean Distance represents the **shortest distance between two points**. Because of this, it represents the Pythagorean Distance between two points, which is calculated using:

**d = √[(x2 – x1)2 + (y2 – y1)2]**

* They are used extensively in the k-nearest neighbour classification systems

1. ***What's the difference between correlation and covariance?***

|  |  |
| --- | --- |
| **CORRELATION** | **COVARIANCE** |
| Correlation is a statistical method used to determine whether a linear relationship between variables exists. That is, there is said to be correlation between two, when change in one results in change in another. | Covariance talks about the direction of the relationship between the two variables (positive or negative). |
| A measure used to reprensent how strongly two random variables are related. Shows a strength of variables. | A measure used to indicate the extent to which two random variables change in random. |
| Correlation is a scaled down version of covariance. | Covariance is considered as a part of correlation. |
| Value here lies between -1 and +1 | Value here lies between -∞ to +∞ |
| Correlation is a unit-free measure. | Covariance value is the product of the units of the variables. |
| There would be no change in correlation due to scale. | Any change in scale affects covariance. |

1. ***What is mean squared error?***

* Mean Squared Error (MSE) is calculated by taking the average of the square of the difference between the original and predicted values of the data.
* It is the most common evaluation metric to analyse the prediction result in machine learning.

1. ***What is the difference between covariance, standard deviation and mean squared error?***

* A measure used to indicate the extent to which two random variables change in random is called **covariance**.
* **Standard deviation** is a measure which shows how much variation such as spread, dispersion from the mean exists. It is a popular measure of variability because it returns to the original units of measure of the data set. If the data points are highly deviated (spread) out from the mean, then it has a high variance.
* **Mean Squared Error (MSE)** is calculated by taking the average of the square of the difference between the original and predicted values of the data. It is the most common evaluation metric to analyse the prediction result in machine learning.

**===\*\*\*===**