**THEORITICAL LEARNING – EXERCISE 1**

**1.what is normalization?**

Normalization is a scaling technique or a mapping technique or a pre-processing stage. **Data normalization** is removing the redundancies and making it easy to store, visualize, and analyze it. For **databases**, there are different levels of **normalization,** each with its own characteristics and definitions. They’re referred to 1 normal form (1NF), 2 normal form (2NF), 3 normal form (3NF), etc.

* **1st Normal form:**

If the domain of each attribute contains only atomic (indivisible) values, and the value of each attribute contains only a single value from that domain.

* **2nd Normal Form:**

if it is in 1NF and all non-key attributes are fully functional dependent on the primary key.

* **3rd Normal Form:**

if it is in 2NF and no transition dependency exists.

**2. 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 helps us to understand a variable by giving us an idea of the values that the variable is most likely to obtain.

**Frequency distribution** is where we will observe how often certain value occurs.

**probability distribution** gives the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences

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

Statistical decisions are**decisions made based on the observations of a phenomenon that obeys probabilistic laws that are not completely known**

Statistical inference is the**process of using data analysis to infer properties of an underlying distribution of probability**. Inferential statistical analysis infers properties of a population

Example: testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

**4. what is Gini in probability and explain in your own terms?**

Gini calculates the probability of a certain randomly selected feature that was classified incorrectly.

The Gini Index varies**between 0 and 1**, where 0 represents purity of the classification and 1 denotes random distribution of elements among various classes.

**5. What is entropy?**

Entropy is the process, that describes, whether the quantity is a spontaneous process and has a probability of occurring in a defined direction, or a non-spontaneous process and will not proceed in the defined direction, but in the reverse direction.

Entropy is basis of mutual information that quantifies the relationship between two variables. Entropy in classification trees.

**6. What is Euclidean distance?**

Euclidean Distance represents the shortest distance between two points.

In machine learning algorithms including K-Means use this distance metric to measure the similarity between observations.

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

**Covariance:**

* + Describes the extent to which the variables are dependent.
  + From covariance, Correlation can be deduced
  + Scalability affects covariance.
  + Covariance ranges from
  + Covariance has definite unit as it is deducted by multiplying 2 numbers and their units.

**Correlation.**

* Describes how strongly the variables are related.
* Correlation provides the measure of covariance on standard scale.
* Scalability does not affect correlation.
* Covariance ranges from -1 to +1
* Covariance has no definite unit.

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

The MSE is the average squared distance between the observed and predicted values. Mean squared error (MSE) measures the amount of error in statistical models.

* If model has no error, the MSE equals zero.
* If model has less errors, the MSE decreases.
* If model has more errors, the MSE increases.
* Formula:

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

**Covariance:**

Describes the joint variability of the variables. Covariance falls under the multivariate analysis. For the covariance is represented joint matrix in which diagonals are variance, lower and upper matrix are covariance in nature.

**Standard deviation:**

Standard deviation is the square root of variance, and it measures the variability of datasets absolutely.it can never be negative.

**Mean Squared Error:**

Describes the difference between observed and predicted value of each datapoint in the dataset.

we obtain the MSE for each value in the dataset and observe which value lies farer or nearer to the straight line.