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

Soln. Normalization is the process of minimizing redundancy from a relation or set of relations. Redundancy in relation may cause insertion, deletion, and update anomalies. So, it helps to minimize the redundancy in relations. Normal forms are used to eliminate or reduce redundancy in database tables.

1NF

- A table is said to be in 1 NF, if every cell contains atomic value
- For every multi-valued attribute, we have to repeat the entire information
- This is important from E-R point of view as, for every multi-valued attribute, a separate table should be created whenever E-R is transferred into relational model
- Every column should contain value of same domain Like; if there is an attribute, Roll No, it should contain only Roll No and nothing else.
- Order of rows and columns is irrelevant
- No two columns should be the same.

2NF

- To be in second normal form, a relation must be in first normal form
- Relation must not contain any partial dependency.
- A relation is in 2NF if it has No Partial Dependency, i.e., no non-prime attribute (attributes which are not part of any candidate key) is dependent on any proper subset of any candidate key of the table.

3NF

- A relation is in third normal form, if there is no transitive dependency for non-prime attributes as well as it is in second normal form.
- A relation is in 3NF if at least one of the following condition holds in every non-trivial functional dependency X –> Y
    1. X is a super key.
    2. Y is a prime attribute (each element of Y is part of some candidate key).

Q2) What is a distribution? What are the uses for frequency and probability distribution?

Soln. A statistical distribution, or probability distribution, describes how values are distributed for a field. In other words, the statistical distribution shows which values

are common and uncommon. There are many kinds of statistical distributions, including the bell-shaped normal distribution.
Frequency distribution is a curve that gives us the frequency of the occurrence of a particular data point in an experiment. This is usually the limit of a histogram of frequencies when the data points are very large and the results can be treated to be varying continuously instead of taking on discrete values.
Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics

Q 3) What is a decision? How's it different from inference?

Soln. Decision theory, in statistics, a set of quantitative methods for reaching optimal decisions. A solvable decision problem must be capable of being tightly formulated in terms of initial conditions and choices or courses of action, with their consequences. 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, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Q4) Google- what is Gini in probability, and explain in your own terms.

Soln. The Gini coefficient (Gini index or Gini ratio) is a statistical measure of economic inequality in a population. The coefficient measures the dispersion of income or distribution of wealth among the members of a population

Q5) What is entropy?

Soln. Entropy is defined as the randomness or measuring the disorder of the information being processed in Machine Learning. Further, in other words, we can say that entropy is the machine learning metric that measures the unpredictability or impurity in the system.

Q6) What is euclidean distance?

Soln. The Euclidean distance formula is used to find the distance between two points on a plane . Euclidean distance is calculated as the square root of the sum of the squared differences between the two vectors.

Q7) What's the difference between correlation and covariance?

Soln.

1. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
2. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.
3. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between $-\infty$ and $+\infty$.

Q8) What does squared error mean?

Soln. The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It's called the mean squared error as you're finding the average of a set of errors. The lower the MSE, the better the forecast.

Q9) What is the difference between covariance, standard deviation and mean squared error?

Soln. Covariance measures the direction of the relationship between two variables. A positive covariance means that both variables tend to be high or low at the same time. A negative covariance means that when one variable is high, the other tends to be low. Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.
Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).