Machine Learning

1.The value of correlation coefficient will always be:

C) between -1 and 1

2.Which of the following cannot be used for dimensionality reduction?

C) Recursive feature elimination

3.Which of the following is not a kernel in Support Vector Machines?

C) hyperplane

4.Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

D) Support Vector Classifier

5.5. In a Linear Regression problem, ‘X’ is independent variable and ‘Y’ is dependent variable, where ‘X’ represents weight in pounds.

If you convert the unit of ‘X’ to kilograms, then new coefficient of ‘X’ will be?.

A) 2.205 × old coefficient of ‘X’

6.As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

C) decreases

7.Which of the following is not an advantage of using random forest instead of decision trees?

C) Random Forests are easy to interpret

8. Which of the following are correct about Principal Components?

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

9.Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.

C) Identifying spam or ham emails

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10.Which of the following is(are) hyper parameters of a decision tree?

A) max_depth

D) min_samples_leaf

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans: ->In statistics, an outlier is a data point that differs significantly from other observations.

An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

The most effective way to find outliers is by using the interquartile range (IQR)

->The interquartile range is a number that indicates the spread of the middle half or the middle 50% of the data.

It is the difference between the third quartile (Q3) and the first quartile (Q1). IQR = Q3 – Q1. The IQR can help determine outliers.

->IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts.

Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of the data.

Q2 represents the 50th percentile of the data.

Q3 represents the 75th percentile of the data.

If a dataset has 2n / 2n+1 data points, then

Q1 = median of the dataset.

Q2 = median of n smallest data points.

Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1.

The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

12. What is the primary difference between bagging and boosting algorithms?

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.

Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

In Bagging the result is obtained by averaging the responses of the N learners (or majority vote).

However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.

13. What is adjusted R2 in linear regression. How is it calculated?

The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model.

In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.

The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

It decreases when a predictor improves the model by less than expected by chance.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

Every time a independent variable is added to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

14. What is the difference between standardization and normalization?

Normalization typically means rescales the values into a range of [0,1].

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1

Normalization is good to use when we know that the distribution of our data does not follow a Gaussian distribution.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution.

Normalization is used to minimize the redundancy from a relation or set of relations.

It is also used to eliminate the undesirable characteristics like Insertion, Update and Deletion Anomalies.

Data standardization is about making sure that data is internally consistent; that is, each data type has the same content and format.

Normalization divides the larger table into the smaller table and links them using relationship.

Standardized values are useful for tracking data that isn't easy to compare otherwise.

The benefits of normalization include: Searching, sorting, and creating indexes is faster, since tables are narrower, and more rows fit on a data page.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent data set.

It is a technique for evaluating machine learning models by training several models on subsets of the available

Input data and evaluating them on the complementary subset of the data.

Cross Validation is a very useful technique for assessing the effectiveness of model, particularly in cases where we need to mitigate overfitting.

4 Types of Cross Validation:

Holdout Method.

K-Fold Cross-Validation.

Stratified K-Fold Cross-Validation.

Leave-P-Out Cross-Validation.

An advantage of using this method is that we make use of all data points and hence it is low bias.

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point.

If the data point is an outlier it can lead to higher variation.

Statistics

1. What is central limit theorem and why is it important?

The central limit theorem states that if we have a population with mean µ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed.

The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases.

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases.

2. What is sampling? How many sampling methods do you know?

A sampling method is a procedure for selecting sample members from a population.

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.

Types of sampling:

1. Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

2. Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

3. What is the difference between type1 and typeII error?

Type1 Error

1. Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.

Type I error is equivalent to false positive.

It is a false rejection of a true hypothesis.

Type I error is denoted by α.

The probability of type I error is equal to the level of significance.

It can be reduced by decreasing the level of significance.

It is caused by luck or chance.

Type I error is similar to a false hit

Type I error is associated with rejecting the null hypothesis.

It happens when the acceptance levels are set too lenient.

Type2 Error

Type II error is the error that occurs when the null hypothesis is accepted when it is not true.

Type II error is equivalent to a false negative.

It is the false acceptance of an incorrect hypothesis.

Type II error is denoted by β.

The probability of type II error is equal to one minus the power of the test.

It can be reduced by increasing the level of significance.

It is caused by a smaller sample size or a less powerful test.

Type II error is similar to a miss.

Type II error is associated with rejecting the alternative hypothesis.

It happens when the acceptance levels are set too stringent.

4. What do you understand by the term Normal distribution?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

In graph form, normal distribution will appear as a bell curve.

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena.

For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

It is also known as the Gaussian distribution and the bell curve.

A normal distribution is the proper term for a probability bell curve.

In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

Normal distributions are symmetrical, but not all symmetrical distributions are normal.

In reality, most pricing distributions are not perfectly normal.

5. What is correlation and covariance in statistics?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).

It's a common tool for describing simple relationships without making a statement about cause and effect.

Correlation is a statistical measure that indicates how strongly two variables are related. A zero correlation indicates that there is no relationship between the variables.

A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down.

Covariance indicates the direction of the linear relationship between variables.

Covariance is a statistical tool that is used to determine the relationship between the movements of two asset prices.

When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

Covariance is a measure of how much two random variables vary together.

6. Differentiate between univariate, Biavariate, and multivariate analysis.

Univariate statistics summarize only one variable at a time.

This type of data consists of only one variable.

The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.

It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Univariate analysis is conducted through several ways which are mostly descriptive in nature

Frequency Distribution Tables

Histograms

Frequency Polygons

Pie Charts

Bar Charts

Bivariate statistics compare two variables.

This type of data involves two different variables.

The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.

Example of bivariate data can be temperature and ice cream sales in summer season.

Bivariate analysis is conducted by:

Correlation coefficients

Regression analysis

Linear regression

Simple regression

Polynomial regression

General linear model

Discrete choice

Binomial regression

Binary regression

Logistic regression

Multivariate statistics compare more than two variables.

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

It is similar to bivariate but contains more than one dependent variable.

The ways to perform analysis on this data depends on the goals to be achieved.

Commonly used multivariate analysis technique include –

Factor Analysis

Cluster Analysis

Variance Analysis

Discriminant Analysis

Multidimensional Scaling

Principal Component Analysis

Redundancy Analysis

7. What do you understand by sensitivity and how would you calculate it?

The technique used to determine how independent variable values will impact a particular dependent variable under a given set of assumptions is defined as sensitive analysis. It’s usage will depend on one or more input variables within the specific boundaries, such as the effect that changes in interest rates will have on a bond’s price.It is also known as the what – if analysis. Sensitivity analysis can be used for any activity or system.

All from planning a family vacation with the variables in mind to the decisions at corporate levels can be done through sensitivity analysis.

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.

The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.

The test provides evidence concerning the plausibility of the hypothesis, given the data.Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

In statistics, a two-tailed test is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. It is used in null-hypothesis testing and testing for statistical significance. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.

9. What is quantitative data and qualitative data?

Quantitative data is information about quantities, and therefore numbers, and qualitative data is descriptive, and regards phenomenon which can be observed but not measured, such as language.

Qualitative data is non-statistical and is typically unstructured or semi-structured. This data isn’t necessarily measured using hard numbers used to develop graphs and charts. Instead, it is categorized based on properties, attributes, labels, and other identifiers

Contrary to qualitative data, quantitative data is statistical and is typically structured in nature – meaning it is more rigid and defined.

This data type is measured using numbers and values, making it a more suitable candidate for data analysis.

Whereas qualitative is open for exploration, quantitative data is much more concise and close-ended.

It can be used to ask the questions “how much” or “how many,” followed by conclusive information.

10. How to calculate range and interquartile range?

The interquartile range is a measure of where the “middle fifty” is in a data set.

Where a range is a measure of where the beginning and end are in a set, an interquartile range is a measure of where the bulk of the values lie.

That’s why it’s preferred over many other measures of spread when reporting things like school performance or SAT scores.

The interquartile range formula is the first quartile subtracted from the third quartile:

IQR = Q3 – Q1.

In statistics, the range is the spread of your data from the lowest to the highest value in the distribution. The interquartile range is the best measure of variability for skewed distributions or data sets with outliers.

11. What do you understand by bell curve distribution?

A bell curve is the informal name of a graph that depicts a normal probability distribution.

The term obtained its name due to the bell-shaped curve of the normal probability distribution graph.

The bell curve is perfectly symmetrical. It is concentrated around the peak and decreases on either side.

In a bell curve, the peak represents the most probable event in the dataset while the other events are equally distributed around the peak.

The peak of the curve corresponds to the mean of the dataset (note that the mean in a normal probability distribution also equals the median and the mode).

12. Mention one method to find outliers.

Using Z-scores to Detect Outliers

Z-scores can quantify the unusualness of an observation when your data follow the normal distribution.

Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that an observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

To calculate the Z-score for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation.

Mathematically, the formula for that process is the following:

The further away an observation’s Z-score is from zero, the more unusual it is.

A standard cut-off value for finding outliers are Z-scores of +/-3 or further from zero. The probability distribution below displays the distribution of Z-scores in a standard normal distribution. Z-scores beyond +/- 3 are so extreme you can barely see the shading under the curve.

13. What is p-value in hypothesis testing?

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

A p-value is a measure of the probability that an observed difference could have occurred just by random chance.

The lower the p-value, the greater the statistical significance of the observed difference.

P-value can be used as an alternative to or in addition to pre-selected confidence levels for hypothesis testing.

The p-value approach to hypothesis testing uses the calculated probability to determine whether there is evidence to reject the null hypothesis.

The null hypothesis, also known as the conjecture, is the initial claim about a population (or data generating process).

The alternative hypothesis states whether the population parameter differs from the value of the population parameter stated in the conjecture.

14. What is the Binomial Probability Formula?

Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment). If the probability of success on an individual trial is p , then the binomial probability is nCx·px·(1−p)n−x .

15. Explain ANOVA and its applications.

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts:

Systematic factors and Random factors.

The systematic factors have a statistical influence on the given data set, while the random factors do not.

Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

One real-life application of analysis of variance is the recommendation of a fertilizer against others for the improvement of a crop yield.

SQL

1. Write a SQL query to show average number of orders shipped in a day (use Orders table).

```
with y as (select shippedDate, count(orderNumber) as total_orders from Orders)

select avg(total_orders) as AverageNumberOfOrdersShipped from y;
```

2. Write a SQL query to show average number of orders placed in a day.

```
with y as(select orderDate, count(orderNumber) as total_orders from Orders)

select avg(total_orders) as AverageNumberOfOrdersPlaced from y;
```

3. Write a SQL query to show the product name with minimum MSRP (use Products table).

```
select productName from products where MSRP=(select MIN(MSRP) from products);
```

4. Write a SQL query to show the product name with maximum value of stockQuantity.

```
select productName from products where quantityInStock=(select MAX(quantityInStock) from products);
```

5. Write a query to show the most ordered product Name (the product with maximum number of orders).

```
select productName from OrderDetails as a inner join Products as b on a.productCode = b.productCode group by b.productCode

order by count(orderNumber) desc limit 1;
```

6. Write a SQL query to show the highest paying customer Name.

```
select customerName, total_payment from

(select customerName, sum(amount) as total_payment from Customers as a inner join Payments b

on a.customerNumber = b.customerNumber group by customerName);
```

7. Write a SQL query to show cutomerNumber, customerName of all the customers who are from Melbourne city.

```
select customerNumber ,customerName from customers

    -> where city="Patna";
```

8. Write a SQL query to show name of all the customers whose name start with “N”.

```
 select customerName from customers

    -> where customerName LIKE 'N%';
```

9. Write a SQL query to show name of all the customers whose phone start with ‘7’ and are from city ‘Las Vegas’.

```
select customerName from customers

    -> where phone LIKE '7%' and city='Las Vegas';
```

10. Write a SQL query to show name of all the customers whose creditLimit < 1000 and city is either “Las Vegas” or ”Nantes” or “Stavern”.

```
       select customerName from customers

    -> where creditLimit<1000 and city IN ("Las Vegas","Nantes","Stavern");
```

11. Write a SQL query to show all the orderNumber in which quantity ordered <10.

```
select orderNumber from orderdetails where qunatityOrdered<10;
```

12. Write a SQL query to show all the orderNumber whose customer Name start with letter ‘N’.

```
select orderNumber from customers as a inner join orders as b on
a.customerNumber = b.customerNumber where customerName LIKE  'B%';
```

13. Write a SQL query to show all the customerName whose orders are “Disputed” in status.

```
Select customerName from orders as a inner join customers as b on
a.customerNumber= b.customerNumber where status= "Disputed";
```

14. Write a SQL query to show the customerName who made payment through cheque with checkNumber starting with H and made payment on “2004-10-19”.

```
select customerName from payments inner join customers
using(customerNumber) where paymentDate = "2004-10-19" and checkNumber
LIKE "H%";
```

15. Write a SQL query to show all the checkNumber whose amount > 1000.

```
   Select checkNumber from payments
  -> Where amount>1000;
```