Machine Learning

1. C
2. C
3. A
4. A
5. C
6. D
7. C
8. B,C
9. A,D
10. A,B,D
11. An extreme high or extreme low data point in relation to the closest data point and the rest of the nearby coexisting values in a dataset is referred to as an outlier. They are extreme values that significantly deviate from the overall pattern of values in a dataset.

    Finding the outlier is difficult, especially when there are several different data sets. As a result, we have many methods for spotting outliers for various data types. We can use the Z-Score approach to analyse data that is normally distributed, and we can use the IQR method to analyse data that is skewed. By dividing a data set into quartiles, the IQR is used to measure variability. The information is divided into 4 equal portions and arranged in ascending order. The values that divide the four equal halves are known as the first, second, and third quartiles, or Q1, Q2, and Q3, respectively. 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 is the median of the dataset, Q2 is the median of n smallest data points, Q3 is the 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?
    - While boosting is a method of combining predictions that belong to distinct types, bagging is the simplest approach to combine predictions that belong to the same type.
    - While boosting tries to reduce bias, not variation, bagging aims to reduce bias, not variance.
    - Every model is given the same weight for bagging, however in boosting, models are weighted in accordance with their performance.
    - Unlike boosting, where new models are influenced by the performance of older ones, bagging involves building each model independently.
    - In Bagging, various training data subsets are selected at random from the total training dataset using replacement. Every new subset in boosting contains the components that earlier models incorrectly classified.
    - While boosting seeks to lessen prejudice, Bagging attempts to address the issue of over-fitting.
    - While boosting is expanded to include gradient boosting, bagging is expanded to include the Random Forest model.
13. Adjusted R^2 quantifies the percentage of variation that can be accounted for by simply the independent variables that have a significant impact on the explanation of the dependent variable. If you include independent variables that have no bearing on predicting the dependent variable, you will be penalised. Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

$$R'^2 = 1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t}$$

    In the above equation, dft is the degrees of freedom n– 1 of the estimates of the population variance of the dependent variable, and dfe is the degrees of freedom n – p – 1 of the estimates of the underlying population error variance.

14. What is the difference between standardisation and normalisation?

    When the data does not adhere to the criteria of the Gaussian Distribution, normalisation is a suitable technique. It can be applied to algorithms like K-Nearest

Neighbours and neural networks that don't presume data distribution. In contrast, standardisation is advantageous when the dataset has a Gaussian distribution. Unlike normalisation, standardisation has no bounding range, therefore it is unaffected by the dataset's outliers. Depending on the issue and the machine learning technique, normalisation or standardisation may be applied. To employ normalisation or standardisation, there are no set rules. The two can be contrasted by fitting the normalised or standardised dataset into the model. It is generally desirable to transform the testing data after fitting the scaler on the training data. This would prevent data leakage when testing the model, and scaling target values is typically not necessary.

15. By using subsets of the data and an understanding of the bias/variance trade-off, cross validation is a type of model validation that seeks to advance the fundamental techniques of hold-out validation. The goal is to gain a better understanding of how the model will actually perform when applied outside of the data it was trained on.

    Advantages of Cross Validation:

    - Reduces Overfitting-We divide the dataset into several folds and train the algorithm on each of the folds in cross validation. By doing this, we avoid having our model overfit the training dataset. Thus, the model achieves generalisation capabilities in this way, which is a positive indicator of a resilient algorithm.

    Disadvantages of Cross Validation:

    - Increases training time-Cross Validation drastically increases the training time. Earlier model training was done only on one training set, but with Cross Validation training is performed on multiple training sets.

SQL

1. SELECT AVG(COUNT(‘shippedDate’))
   FROM orders
2. SELECT AVG(COUNT(‘quantityOrdered’))
   FROM orderdetails
3. SELECT ‘productName’,’MSRP’
   FROM products
   WHERE ‘MSRP’=(SELECT MIN(‘MSRP’) FROM products)
4. SELECT ‘productName’, ‘quantityInStock’
   FROM products
   WHERE ‘quantityInStock’=(SELECT MAX(‘quantityInStock’) FROM products)
5. SELECT ‘productName’,’productDescription’ FROM products
   INNER JOIN orderdetails
   WHERE products.’productCode’ = orderdetails.’productCode’
   GROUP BY products.’productCode’
   ORDER BY SUM(‘quantityOrdered’) DESC
   LIMIT 1;
6. SELECT ‘customerName’ FROM customer
   INNER JOIN payments
   WHERE customer.’customerNumber’=payments.’customerNumber’
   GROUP BY payments.’customerNumber’
   ORDER BY SUM(‘amount’) DESC
   LIMIT 1;
7. SELECT ‘customerName’,’customerNumber’ FROM customer
   WHERE ‘city’=”Melbourne”;
8. SELECT ‘customerName’
   FROM customers
   WHERE ‘customerName’ LIKE ‘N%’
9. SELECT ‘customerName’
   FROM customers
   WHERE ‘customerNumber’ LIKE ‘7%’

AND ‘city’=”LasVegas”;

10. SELECT ‘customerName’
    FROM customers
    WHERE ‘creditLimit’<1000
    AND ‘city’=’LasVegas’ OR ‘city’=’ Nantes’ OR ‘city’=’Stavern’
11. SELECT ‘orderNumber’
    FROM ‘orderdetails’
    WHERE ‘quantityOrdered’<10;
12. SELECT orders.’orderNumber’
    FROM orders
    LEFT JOIN customers
    ON customers.’customerNumber’ = orders.’customerNumber’
    WHERE customers.’customerName’ LIKE 'N%'
13. SELECT customers.’customerName’
    FROM customers
    LEFT JOIN orders
    ON customers.’customerNumber’=orders.’customerNumber’
    WHERE ‘status’=’Disputed’;
14. SELECT customers.’customerName’
    FROM customers
    RIGHT JOIN payments
    ON customers.’customerNumber’=payments.’customerNumber’
    WHERE ‘checkNumber’ LIKE ‘H%’
    AND ‘paymentDate’=’2004-10-19’;
15. SELECT ‘checkNumber’
    FROM payments
    WHERE ‘amount’>1000;

Statistics

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

   According to the central limit theory, the mean of all samples from a population with a finite level of variance will be roughly equal to the population mean if you take a sufficiently large sample size from that population. The normalcy assumption and the accuracy of the estimates are the two key factors that make the central limit theorem important in statistics.

   - It has important ramifications that sampling distributions can resemble a normal distribution. The normalcy assumption in statistics is essential for parametric hypothesis testing of the mean, like the t-test. However, the central limit theorem comes into play and generates sampling distributions that resemble a normal distribution when the sample size is large enough. Given that the sample size is high enough, this fact enables us to utilise these hypothesis tests even when the data are not regularly distributed.
   - As sample numbers grow, the sampling distributions of the mean huddle more together around the population mean. When estimating the population mean from a sample, the central limit theorem's property becomes important. The sample mean is more likely to be close to the actual population mean when the sample size is bigger. The estimate is therefore more accurate. On the other hand, for smaller sample sizes, the sampling distributions of the mean are substantially wider. It is common for sample means to deviate further from the true population mean when sample sizes are small. As a result, less accurate estimates are produced.

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

   In statistical analysis, sampling is the procedure by which researchers select a specific number of observations from a larger population. The sampling strategy is determined by the sort of analysis being done. There are two primary types of sampling methods that are used in research- Probability Sampling and Non-Probability Sampling

   - Probability sampling relies heavily on random selection to provide for the rigorous statistical analysis of group results.
   - Non-probability sampling entails non-random selection based on practicality or other factors, making it simple to gather data.

3. What is the difference between type1 and type II error?

   A sample may not always be representative of the population due to pure chance. As a result, the sample's findings do not accurately represent the population's situation, and the random mistake causes an incorrect conclusion to be drawn. When a null hypothesis that is actually true in the population is rejected, it is referred to as a type I error (false-positive); when one that is genuinely untrue in the population is not rejected, it is referred to as a type II error (false-negative). The investigator can lessen the chance of type I and type II errors by increasing the sample size, even though they cannot be completely prevented. The less likely it is that the sample will significantly diverge from the population, the larger it is.

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

   An example of a continuous probability distribution is the normal distribution, in which the majority of data points cluster in the middle of the range while the remaining ones taper off symmetrically toward either extreme. The distribution's mean is another name for the centre of the range. A Gaussian distribution is another name for the normal distribution. Because it is symmetric around the mean, it shows that values close to the mean happen more frequently than those distant from the mean.

   A bell curve represents a normal distribution graphically due to its flared appearance. Depending on how the population's values are distributed, the exact shape may change. The total number of data points that make up the distribution is known as the population. A normal distribution bell curve is always symmetrical around the mean, regardless of its precise shape. A symmetrical distribution has two mirror images on either side of a vertical dividing line that is drawn across the maximum/mean value, with half of the population being less than the mean and half being more. The converse, i.e., that all symmetrical distributions are normal, is not necessarily true. The mean, mode, and median of a bell curve are all the same, and the peak is always in the middle.

5. What is correlation and covariance in statistics?

   Both statistics and regression analysis employ the phrases covariance and correlation.

   A change in one variable reflects a change in the other, which is referred to statistically as covariance. Covariance describes a systematic link between two

random variables. A negative number for the covariance value indicates a negative association, whereas a positive value indicates a positive relationship. The covariance value can vary from -infinity to +infinity. The relationship is more dependent the higher this value is. A positive figure for positive covariance indicates a direct relationship. An inverse link between the two variables is indicated by a negative value, which signifies negative covariance. For identifying the type of link, covariance is excellent.

In statistics, correlation is a metric that assesses how closely two or more random variables follow one another. The variables are said to be correlated when, during the study of two variables, an analogous movement of one variable reciprocates the movement of the other variable in some manner. When two variables move in the same direction, there is a positive correlation. Variables are said to be negatively linked when they move in the opposite way.

6. Differentiate between univariate, Bi-variate and multivariate analysis

Univariate data consists of only one variable. Since there is only one variable that varies, univariate data analysis is the most straightforward type of analysis. The analysis's primary goal is to explain the data and identify any patterns in it; it does not deal with causes or correlations.

Two distinct variables are present in bivariate data. The analysis of this kind of data focuses on linkages and causes, and it seeks to understand the causal connection between the two variables. The summertime temperature and ice cream sales are two examples of bivariate data.

Multivariate data refers to data that has three or more variables. For instance, if an online marketer wanted to compare the popularity of four ads, they could analyse the click rates for men and women and then look at the correlations between the variables. It is comparable to bivariate but has more dependent variables than that. The methods used to analyse this data depend on the objectives to be met. Regression analysis, path analysis, factor analysis, and multivariate analysis of variance are a few of the methodologies.

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

   Sensitivity is commonly referred to as the true positive rate, it is the likelihood that a test result will actually be positive. Formula used to calculate sensitivity is given below,

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

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

   In statistics, the process of hypothesis testing involves putting an analyst's presumption about a population parameter to the test. The type of data used and the purpose of the study will determine the methodology the analyst uses. Using sample data, hypothesis testing is done to determine whether a claim is plausible. These data could originate from a broader population or a process that creates data. A random sample of the population being studied is measured and examined by statistical analysts in order to test a hypothesis. To test the null hypothesis (H0) and the alternative hypothesis (H1), all analysts employ a random population sample. The null hypothesis typically states that two population parameters are identical; for instance, it can claim that the population mean return is equal to zero. In essence, the alternative hypothesis is the opposite of the null hypothesis (e.g., the population mean return is not equal to zero). Since they contradict one other, only one of them can be true. One of the two possibilities will, however, always be accurate. For two-tailed test the value of null hypothesis

   Due to the ability to test for effects in both directions, two-tailed hypothesis tests are often referred to as nondirectional and two-sided tests. A two-tailed test divides the significance threshold between the distribution's two tails. Sample data are sufficiently incompatible with the null hypothesis when a test statistic is in either of the crucial regions, allowing it to be rejected for the population. The general null and alternate hypotheses for a two-tailed test are as follows: Alternative: The effect does not equal zero; null: The effect equals zero.

9. What is quantitative data and qualitative data?

   Quantitative Data- Numbers are used to express quantitative data, which are measures of values or counts. Numerical variables are the subject of quantitative data (e.g. how many; how much; or how often).

   Qualitative Data- Qualitative data can be represented by a name, symbol, or number code and are measures of "types." Data concerning categorical variables make up qualitative data (e.g. what type).

10. How to calculate range and interquartile range?

    The smallest observed value is subtracted from the biggest observed value to determine the range. The data points between the distribution's two extremes are not taken into account by the range; just these two values are. Because of its sensitivity to extreme values, it is frequently employed as a supplement to other measures of dispersion rather than as the single measure.

    When values are ranked from lowest to highest, the IQR designates the middle 50% of those values. The median (middle value) of the lower and upper half of the data is calculated to obtain the interquartile range (IQR). Quartile 1 (Q1) and Quartile 3 are these values (Q3). The IQR represents the variation between Q3 and Q1.

11. What do you understand by bell curve distribution?

    The normal distribution, sometimes referred to as the bell curve, is a typical sort of distribution for a variable. The graph that is used to represent a normal distribution has a symmetrical bell-shaped curve, hence the name "bell curve”. All other potential occurrences are symmetrically distributed around the mean, resulting in a downward-sloping curve on either side of the peak, while the highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and median in this case). The standard deviation of the bell curve serves as a measure of its width.

12. Mention one method to find outliers.

    Statistical outlier detection which involves applying statistical test to identify extreme values.

13. What is p-value in hypothesis testing?

    The p value is a number, obtained from a statistical test, that describes how likely a certain set of observations are determined if the null hypothesis were true. In order

to determine whether to reject the null hypothesis, P values are utilised in hypothesis testing. Smaller the p value, more likely to reject the null hypothesis.

14. What is Binomial Probability formula?

    In an experiment with two possible outcomes, the likelihood of exactly x successes on n repeated trials is known as the binomial probability. If the probability of success on an individual trial is p, then the binomial probability is,

$$nCx.p^x.(1-p)^{n-x}$$

    Here nCx indicates the number of different combinations of x objects selected from a set of n objects.

15. Explain ANOVA and its applications.

    ANOVA is a statistical analysis technique that divides systematic components from random factors to account for the observed aggregate variability within a data set. The presented data set is statistically affected by the systematic factors but not by the random ones. The ANOVA test is used by analysts to evaluate the impact of independent factors on the dependent variable in a regression analysis.

    Applications-

    - For instance, a researcher might administer tests to students from various universities to determine whether any of the colleges regularly produces better performers than the others. To determine whether product creation technique is more cost-effective than the other in a corporate setting, an R&D researcher may examine two different methods.
    - ANOVA test types are determined by a variety of variables. It is used when experimental data is required. If statistical software is unavailable, analysis of variance is used, requiring manual computation of ANOVA. It is easy to use and works best with little samples. The sample sizes must be the same for all possible factor level combinations in many experimental designs.
    - When examining three or more variables, an ANOVA is useful. It resembles numerous two-sample t-tests. But it produces fewer type I errors and is suitable for a variety of problems. ANOVA includes dispersing the variation among many sources and groups differences by comparing the means of each

group. It is used with test groups, subjects, as well as between and within groups.