MACHINE LEARNING

ASSIGNMENT SET-4

Question No	Answers
1)	C) Between -1 and 1.
2)	D) Ridge Regularisation.
3)	C) Hyperplane.
4)	A) Logistic Regression.
5)	A) 2.205 X old coefficient of ‘X’
6)	B) Increase.
7)	B) Random forests explain more variance in data then decision trees.
8)	B) Principal components are calculated using unsupervised learning techniques. C) Principal components are linear combinations of linear variables.
9)	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 bank on the basis of previous years data loan accounts. C) identifying spam or ham emails. D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.
10)	A) max_depth B) max_features D) Min_samples_leaf

Q11 Answer: -

Outliers are datapoints in dataset in which are abnormal observations amongst the normal observations and can lead to weird accuracy scores which can skew measurements as the results do not present the actual results. Outliers are an observation that appears far away from an overall pattern in a sample. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimate poorer results.

What outliers are actually: -

I. Data entry errors (human errors).
II. Measurement errors (instrument errors).

III. Experimental errors (execution errors).
IV. Intentional (dummy outliers made to test detection methods).
V. Data processing errors (Data manipulation error).
VI. Sampling errors (extracting or mixing data from wrong or various sources).

We can use the IQR method of identifying outliers to set up a fence outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence ports that we compare each observation to. Any observation that are more than 1.5IQR below Q1 or more than 1.5 IQR above Q3 are consider outliers.

IQR = Q3 -Q1

Q1 is the first quartile of the data, 25% of data lies between minimum and Q1.

Q3 is the third quartile of the data to say 75% of data lies between minimum and Q3.

Lower bound: - (Q1 – 1.5 * IQR)

Q12 Answer: -

The primary difference bagging and boosting is that bagging involve fitting many decisions trees on different samples of the dataset and averaging the predictions. Whereas boosting involves adding ensemble members sequentially to correct the predictions made by prior models and output a weighted average of the predictions.

Q13 Answer: -

Adjusted R squared or modified R^2 determines the extent of the variance of the dependent variable, which the independent variable can explain. The speciality of the modified R^2 is that it does not consider the impact of all independent variables but only those which impact the variation of the dependent variable. Therefore, the value of the modified R^2 can also be negative, though it is not always negative.

$$R^2_{adj} = 1 - [\frac{(1-R^2)(n-1)}{n-k-1}]$$

Here, k is the no. of regressors and n is the sample size. if the newly added variable is good enough to improve model’s performance, then it will overwhelm the decrease due to k. Otherwise, an increase in k will decrease adjusted r-square value.

Q14 Answer: -

Sr. No.	Normalisation	Standardisation
1)	It is feature scaling method to bring the data into common range such as (0,1) & (-1,1) etc.	It is also feature scaling method bring the data with mean 0 & and unit variance.
2)	Scikit learn provides MinMaxScaler, MaxAbsScaler and RobustScaler methods for Normalisation.	Scikit learn provides StandardScaler for standardization.
3)	MinMaxScaler & MaxAbsScaler are sensitive to outliers whereas RobustScaler is more robust to outliers.	Standardisation is less sensitive to outliers.
4)	It is useful when we do not know about the distribution of features and there are no or little outliers.	Useful when we know features are normally distributed.

Q15 Answer: -

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

Advantages of Cross Validation

- Reduces Overfitting: - In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantages of Cross Validation: -

- Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.

STATISTICS

ASSIGNMENT SET-4

Q1 Answer: -

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 mean is asymptotically normal. We can calculate the mean of the sample means for the random samples we choose from the population:

$\mu X = \mu$

As well as the standard deviation of sample means:

$\sigma X = \sigma n$

According to the central limit theorem, the form of the sampling distribution will approach normalcy as the sample size is sufficiently large (usually $n>30$). regardless of the population distribution.

Importance of Central Limit Theorem:

This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

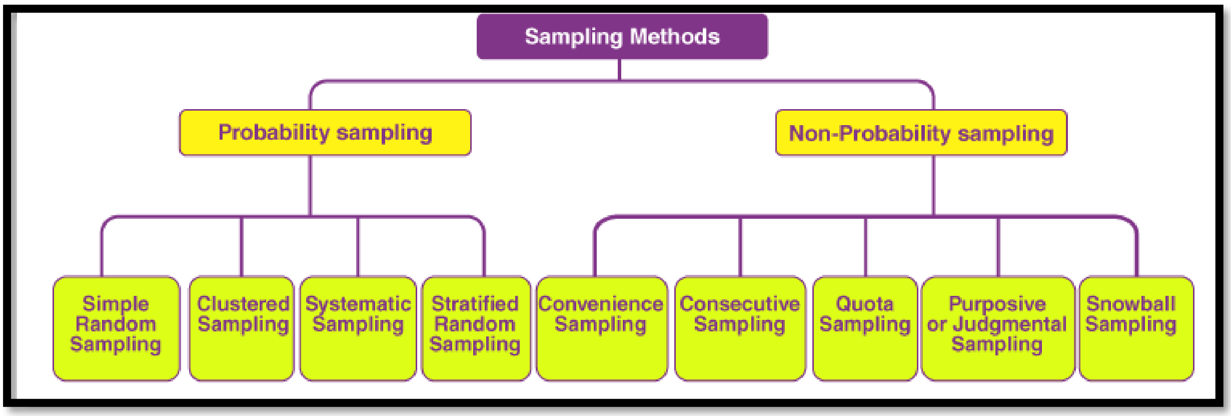
Q2 Answer: -

In Statistics, the sampling method or sampling technique is the process of studying the population by gathering information and analysing that data. It is the basis of the data where the sample space is enormous. There are several different sampling techniques available, and they can be subdivided into two groups. All these methods of sampling may involve specifically targeting hard or approach to reach groups.

Types of Sampling Method

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are as below,

- **Probability Sampling: -** The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population.
- **Non-probability Sampling: -** The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

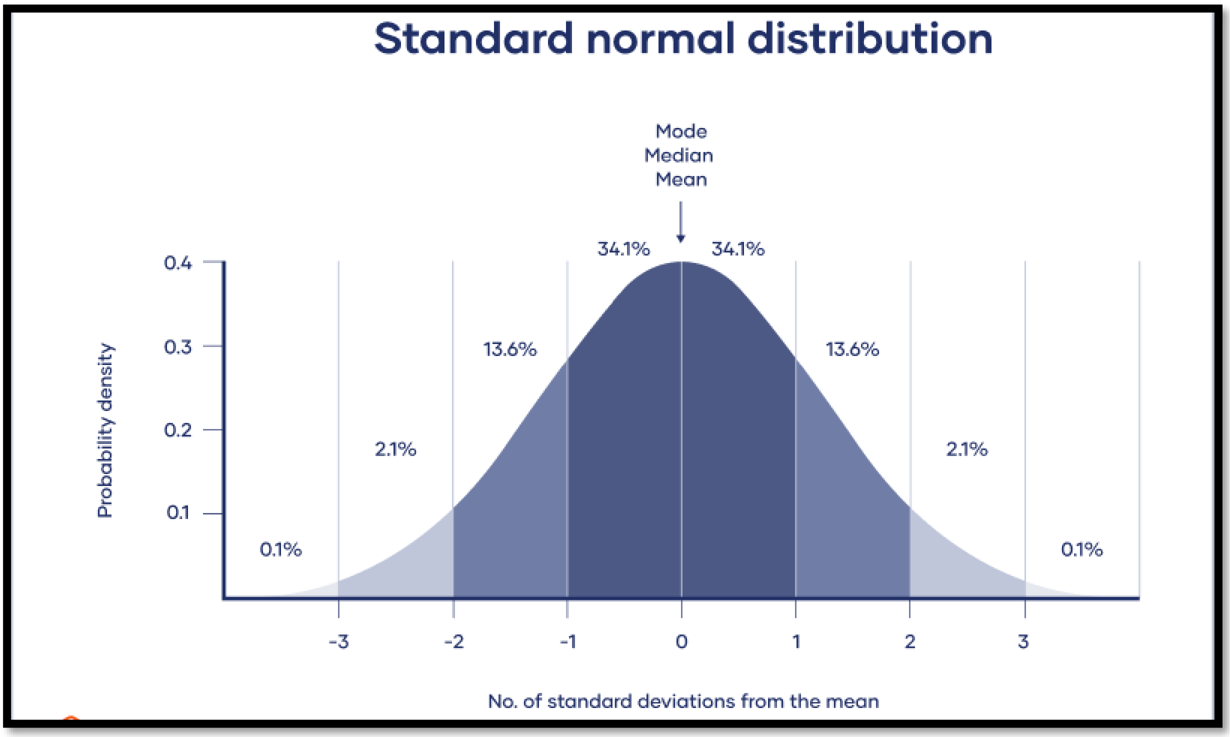


Q3 Answer: -

1. There is a rejection of reality by the researcher in type one error, whereas the researcher accepts the false reality in type two error.
2. In type 1 error, the null hypothesis, in reality, is true, whereas in type 2 error, the null hypothesis, in reality, is false.
3. The probability of type 1 error taking place is alpha, whereas that of type 2 error is beta.
4. Many refer to type 1 error as an error of the first kind and type 2 error as an error of the second kind.
5. Type 2 error can be reduced to a certain extent by decreasing the level of alpha, whereas type 2 error can be reduced by increasing the alpha level.

Q4 Answer: -

A normal distribution or Gaussian distribution refers to a probability distribution where the values of a random variable are distributed symmetrically. These values are equally distributed on the left and the right side of the central tendency. Thus, a bell-shaped curve is formed. The mean, median and mode are exactly the same. The distribution is symmetric about the mean—half the values fall below the mean and half above the mean. The distribution can be described by two values: the mean and the standard deviation.



Q5 Answer: -

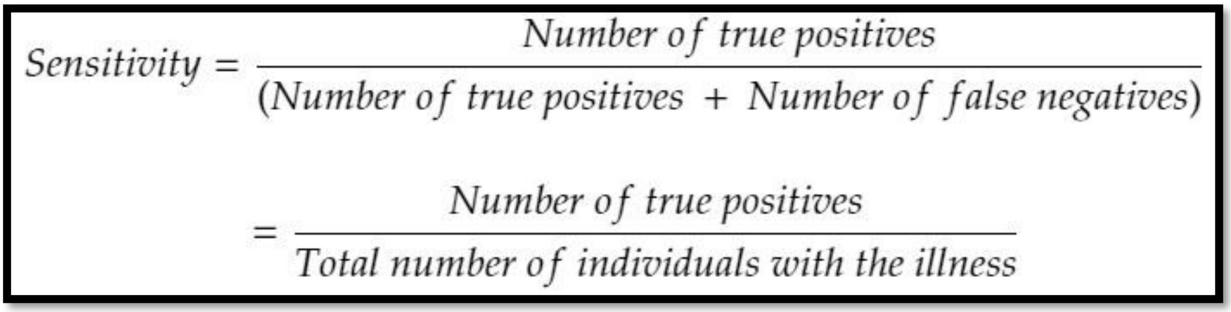
- **Correlation: -** Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It is a common tool for describing simple relationships without making a statement about cause and effect.
- **Covariance: -** In mathematics and statistics, covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables. Unlike the correlation coefficient, covariance is measured in units. The units are computed by multiplying the units of the two variables.

Q6 Answer: -

- **Univariate Analysis: -** The examination of the distribution of cases on only one variable at a time. (Example: - Weight of the college student).
- **Bivariate analysis: -** The examination of two variables simultaneously (example: - relation between gender and weight of college students.)
- **Multivariate Analysis: -** The examination of more than two variables simultaneously (example: - The relationship between gender, race, and weight of college students.)

Q7 Answer: -

Sensitivity (true positive rate) refers to the probability of a positive test, conditioned on truly being positive. It is calculated by using below formula,



$$Sensitivity = \frac{Number\ of\ true\ positives}{(Number\ of\ true\ positives\ +\ Number\ of\ false\ negatives)}$$

$$= \frac{Number\ of\ true\ positives}{Total\ number\ of\ individuals\ with\ the\ illness}$$

Q8 Answer: -

Hypothesis testing is a statistical interpretation that examines a sample to determine whether the results stand true for the population. The test allows two explanations for the data the null hypothesis or the alternative hypothesis.

- Null hypothesis (H0): - If the sample mean matches the population mean, the null hypothesis is proven true.
- Alternative Hypothesis (H1): - Alternatively, if the sample mean is not equal to the population mean, the alternate hypothesis is accepted.

A two-sample t-test always uses the following null hypothesis,

- H0: μ1 = μ2 (the two-population means are equal)

The alternative hypothesis can be either two-tailed, left-tailed, or right-tailed,

- H1 (two-tailed): μ1 ≠ μ2 (the two-population means are not equal)

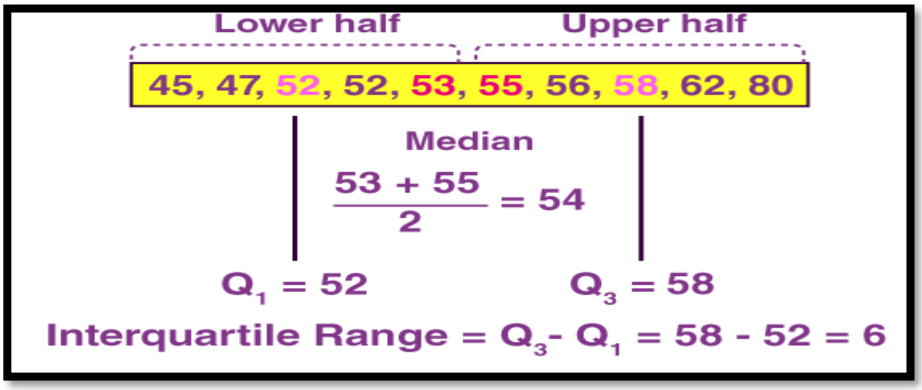
Q9 Answer: -

- **Qualitative Data: -** Qualitative data is a set of information which cannot be measured using numbers. It generally consists of words, subjective narratives. Result of a qualitative data analysis can come in form of highlighting key words, extracting information and concepts elaboration. For example, a study on parents perception about the current education system for their kids.
- **Quantitative Data: -** Quantitative data is a set of numbers collected from a group of people and involves statistical analysis. For example, if you conduct a satisfaction survey from participants and ask them to rate their experience on a scale of 1 to 5. You can collect the ratings and being numerical in nature, you will use statistical techniques to draw conclusions about participants satisfaction.

Q10 Answer: -

- **Range: -** To find the range in statistics, we need to arrange the given values or set of data or set of observations in ascending order. Thus, the range can be defined as the difference between the highest observation and lowest observation.
  (For example, if the given data set is {2,5,8,10,3}, then the range will be 10 – 2 = 8.)
- **Interquartile Range: -** The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.
  (Interquartile range = Upper Quartile – Lower Quartile = Q3 – Q1)

Example: -

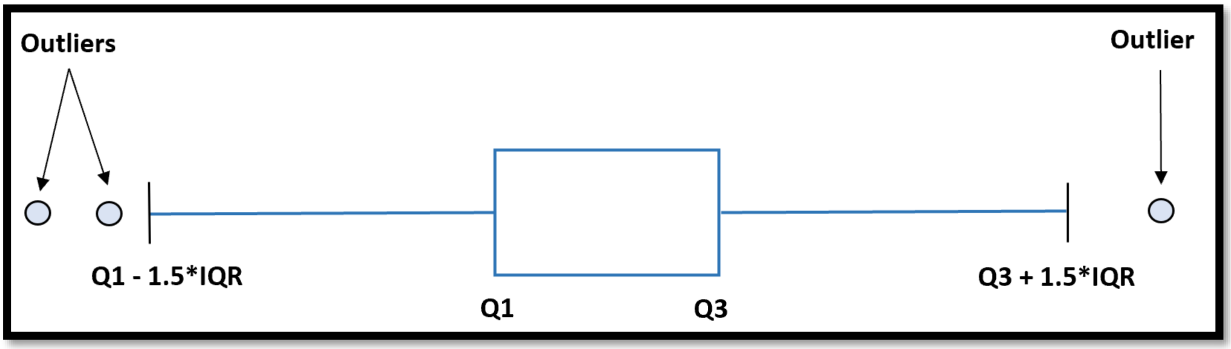


Q11 Answer: -

A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean. The width of a bell curve is determined by the standard deviation 68% of the data points are within one standard deviation of the mean, 95% of the data are within two standard deviations, and 99.7% of the data points are within three standard deviations of the mean.

Q12 Answer: -

An outlier is an observation that lies abnormally far away from other values in a dataset. Outliers can be problematic because they can affect the results of an analysis. One common way to find outliers in a dataset is to use the interquartile range. The interquartile range, often abbreviated IQR, is the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset. It measures the spread of the middle 50% of values. One popular method is to declare an observation to be an outlier if it has a value 1.5 times greater than the IQR or 1.5 times less than the IQR.



Q13 Answer: -

In statistical hypothesis testing, P-Value or probability value can be defined as the measure of the probability that a real-valued test statistic is at least as extreme as the value actually obtained. P-value shows how likely it is that your set of observations could have occurred under the null hypothesis. When you run the hypothesis test, if you get,

A small p value (<=0.05), you should reject the null hypothesis.

A large p value (>0.05), you should not reject the null hypothesis.

Q14 Answer: -

The Binomial Probability distribution of exactly x successes from n number of trials is given by the below formula,

P (X) = nCx px qn – x

Where,

n = Total number of trials

x = Total number of successful trials

p = probability of success in a single trial

q = probability of failure in a single trial = 1-p

Q15 Answer: -

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables. If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

Application 1: - A grocery chain wants to know if three different types of advertisements affect mean sales differently. They use each type of advertisement at 10 different stores for one month and measure total sales for each store at the end of the month. to sees if there is a statistically significant difference in mean sales between these three types of advertisements, researchers can conduct a one-way ANOVA, using “type of advertisement” as the factor and “sales” as the response variable. If the overall p-value of the ANOVA is lower than our significance level, then we can conclude that there is a statistically significant difference in mean sales between the three types of advertisements. We can then conduct post hoc tests to determine exactly which types of advertisements lead to significantly different results.

Application 2: - Biologists want to know how different levels of sunlight exposure (no sunlight, low sunlight, medium sunlight, high sunlight) and watering frequency (daily, weekly) impact the growth of a certain plant. In this case, two factors are involved (level of sunlight exposure and water frequency), so they will conduct a two-way ANOVA to see if either factor significantly impacts plant growth and whether or not the two factors are related to each other. The results of the ANOVA will tell us whether each individual factor has a significant effect on plant growth. Using this information, the biologists can better understand which level of sunlight exposure and/or watering frequency leads to optimal growth.