**STATISTICS WORKSHEET-4**

Q1to Q15 are descriptive types. Answer in brief.

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

**Answer**. Central limit theorem is a statistical theory which states that if we take large random samples with replacement from a population with mean μ and standard deviation σ , the distribution of all samples mean from the population would be approximately normally distributed. This will hold true for a skewed or a normal population, provided the sample size is sufficiently large (n>=30). Therefore, as the sample size increases the sample mean and standard deviation will be equal to the population mean and standard deviation.

Central limit theorem is important as it allows us to assume the sampling distribution of the mean will be normal in most cases. This means, we can easily use the statistical techniques that assume a normal distribution.

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

**Answer**. Sampling is a method through which we can get information about the population based on the statistics from a subset of the population without actually investing the whole population. There are two methods of sampling:

1. Probability sampling- In this, every element of the population has a fair chance of getting selected. Probability sampling gives us the best chance to select the sample that is a true representation of the population.

2. Non-probability sampling- In this all the element do not have an equal chance of getting selected in the sample of the population. There is significant risk of ending up with a non-representative sample that would fail to produce generalized results.

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

**Answer**. Type 1 error also referred to as false positives occurs when we reject a true null-hypothesis. Type 1 error is denoted by alpha.

Type 2 error also called true negatives occurs when we fail to reject a false null-hypothesis. Type 2 error is denoted by beta.

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

**Answer**. Normal distribution, or Gaussian distribution, is a probability distribution that is symmetric about the mean i.e., most values cluster in the middle of the range rest tapers symmetrically towards either extreme. Probability density of Normal distribution will appear as a bell curve in graphical representation.

5. What is correlation and covariance in statistics?

Answer. Correlation in statistics is the association between two quantitative variables. This association is linear i.e., one variable increase or decreases for a unit increase/decrease in another variable. Correlation can be either a positive correlation or a negative correlation.

Covariance is the measure of directional relationship between two random variables. It is the measure of how two random variables would change when they are compared to each other.

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

**Answer**. Univariate analysis is the basic form of statistical data analysis where the data to be analysed contains only one variable. Since it is a single variable, it doesn’t deal with cause or relationship. The main purpose of this analysis is to describe the data and find patterns that exists within it.

Bivariate analysis is the analysis of two variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among two variables. Bivariate analysis is used to find if there is any relationship between two independent variables.

Multivariate analysis is the analysis of three or more variables. When the data contains three or more variables it is categorized under Multivariate.

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

**Answer**. Sensitivity analysis is the technique to determine the impact of independent variable over the dependant variable under certain assumptions. Sensitivity is calculated by dividing percent change in output by the percent change in the input.

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

**Answer**. Hypothesis testing is a statistical method to select samples from a group or population with an intent of making determinations about the expected behaviour of the group. Hypothesis testing is basically an assumption that we make about a population parameter and test whether it is true or false.

H0 also called null hypothesis is the hypothesis that is to be tested. H0 is the statistical hypothesis that states that there is no difference between a parameter and a specific value. The null hypothesis is assumed to be true unless we have some strong evidence to the contrary.

H1 also called alternate hypothesis is the complement of alternative hypothesis. H1 is the statistical hypothesis that states there is a difference between a parameter and a specific value. It is assumed to be true when the null hypothesis is false.

Null hypothesis in two-tailed test- The effect equals zero.

Alternative hypothesis in two-tailed test- The effect does not equal zero.

9. What is quantitative data and qualitative data?

**Answer**. Quantitative data- Quantitative data is information about quantities. Quantitative data can be counted, measured and expressed using numbers.

Qualitative data- Qualitative data is a set of information which cannot be measured using numbers. It generally consists of words, subjective narrative. Qualitative data is descriptive, non-numerical, and regards phenomenon which can be observed but not measured, such as language.

10. How to calculate range and interquartile range?

**Answer**. IQR, Interquartile range is the best measure of variability for skewed distributions or datasets with outliers.

IQR= Q3-Q1 where

Q1 is first quartile, median of lower half of the dataset.

Q3 the third quartile, is the median of upper half of the dataset.

Range is the spread of the data from the lowest to the highest value in the distribution.

Range = highest-lowest

11. What do you understand by bell curve distribution?

**Answer**. The bell curve is informal name of a graph that describes a mathematical concept called Normal distribution also referred to as Gaussian distribution. The term obtained its name due to the bell-shaped curve of the normal probability distribution graph. A bell curve signifies that the data is symmetrical. The top of the curve shows the mean, median, mode of the data collected. It’s standard deviation depicts the bell curve’s relative width around the mean.

A good example of a bell curve is the roll of two die in which the distribution is centred around number 7 and the probability decreases as we move away from the centre.

12. Mention one method to find outliers.

**Answer**. We can find outliers using IQR (inter quartile range), as IQR contains middle bulk of the data so outliers can be easily traced once we know the IQR. An outlier is a point that lies over 1.5IQR below the first quartile(Q1) or 1.5IQR above the third quartile(Q3) in a dataset.

13. What is p-value in hypothesis testing?

**Answer**. p-value is the level of marginal significance within a statistical hypothesis testing, representing the probability of the occurrence of a given event. A p-value is used in hypothesis testing to help support or reject the null hypothesis. P-value is evidence against the null-hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null-hypothesis.

14. What is the Binomial Probability Formula?

**Answer**. Binomial probability refers to the probability of exactly x successes on n repeated trials of an experiment which has two possible outcomes (called a binomial experiment). If the probability of successes of an individual trial is p, the binomial probability is given by nCx.px.(1-p)n-x .

Here nCx is the number of different combinations of x objects selected from a set of n objects. As p is probability of success of a trial (1-p) is the probability of failure of a trial.

15. Explain ANOVA and its applications.

**Answer**. A statistical technique which is used to check if the means of two or more groups are significantly different from one another. compares samples on the basis of their mean is ANOVA or analysis of variance.

ANOVA is helpful in comparing three or more variables. For example, ANOVA can be used to compare three medicines working upon patients with same disease.

**MACHINE LEARNING ASSIGNMENT - 4**

1 In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:

A) between 0 and 1

B) greater than -1

C) between -1 and 1

D) between 0 and -1

**Answer**. C) between -1 and 1

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

A) Lasso Regularisation

B) PCA

C) Recursive feature elimination

D) Ridge Regularisation

**Answer**. D) Ridge regularization

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

A) linear B) Radial Basis Function C) hyperplane D) polynomial

**Answer**. C) hyperplane

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

A) Logistic Regression

B) Naïve Bayes Classifier

C) Decision Tree Classifier

D) Support Vector Classifier

**Answer**. D) Support Vector Classifier

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? (1 kilogram = 2.205 pounds)

A) 2.205 × old coefficient of ‘X’

B) same as old coefficient of ‘X’

C) old coefficient of ‘X’ ÷ 2.205

D) Cannot be determined

**Answer**. C) old co-efficient of x/2.205

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

A) remains same B) increases C) decreases D) none of the above

**Answer**. C) decreases

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

A) Random Forests reduce overfitting

B) Random Forests explains more variance in data then decision trees

C) Random Forests are easy to interpret

D) Random Forests provide a reliable feature importance estimate

**Answer**. A) Random forests reduce overfitting

In Q8 to Q10, more than one options are correct, Choose all the correct options:

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

A) Principal Components are calculated using supervised learning techniques

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

D) All of the above

Answer. 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.

**Answer**. B) identifying loan defaulter in a bank on the basis of their previous year’s 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 B) max\_features C) n\_estimators D) min\_samples\_leaf

**Answer**. A) max\_depth, B) max\_features, D) min\_samples\_leaf

Q11 to Q15 are subjective answer type questions, Answer them briefly.

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

**Answer**. An outlier is an observation in the data that is unlike the other observations. Simply, an outlier is an odd one out or one that is different from the crowd.

IQR shows the spread of data about the median. IQR can be used to detect the outliers in the following ways:

1. Calculate the IQR for the data IQR=Q3-Q1.

2. Multiply IQR by 1.5.

3. Add 1.5\*IQR to third quartile Q3, any number greater than this value is a suspected outlier.

4. Subtract 1.5\*IQR by Q! the first quartile, any number lesser than this value is a suspected outlier.

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

**Answer**.

|  |  |
| --- | --- |
| BAGGING ALGORITHM | BOOSTING ALGORITHM |
| 1. Bagging is an ensemble technique. | 1. boosting is an ensemble technique. |
| 2. It is a way of combining the predictions that belong to the same type. | 2. A way of combining predictions that belong to the different types. |
| 3. Aims to decrease variance, not bias. | 3. Aims to decrease bias, not variance. |
| 4. Each model is built independently. | 4. New models are built under influence of previously built models. |
| 5. Different training data subsets are randomly drawn with replacement from the training dataset. | 5. New subset contains the misclassified elements by the previous classifier. |
| 6. Random forest is a bagging algorithm. | 6. Gradient boosting is a boosting algorithm. |

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

**Answer**. R2- R squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. R-squared is a goodness of fit measure for linear regression models.

R2= Variance explained by the model/ total variance

R2 is between 0 to 100%.

0% means the model cannot explain the variance in the response variable around its mean.

100% represents a model that explains all the variation in the response variable around its mean.

14. What is the difference between standardisation and normalisation?

**Answer**. Normalization- It is a scaling technique in which values of the data are re-scaled in the range 0-1. It is also called Min max scaling.

Standardization- Standardization is also a scaling technique where the values are centred around the mean with a unit standard deviation.

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

**Answer**. Cross validation is a resampling technique for evaluating a machine learning model and helps to make our model sure about its efficiency and accuracy on the unseen data. Cross validation tends to have lower bias than other methods used to count model’s efficiency scores.

Advantage: 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.

Disadvantage: Increases training time- Cross validation drastically increases the training time. In cross validation you have to train the model on multiple training sets.

**SQL WORKSHEET 4**

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

**Answer**. select avg(cnt) from(select count(shippedDate) as cnt from orders group by shippedDate) as average;

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

**Answer**. select avg(totalorder) from(select count(orderNumber) as totalorder from orders group by orderDate) as average;

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

**Answer**. 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.

**Answer**. select productName from products order by stockQuantity desc limit 1;

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

Answer. select count(name) from products inner join orderdetails on products.code=orderdetails.code group by name limit 1;

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

**Answer**. select customeName from customers inner join payments on customers.customerNumber=payments.customerNumber group by customerName order by amount desc limit 1;

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

**Answer**. select customerNumber,customerName from customers where city=’Melbourne’;

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

**Answer**. 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 ‘LasVegas’.

**Answer**. select customerName from customers where customerName like 'N%' 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”.

**Answer**. select customerName from customers where creditLimit<1000 and (city=’Las Vegas’ or city=’Nantes’ or city=’Stavern’);

11. Write a SQL query to show all the orderNumber in which quantity ordered 1000.

**Answer**. select orderNumber from orderdetails where quantityOrdered=1000;

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

**Answer**. select orderNumber from orders inner join customers on orders.customerNumber=customers.customerNumber where customerName like ‘N%’;

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

**Answer**. select customerName from customers inner join orders on customers.customerNumber=orders.customerNumber where status=’Disputed’;

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

**Answer**. select customerName from customers inner join payments on customers.customerNumber=payments.customerNumber where checkNumber like ‘H%’ and paymentDate=’2004-10-19’;

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

**Answer**. select checkNumber from payments where amount >1000;