WORKSHEET SET-4

MACHINE LEARNING

1. C
2. D
3. C
4. A
5. B **6.** B
7. C
8. B & C
9. A & D
10. A & D

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. ... These points are often referred to as outliers.
The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

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 overfitting issues in a model. Boosting decreases bias, not variance.

13. What is adjusted R2 in linear regression. How is it calculated?
It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

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. 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.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. R^2adjusted=1-(1-R^2)(N-1)/(N-P-1)

WORKSHEET SET-4

14. What is the difference between standardisation and normalization?

Normalisation	***Standardisation***
• Minimum and maximum value of features are used for scaling	• Mean and standard deviation is used for scaling.
• It is used when features are of different scales	• It is used when we want to ensure zero mean and unit standard deviation.
• Scales values between [0, 1] or [-1, 1].	• It is not bounded to a certain range
• It is really affected by outliers	• It is much less affected by outliers.
• Scikit-Learn provides a transformer called MinMaxScaler for Normalization	• Scikit-Learn provides a transformer called StandardScaler for standardization.
• This transformation squishes the n dimensional data into an n-dimensional unit hypercube	• It translates the data to the mean vector of original data to the origin and squishes or expands.
• It is useful when we don't know about the distribution	• It is useful when the feature distribution is Normal or Gaussian.
• It is a often called as Scaling Normalization	• It is a often called as Z-Score Normalization.

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

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set. The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Advantages of Cross Validation:

1.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.
2.Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

1.Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.
2.Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.

WORKSHEET SET-4

SQL

1. Select AVG(orderNumber) from order Where ShippedDate =---’

2. Select AVG(orderNumber) from order Where orderDate = “----”

3. Select ProductName from products
Where MSRP=(select min(MSRP) from products

4. Select products.productName from products
Innerjoin products.productCode=orderdetails.productCode
Where ordersdetails.quantityOrdered=(select max(orderdetails.quantityordered)from orderdetails)

5. Select products.productName from products
Innerjoin ordersdetails ON products.productCode=orderdetails.productCode Where orderdetails.quantityordered = (select max(orderdetails.quantityordered)from orderdetails)

6. Select cutomers.CustomerName from customers
Innerjoin payments ON customers.customerNumber=payments.customerNumber
Where payments.amount=(select max(payments.amount) from payments

7. Select customerNumber,customerName from customers Where city=“ Melbourane city”

8. Select customerName from customers Where customerName Like “N%”

9. Select customerName from customers
Where (phone like “7%” AND city=“ Lasvegas”)

10. Select customerName from cutomers
Where (credit limit<1000 AND city IN (“Lasvegas”, “Nates”, “Steven”)

11. Select orders.orderNumber from orders
Innerjoin orderdetails ON orders.orderNumber=orderdetails.orderNumber Where orderdetails.quantityordered<10

12. Select orders.orderNumber from orders
Innerjoin customers ON orders.customerNumber=customer.customerNumber
Where customers.customerName like “N%”

13. Select customers.customerName from customers
Innerjoin orders ON orders.customerNumber=customer.customerNumber Where orders.status=“Disputed”

14. Select customers.customerName from customers
Innerjoin payments ON customers.customerNumber=payments.customerNumber

15. Select checkNumber from payments Where amount>1000

WORKSHEET SET-4

STATISTICS

1. **What is central limit theorem and why is it important?**
   The central limit theorem states that the sampling distribution of the mean approaches a normal distribution, as the sample size increases. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ .
2. **What is sampling? How many sampling methods do you know?**
   When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.

There are two types of sampling methods:

- Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.
- Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

3. **What is the difference between type1 and typeII error?**
   A **type I error** (false-positive) occurs if an investigator rejects a null hypothesis that is actually true **in the** population; a **type II error** (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false **in the** population.

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.
5. **What is correlation and covariance in statistics?**
   Covariance is a statistical tool that is used to determine the relationship between the movement 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.
   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.
6. **Differentiate between univariate ,Biavariate,and multivariate analysis.**

WORKSHEET SET-4

Univariate statistics summarize only one variable at a time.

Bivariate statistics compare two variables.
Multivariate statistics compare more than two variables.

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

   A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty.

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. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.
   H0:Null hypothysis
   H1:Alternative hypothysis
   In statistics, a two-tailed test is a method in which the critical area of a distribution is twosided and tests whether a sample is greater or less than a range of values. ... 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 can be counted, measured, and expressed using numbers. Qualitative data is descriptive and conceptual. Qualitative data can be categorized based on traits and characteristics.

10. **How to calculate range and interquartile range?**
    To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.
    The interquartile range formula is the first quartile subtracted from the third quartile:

    $$IQR = Q_3 – Q_1.$$

11. **What do you understand by bell curve distribution ?** The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape. A standard deviation is a measurement used to quantify the variability of data dispersion, in a set of given values around the mean.

12. **Mention one method to find outliers.**
    The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR.

13. **What is p-value in hypothesis testing?**

WORKSHEET SET-4

The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true – the definition of ‘extreme’ depends on how the hypothesis is being tested. P is also described in terms of rejecting H_0 when it is actually true, however, it is not a direct probability of this state.

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 it’s applications.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. Another measure to compare the samples is called a t-test.