Machine Learning

1. A
2. D
3. C
4. A
5. C
6. C
7. C
8. B , C
9. A ,B ,C, D
10. A , B, D
11. Outliers are data points which lie outside the normal range of data than most of the other data points.

Interquartile range is the range between the 75th and 25th percentile of data. Any point below Q1-1.5 IQR and above Q3+1.5IQR is considered as outlier.

1. In Boosting methods output of one estimator is fed as input to the other to solve the error received from previous estimator whereas in Bagging output of n-estimators is averaged to get final output.

Bagging reduces variance and not bias whereas Boosting reduces Bias not variance.

1. Adjusted-R2 is a method used to get the R2 score by eliminating the independent features which are non correlated (least correlated) with the target attribute.

It is calculated by using below formula:

R2 (adjusted)= (1-R2) (N-1)

1. \_\_\_\_\_\_\_\_\_\_\_\_

N-p-1

Where p= no. of independent features (which are not correlated with target)

N= total sample size

If ‘p’ gets increased then N-p-1 will be decresed and hence (N-1)/ (N-p-1) will be increased and therefore R2(adjusted ) will be decreased.

1. **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 (unit variance.
2. Cross validation is a concept of validating the prepared model on different combination of data to see it the model is not just specific to the data on which it was prepared.

The **disadvantage** of this **method** is that the training **algorithm** has to be rerun from scratch **k** times, which means it takes **k** times as much computation to make an evaluation

The advantage is it reduce overfitting since 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.

SQL

1. Select avg(ordercount) from ( Select count(orderNumber) as ordercount from orders group by shippedDate);
2. Select avg(ordercount) from ( Select count(orderNumber) as ordercount from orders group by orderDate);
3. Select productName from products where MSRP= (select min(MSRP) from products);
4. Select productName from products where quantityInStock= (select max(quantityInStock) from products);
5. Select productName from **(**Select productName , count(od.orderNumber) as ordercount from products p inner join orderdetails od on p.productCode=od.productCode group by od.productCode order by ordercount desc limit 1));
6. Select c.customerName from customers c inner join payments p on p.customerNumber=c.customerNumber where p.amount= (select max(amount) from payments);
7. Select customerNumber , cutomerName from customers where city=’Melbourne’;
8. Select customerName from customers where customerName like ‘N%’;
9. Select customerNames from customers where city=’Las Vegas’ and phone like ‘7%’;
10. Select customerNames from customers where creditLimt < 2000 and city in (‘Las Vegas’, ‘Nantes’, ‘Stavern’);
11. Select orderNumber from orderdetails where quantityOrdered< 20;
12. Select o.orderNumber from customers c inner join orders o on o.customerNumber=c.customerNumber where c.customerName like ‘N%’;
13. Select c.customerName from customers c inner join orders o on o.customerNumber=c.customerNumber where o.status = ‘Disputed’;
14. Select c.customerName from customers c inner join payments p on c.customerNumber=p.customerNumber where p.paymentDate=’2004-10-19’ and p.checkNumber like ‘H%’;
15. Select checkNumber from payments where amount >1000;

STATISTICS

1. Central Limit Theorem states that ‘mean of sample distribution of sample mean approaches normal distribution ‘ when sample size is large (especially when sample size>30). 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. This means that we can take advantage of statistical techniques that assume a normal distribution.
2. Sampling is the process of selecting random observations from population.

There are various methods of selecting samples from population some of them are:

1, simple random sampling – samples are selected randomly and any sample may be selected by chance

2, Systematic sampling: samples are selected at regular intervals from population.

3, Clustered sampling: subgroups of population are formed which are used as sampling unit . Any random cluster is considered for study.

4, Stratified sampling: Population is first divided into subgroups having sharing similar characteristics and then equal number of samples are taken from each sub group.

1. Type I error – When a true null hypothesis is rejected

Type II error – When we fail to reject a false null hypothesis

1. Normal distribution is the probability distribution which is symmetric about the mean and forms a bell shaped curve. This shows that the data near to the mean occurs more frequently than the data far from mean.
2. **Covariance** is a measure of how much two random variables vary together. **Correlation** is a statistical measure that indicates how strongly two variables are related.
3. **Univariate** Analysis- Analysis conducted on using single variable/attribute/

**Bivariate** Analysis- Analysis done by taking two variable/attributes at a time.

**Multivariate** Analysis- Analysis done by taking more than 2 variables .

1. Sensitivity means “out of total actual positive values , how many were predicted correctly”. It is also known as True positive rate or recall.

It is calculated using below formula:

Sensitivity= TruePositives / (TruePositives+FalseNegatives)

1. **Hypothesis testing** in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results or not.

**H0- Null hypothesis**

**H1- Alternate hypothesis**

In two tail test our H0 aims to claim that the there is no significant difference in the mean OR mean remains same. Whereas H1 aims to claim that the mean is not same i.e either it is significantly lower or significantly higher.

1. Quantitative data some is information about quantities (measurable quatity) example : length ,mass,temperature.

Qualitative data gives description about quality or characteristics. Exmaple: Name ,address of person.

1. Range simply equals to the difference of maximum and minimum data points. (Range=Max-Min)

Interquartile range is the difference between 75th and 25th percentile of data.

1. The bell curve distribution is nothing but the normal distribution which is symmetric about the mean and where data close to mean occurs more frequently than that of data far from mean.
2. Z score method can be used to find outliers .In this method normally the data points having z score between -3 and 3 are considered and rest all are considered as outliers.
3. The **p**-**value**, or probability **value**, **tells you** how likely are the chances that your data could have occurred under the null hypothesis.

OR p value is the probability that tells that the change in phenomena (as per the user case) is just by chance.

Now if p value is less than significance level we can conclude that there are significantly less chances that our data occur under null hypothesis OR the probability that change in phenomena is just by chance is significantly low which mean it is not just by chance and hence null hypothesis is rejected.

1. Binomial distribution is the probability distribution of getting success in N-number of Bernaulli trails.



Here n= number of trails

X= total number of successes

P= probability of success on individual trial

1. **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 mean of different samples.ANOVA can be used to prove/disprove a newly introduced medicine was effective or not.

In case of only two samples, t-test and ANOVA will give same result.