

MACHINE LEARNING ASSIGNMENT – 7

1. Which of the following in sk-learn library is used for hyper parameter tuning?

- A) GridSearchCV()
- B) RandomizedCV()
- C) K-fold Cross Validation

D) All of the above

2. In which of the below ensemble techniques trees are trained in parallel?

A) Random forest

- B) Adaboost
- C) Gradient Boosting
- D) All of the above

3. In machine learning, if in the below line of code:

```
sklearn.svm.SVC (C=1.0, kernel='rbf', degree=3)
```

we increasing the C hyper parameter, what will happen?

A) The regularization will increase

B) The regularization will decrease

- C) No effect on regularization
- D) kernel will be changed to linear

4. Check the below line of code and answer the following questions:

```
sklearn.tree.DecisionTreeClassifier(*criterion='gini',splitter='best',max_depth=None,  
min_samples_split=2)
```

Which of the following is true regarding max_depth hyper parameter?

- A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.
- B) It denotes the number of children a node can have.

C) both A & B

D) None of the above

5. Which of the following is true regarding Random Forests?

A) It's an ensemble of weak learners.

B) The component trees are trained in series

C) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees.

D)None of the above

6. What can be the disadvantage if the learning rate is very high in gradient descent?

A) Gradient Descent algorithm can diverge from the optimal solution.

B) Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.

C) Both of them

D) None of them

7. As the model complexity increases, what will happen?

A) Bias will increase, Variance decrease

B) Bias will decrease, Variance increase

C)both bias and variance increase

D) Both bias and variance decrease.

8. Suppose I have a linear regression model which is performing as follows: Train accuracy=0.95 and Test accuracy=0.75

Which of the following is true regarding the model?

A) model is underfitting

B) model is overfitting

C) model is performing good

D) None of the above

Q9 to Q15 are subjective answer type questions, Answer them briefly

QUESTION 9

Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

ANSWER

Gini Index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

If all the elements belong to a single class, then it can be called pure. The degree of Gini Index varies between 0 and 1,

'0' denotes that all elements belong to a certain class or there exists only one class (pure), and

'1' denotes that the elements are randomly distributed across various classes (impure).

A Gini Index of '0.5' denotes equally distributed elements into some classes.

QUESTION 10

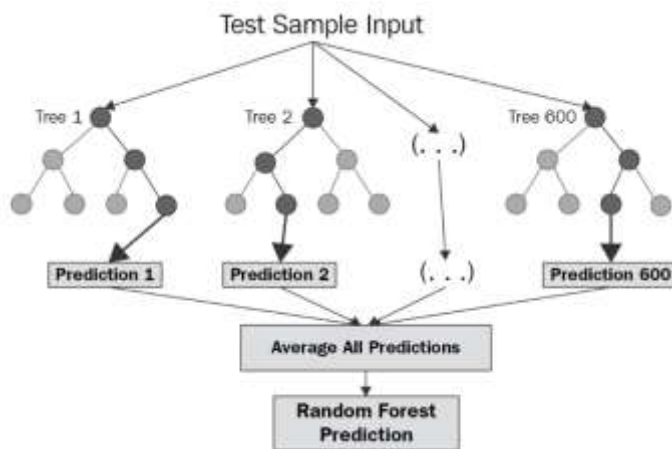
What are the advantages of Random Forests over Decision Tree?

ANSWER

Random forest is yet another powerful and most used supervised learning algorithm. It allows quick identification of significant information from vast datasets. The biggest advantage of Random forest is that it relies on collecting various decision trees to arrive at any solution.

This is an ensemble algorithm that considers the results of more than one algorithms of the same or different kind of classification.

1. Randomly chose " k " features satisfying condition $k < m$.
2. Among the k features, calculate the root node by choosing a node with *the highest Information gain*.
3. Split the node into child nodes.
4. Repeat the previous steps n times.
5. You end up with a forest constituting n trees.
6. Perform *Bootstrapping*, i.e., combining the results of all Decision Trees.



Advantages of random forest

- Random forests are found to be biased while dealing with categorical variables.
- Slow Training.
- Not suitable for linear methods with a lot of sparse features
- It can perform both regression and classification tasks.
- A random forest produces good predictions that can be understood easily.
- It can handle large datasets efficiently.
- The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.
- decision tree is fast
- operates easily on large data sets, especially the linear one.
- The random forest model needs rigorous training

QUESTION 11

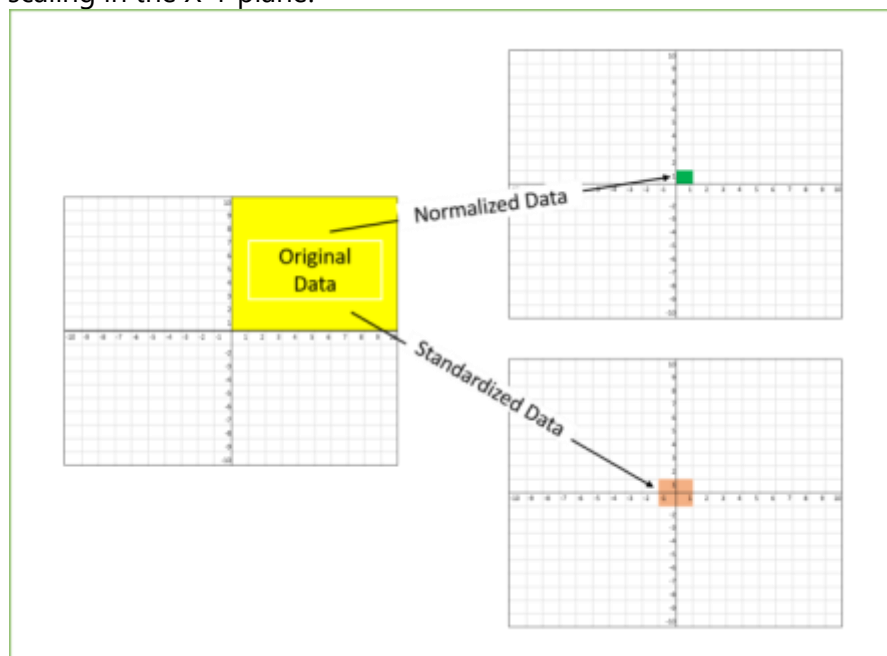
What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

ANSWER

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

Normalization is used when we want to bound our values between two numbers, typically, between $[0,1]$ or $[-1,1]$. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless. Refer to the below diagram, which shows how data looks after scaling in the X-Y plane.



Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

The ML algorithm is sensitive to the “relative scales of features,” which usually happens when it uses the numeric values of the features rather than say their rank.

Normalization is a scaling technique in Machine Learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model. It is required only when features of machine learning models have different ranges.

Normalization techniques in Machine Learning

Although there are so many feature normalization techniques in Machine Learning, few of them are most frequently used. These are as follows:

- **Min-Max Scaling:** This technique is also referred to as scaling. As we have already discussed above, the Min-Max scaling method helps the dataset to shift and rescale the values of their attributes, so they end up ranging between 0 and 1.
- **Standardization scaling:**

Standardization scaling is also known as Z-score normalization, in which values are centered around the mean with a unit standard deviation, which means the attribute becomes zero and the resultant distribution has a unit standard deviation. Mathematically, we can calculate the standardization by subtracting the feature value from the mean and dividing it by standard deviation.

Standardization

Data standardization is the process of rescaling the attributes so that they have mean as 0 and variance as 1. The ultimate goal to perform standardization is to bring down all the features to a common scale without distorting the differences in the range of the values

The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively. The equation is shown below:

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$$

This technique is to re-scale features value with the distribution value between 0 and 1 is useful for the optimization algorithms, such as gradient descent, that are used within machine learning algorithms that weight inputs (e.g., regression and neural networks). Rescaling is also used for algorithms that use distance measurements, for example, K-Nearest-Neighbours (KNN).

QUESTION 12

Write down some advantages which scaling provides in optimization using gradient descent algorithm.

ANSWER

Optimization refers to the task of minimizing/maximizing an objective function $f(x)$ parameterized by x . In machine/deep learning terminology, it's the task of minimizing the cost/loss function $J(w)$ parameterized by the model's parameters $w \in \mathbb{R}^d$. Optimization algorithms (in case of minimization) have one of the following goals:

- Find the global minimum of the objective function. This is feasible if the objective function is convex, i.e. any local minimum is a global minimum.

- Find the lowest possible value of the objective function within its neighborhood. That's usually the case if the objective function is not convex as the case in most deep learning problems.

There are three kinds of optimization algorithms:

- Optimization algorithm that is not iterative and simply solves for one point.
- Optimization algorithm that is iterative in nature and converges to acceptable solution regardless of the parameters initialization such as gradient descent applied to logistic regression.
- Optimization algorithm that is iterative in nature and applied to a set of problems that have non-convex cost functions such as neural networks. Therefore, parameters' initialization plays a critical role in speeding up convergence and achieving lower error rates

Gradient Descent is the most common optimization algorithm in *machine learning* and *deep learning*. It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function $J(w)$ w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α . Therefore, we follow the direction of the slope downhill until we reach a local minimum.

In this article, we'll cover gradient descent algorithm and its variants: *Batch Gradient Descent*, *Mini-batch Gradient Descent*, and *Stochastic Gradient Descent*.

The main advantages:

- Faster than Batch version because it goes through a lot less examples than Batch (all examples).
- Randomly selecting examples will help avoid redundant examples or examples that are very similar that don't contribute much to the learning.
- With batch size $<$ size of training set, it adds noise to the learning process that helps improving generalization error.
- Even though with more examples the estimate would have lower standard error, the return is less than linear compared to the computational burden we incur.

QUESTION 13

In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

ANSWER

Classification accuracy is a metric that summarizes the performance of a classification model as the number of correct predictions divided by the total number of predictions.

It is easy to calculate and intuitive to understand, making it the most common metric used for evaluating classifier models. This intuition breaks down when the distribution of examples to classes is severely skewed.

Intuitions developed by practitioners on balanced datasets, such as 99 percent representing a skillful model, can be incorrect and dangerously misleading on imbalanced classification predictive modeling problems.

In this tutorial, you will discover the failure of classification accuracy for imbalanced classification problems.

After completing this tutorial, you will know:

- Accuracy and error rate are the de facto standard metrics for summarizing the performance of classification models.
- Classification accuracy fails on classification problems with a skewed class distribution because of the intuitions developed by practitioners on datasets with an equal class distribution.
- Intuition for the failure of accuracy for skewed class distributions with a worked example.

Classification predictive modeling involves predicting a class label given examples in a problem domain.

The most common metric used to evaluate the performance of a classification predictive model is classification accuracy. Typically, the accuracy of a predictive model is good (above 90% accuracy), therefore it is also very common to summarize the performance of a model in terms of the error rate of the model.

Classification accuracy involves first using a classification model to make a prediction for each example in a test dataset. The predictions are then compared to the known labels for those examples in the test set. Accuracy is then calculated as the proportion of examples in the test set that were predicted correctly, divided by all predictions that were made on the test set.

- $\text{Accuracy} = \text{Correct Predictions} / \text{Total Predictions}$
Conversely, the error rate can be calculated as the total number of incorrect predictions made on the test set divided by all predictions made on the test set.
- $\text{Error Rate} = \text{Incorrect Predictions} / \text{Total Predictions}$
The accuracy and error rate are complements of each other, meaning that we can always calculate one from the other. For example:

- Accuracy = 1 – Error Rate
- Error Rate = 1 – Accuracy

Another valuable way to think about accuracy is in terms of the confusion matrix.

A confusion matrix is a summary of the predictions made by a classification model organized into a table by class. Each row of the table indicates the actual class and each column represents the predicted class. A value in the cell is a count of the number of predictions made for a class that are actually for a given class. The cells on the diagonal represent correct predictions, where a predicted and expected class align

The confusion matrix provides more insight into not only the accuracy of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made.

The simplest confusion matrix is for a two-class classification problem, with negative (class 0) and positive (class 1) classes.

In this type of confusion matrix, each cell in the table has a specific and well-understood name, summarized as follows:

- 1 | Positive Prediction | Negative Prediction
- 2 Positive Class | True Positive (TP) | False Negative (FN)
- 3 Negative Class | False Positive (FP) | True Negative (TN)

The classification accuracy can be calculated from this confusion matrix as the sum of correct cells in the table (true positives and true negatives) divided by all cells in the table.

- Accuracy = $(TP + TN) / (TP + FN + FP + TN)$

Similarly, the error rate can also be calculated from the confusion matrix as the sum of incorrect cells of the table (false positives and false negatives) divided by all cells of the table.

- Error Rate = $(FP + FN) / (TP + FN + FP + TN)$

Accuracy Fails for Imbalanced Classification

Classification accuracy is the most-used metric for evaluating classification models.

The reason for its wide use is because it is easy to calculate, easy to interpret, and is a single number to summarize the model's capability.

As such, it is natural to use it on imbalanced classification problems, where the distribution of examples in the training dataset across the classes is not equal.

This is the most common mistake made by beginners to imbalanced classification.

When the class distribution is slightly skewed, accuracy can still be a useful metric. When the skew in the class distributions are severe, accuracy can become an unreliable measure of model performance.

The reason for this unreliability is centered around the average machine learning practitioner and the intuitions for classification accuracy.

Typically, classification predictive modeling is practiced with small datasets where the class distribution is equal or very close to equal. Therefore, most practitioners develop an intuition that large accuracy score (or conversely small error rate scores) are good, and values above 90 percent are great.

Achieving 90 percent classification accuracy, or even 99 percent classification accuracy, may be trivial on an imbalanced classification problem

This means that intuitions for classification accuracy developed on balanced class distributions will be applied and will be wrong, misleading the practitioner into thinking that a model has good or even excellent performance when it, in fact, does not.

QUESTION 14

What is "f-score" metric? Write its mathematical formula.

ANSWER

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

It is possible to adjust the F-score to give more importance to precision over recall, or vice-versa. Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score.

F1 score is an alternative machine learning evaluation metric that assesses the predictive skill of a model by elaborating on its class-wise performance rather than an overall performance as done by accuracy. F1 score combines two competing metrics- precision and recall scores of a model,

How to calculate F1 score?

To understand the calculation of the F1 score, we first need to look at a confusion matrix.

A confusion matrix represents the predictive performance of a model on a dataset. For a binary class dataset (which consists of, suppose, "positive" and "negative" classes), a confusion matrix has four essential components:

1. True Positives (TP): Number of samples *correctly* predicted as "positive."
2. False Positives (FP): Number of samples *wrongly* predicted as "positive."
3. True Negatives (TN): Number of samples *correctly* predicted as "negative."
4. False Negatives (FN): Number of samples *wrongly* predicted as "negative."

		actual	
		+ve	-ve
Predicted	+ve	TP	FP
	-ve	FN	TN

f1 is calculated by following formula = $2 \times \left[\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right]$.

QUESTION 15

What is the difference between `fit()`, `transform()` and `fit_transform()`?

ANSWER

Transformers are a commonly used object seen on Scikit-learn. The function of a transformer is to execute the feature transformation process, which is a part of data pre-processing; however, for model training, we require objects referred to as models, such as linear regression, classification, etc. Some examples of the transformer-like objects used for feature selection are StandardScaler, PCA, Imputer, MinMaxScaler, etc... We use these tools to perform some pre-processing on the raw data, such as changing the format of the input data and feature scaling. Further, this data is used for model training.

We use a standardization procedure that takes a feature F and changes it into F' . By utilizing a standardization formula for f_1 , f_2 , f_3 , and f_4 features, f_1 , f_2 , f_3 , and f_4 are the independent features, and f_4 is the dependent feature; we change these features. We can transform an input feature F into another input feature F' with the help of three distinct operations. These operations are:

1. `fit()`
2. `transform()`
3. `fit_transform()`

`fit()` Method

In the `fit()` method, we apply the necessary formula to the feature of the input data we want to change and compute the result before fitting the result to the transformer. We must use the `.fit()` method after the transformer object.

If the StandardScaler object `sc` is created, then applying the `.fit()` method will calculate the mean (μ) and the standard deviation (σ) of the particular feature F . We can use these parameters later for analysis.

Let's use the pre-processing transformer known as StandardScaler as an example and assume that we have to scale the features of self-created data. The example dataset in the code below is created using the `arrange` method and then divided into the training and testing datasets. After that, we create a StandardScaler instance and fit the feature of the training data to it to determine the mean and standard deviation to be utilized for scaling in the future.

The significance of separating the dataset into the train and test datasets before using any pre-processing process, such as scaling, must be emphasized. Test data points represent real-world data. Therefore, we must only execute `fit()` to the training feature to prevent future data to our model.

transform() Method

To change the data, we most likely use the transform() function, where we perform the calculations from fit() to each value in feature F. We transform the fit computations. Hence we must use .transform() after we have applied the fit object.

When we make an object using the fit method, we utilize the example from the section above and place the object in front of the.

The scale of the data points is transformed using the transform and fit_transform method, and the output we receive is always a sparse matrix or array.

fit_transform() Method

The training data is scaled, and its scaling parameters are determined by applying a fit_transform() to the training data. The model we created, in this case, will discover the mean and variance of the characteristics in the training set.

The mean and variance of every feature reported in our data are calculated using the fit approach. The transform method transforms all features using the corresponding means and variances.

We wish scaling to be implemented in our testing data, but we also don't want our model to be biased. We expect our test set of data to be entirely fresh and unexpected for our model. In this situation, the convert approach is useful.

explored the three sklearn transformer functions, fit(), transform(), and fit_transform(), that are most frequently used. We looked at what each performs, how they differ, and in what situations we should choose one over the other. In simple language, the fit() method will allow us to get the parameters of the scaling function. The transform() method will transform the dataset to proceed with further data analysis steps. The fit_transform() method will determine the parameters and transform the dataset.

ASSIGNMENT - 7 SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. The primary key is selected from the

A. Composite keys

B. Candidate keys

C. Foreign keys

D. Determinants

2. Which is/are correct statements about primary key of a table?

A. Primary keys can contain NULL values.

B. Primary keys cannot contain NULL values...

C. A table can have only one primary key with single or multiple fields....

D. A table can have multiple primary keys with single or multiple fields.

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

3. Which SQL command is used to insert a row in a table?

A. Select

B. Create

C. Insert

D. Drop

4. Which one of the following sorts rows in SQL?

A. SORTBY

B. ALIGNBY

C. ORDERBY

D. GROUPBY

5. The SQL statement that queries or reads data from a table is

A. QUERY

B. READ

C. SELECT

D. QUERY

6. Which normal form is considered adequate for relational database design?

A. 1NF

B. 2NF

C. 3NF

D. 4NF

7. SQL can be used to

A. Create database structures only

B. Modify database data only

C. All of the above can be done by SQL

D. Query database data only ASSIGNMENT

8. SQL query and modification commands make up

A. DDL

B. DML

C. HTML

D. XML

9. The result of a SQL SELECT statement is a(n)

A. File

B. Table

C. Report

D. Form

10. Second normal form should meet all the rules for

A. 1 NF

B. 2 NF

C. 3 NF

D. 4 NF

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

QUESTION 11

What are joins in SQL?

ANSWER

A JOIN clause is used to combine rows from two or more tables, based on a related column between them.

Different Types of SQL JOINS

Here are the different types of the JOINS in SQL:

- (INNER) JOIN: Returns records that have matching values in both tables

The INNER JOIN keyword selects records that have matching values in both tables.

INNER JOIN Syntax

SELECT *column_name(s)*

FROM *table1*

INNER JOIN *table2*

ON *table1.column_name = table2.column_name;*

LEFT (OUTER) JOIN: Returns all records from the left table, and the matched records from the right table

The LEFT JOIN keyword returns all records from the left table (table1), and the matching records from the right table (table2). The result is 0 records from the right side, if there is no match.

LEFT JOIN Syntax

```
SELECT column_name(s)
FROM table1
LEFT JOIN table2
ON table1.column_name = table2.column_name;
```

RIGHT (OUTER) JOIN: Returns all records from the right table, and the matched records from the left table

SQL RIGHT JOIN Keyword

The RIGHT JOIN keyword returns all records from the right table (table2), and the matching records from the left table (table1). The result is 0 records from the left side, if there is no match.

RIGHT JOIN Syntax

```
SELECT column_name(s)
FROM table1
RIGHT JOIN table2
ON table1.column_name = table2.column_name;
```

FULL (OUTER) JOIN: Returns all records when there is a match in either left or right table

SQL FULL OUTER JOIN Keyword

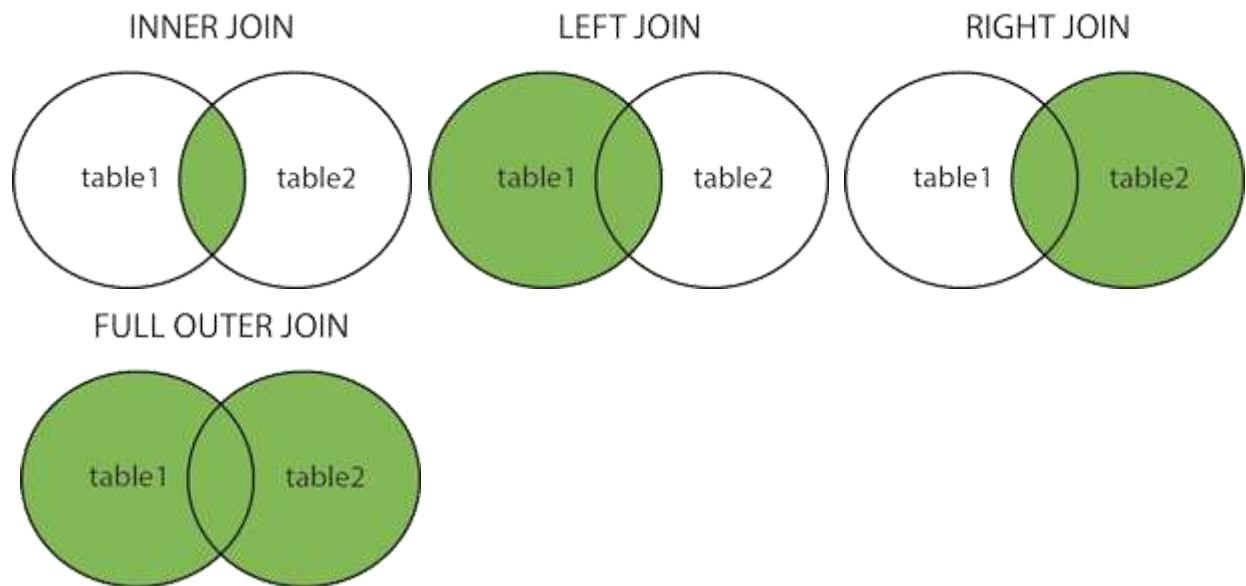
The FULL OUTER JOIN keyword returns all records when there is a match in left (table1) or right (table2) table records.

Tip: FULL OUTER JOIN and FULL JOIN are the same.

FULL OUTER JOIN Syntax

```
SELECT column_name(s)
FROM table1
FULL OUTER JOIN table2
```

ON *table1.column_name = table2.column_name*
WHERE *condition*;



QUESTION 12

What are the different types of joins in SQL?

ANSWER

A JOIN clause is used to combine rows from two or more tables, based on a related column between them.

Different Types of SQL JOINS

Here are the different types of the JOINS in SQL:

- (INNER) JOIN: Returns records that have matching values in both tables

The INNER JOIN keyword selects records that have matching values in both tables.

INNER JOIN Syntax

```
SELECT column_name(s)
FROM table1
INNER JOIN table2
ON table1.column_name = table2.column_name;
```

LEFT (OUTER) JOIN: Returns all records from the left table, and the matched records from the right table

The LEFT JOIN keyword returns all records from the left table (table1), and the matching records from the right table (table2). The result is 0 records from the right side, if there is no match.

LEFT JOIN Syntax

```
SELECT column_name(s)
FROM table1
LEFT JOIN table2
ON table1.column_name = table2.column_name;
```

RIGHT (OUTER) JOIN: Returns all records from the right table, and the matched records from the left table

SQL RIGHT JOIN Keyword

The RIGHT JOIN keyword returns all records from the right table (table2), and the matching records from the left table (table1). The result is 0 records from the left side, if there is no match.

RIGHT JOIN Syntax

```
SELECT column_name(s)
FROM table1
RIGHT JOIN table2
ON table1.column_name = table2.column_name;
```

FULL (OUTER) JOIN: Returns all records when there is a match in either left or right table

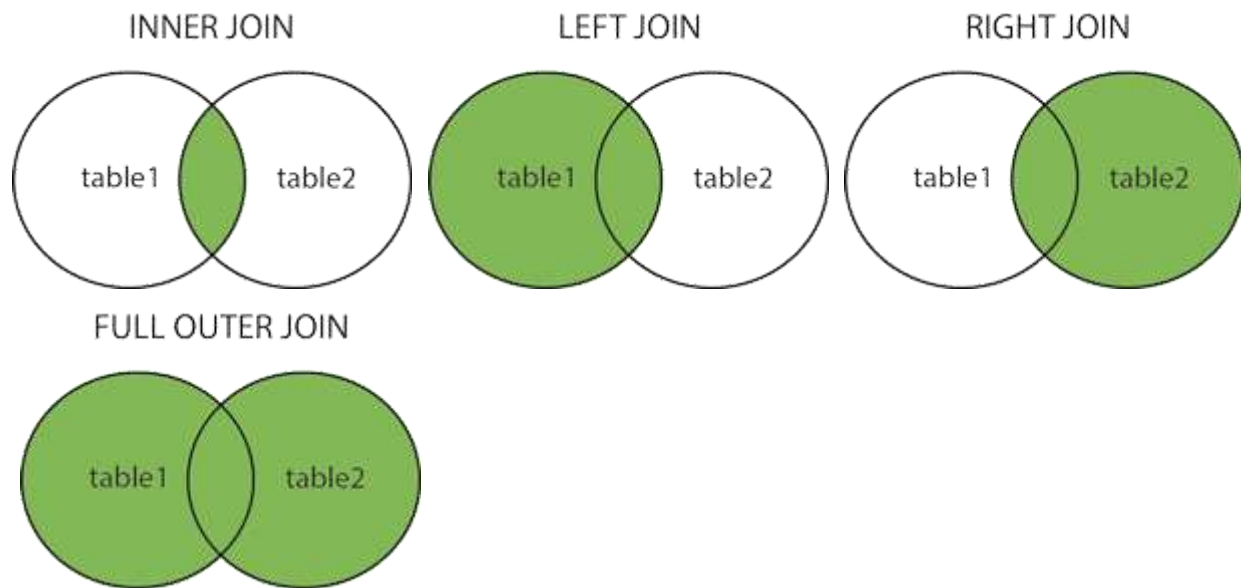
SQL FULL OUTER JOIN Keyword

The FULL OUTER JOIN keyword returns all records when there is a match in left (table1) or right (table2) table records.

Tip: FULL OUTER JOIN and FULL JOIN are the same.

FULL OUTER JOIN Syntax

```
SELECT column_name(s)
FROM table1
FULL OUTER JOIN table2
ON table1.column_name = table2.column_name
WHERE condition;
```



QUESTION 13

What is SQL Server?

ANSWER

SQL Server is software (A Relational Database Management System) developed by Microsoft. It is also called MS SQL Server. It is implemented from the specification of RDBMS.

Our SQL Server Tutorial includes all topics of SQL Server such as SQL Server tutorial with SQL Server, install visual studio, install SQL Server, architecture, management studio, datatypes, db operations, login database, create database, select database, drop database, create table, delete tabel, update table, min function, max function, sum function, sql operators, advance operator, clauses, create view, keys constraints and indexes, primary keys, foreign keys, indexes etc.

The relational database management system (RDBMS) is a Microsoft software product mainly used to store and retrieve data for the same or other applications. We can run these applications on the same computer or a different one.

Microsoft developed and marketed the SQL Server relational database management system (RDBMS) to primarily compete with the MySQL and Oracle databases. It is also called MS SQL Server, which is an ORDBMS, platform-dependent, and can work on GUI and command-based software. The key interface tool for SQL Server is SQL Server Management Studio (SSMS), which operates in both 32-bit and 64-bit environments.

If we want to understand the SQL Server completely, we must first learn the SQL language. SQL is a query processing language used for dealing with data in relational databases. According to the client-server model, a database server is a computer program that provides several services for our database to other programs or computers. As a result, we referred to a SQL Server as a database server that uses SQL as its query language.

Microsoft SQL Server comes in several versions, each corresponding to various workloads and demands. The data center edition is optimized for higher application support and scalability levels, while the Express edition is a free, scaled-down version of the software.

Usage of SQL Server

The following are the key usage of MS SQL Server:

- Its main purpose is to build and maintain databases.
- It is used to analyze the data using SQL Server Analysis Services (SSAS).
- It is used to generate reports using SQL Server Reporting Services (SSRS).
- It is used to perform ETL operations using SQL Server Integration Services (SSIS).

SQL Server comprises five editions with different bundled services and tools and pricing options to meet the user needs. Microsoft provides two editions of SQL Server free of charge, which are given below:

SQL Server Developer: This edition was released mainly for use in the non-production environment, i.e., database development and testing. It allows to build, test, and demo purpose.

SQL Server Express: It is used for small-scale applications and databases with disc storage capacities of up to 10 GB.

For commercial purposes, the following editions are used:

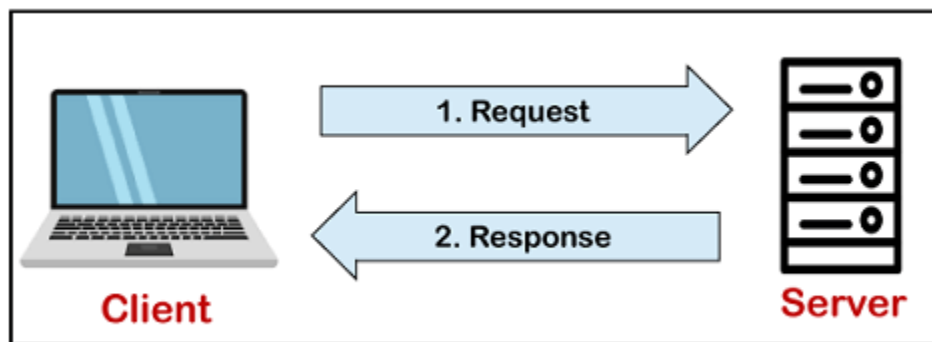
SQL Server Enterprise: It is used in high-end, larger, and more critical businesses. All SQL Server features, such as high-end security, advanced analytics, and machine learning, are included in this version.

SQL Server Standard: This edition is suitable for data marts and mid-tier applications that included basic reporting and analytics. It supports partial enterprise edition's feature, as well as server limitations on the number of processor cores and memory that we can configure.

SQL Server WEB: This edition is suitable for Web hosts who want a low overall ownership cost. It has features of scalability, manageability capabilities, and affordability for small to large-scale web properties.

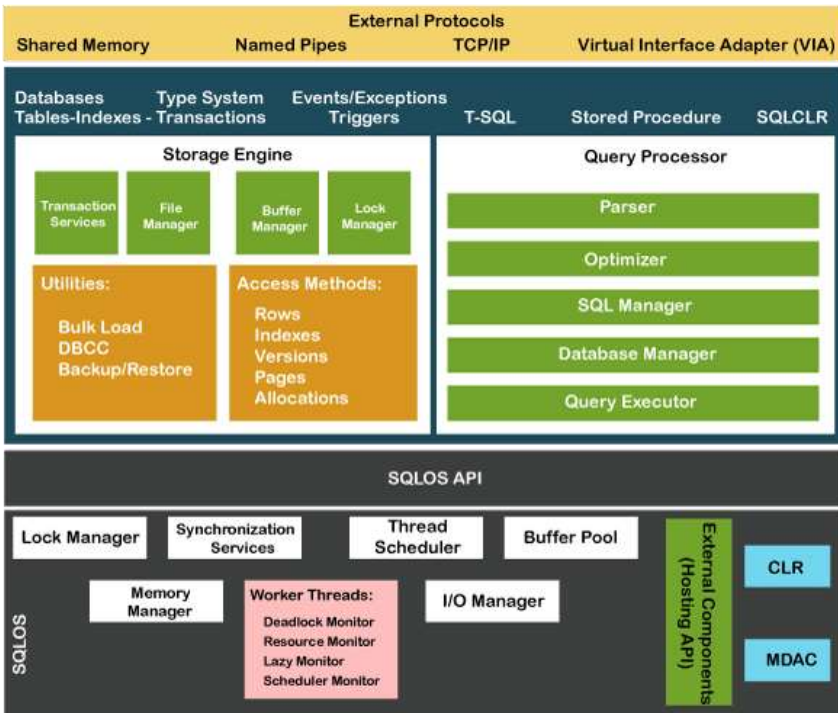
SQL Server as Client-Server Architecture

SQL Server is based on a Client-Server Architecture and is intended for end-users known as clients who send requests to the MS SQL Server installed on a particular computer. The server will give the desired output as soon as the processing input data is requested. This server is available as a separate program and responsible for handling all the database instructions, statements, or commands. The SQL Server Database Engine, which controls data storage, processing, and security, is thus the core component of MS SQL Server.



SQL Server Architecture

The below diagram explains the basic overview of the SQL Server architecture:



SQL Server works on a client-server architecture. It looks very simple from the front end, but internally, multiple processes run in the background to fulfill this request. Based on the architecture, the SQL Server mainly has three major components:

1. Network Protocols (SNI- SQL Server Network Interface)
2. Database Engine
3. SQLOS

Network Protocol

It is entirely responsible for the SQL Server database engine's client connectivity. For client connectivity to the SQL Server database engine. It also has one more protocol named VIA. VIA is a hardware-based protocol that is now obsolete by Microsoft. In the latest SQL Server Configuration Manager, we will not see this protocol.

It supports three primary protocols for network connectivity:

- **Shared Memory:** It is the simplest protocol that does not require any configuration. It works on the same system where SQL Server is installed. There is no communication between the client and the server.

- **TCP/IP:** This is the most commonly used client-server communication protocol. We can use the SQL Server Configuration Manager to enable it.
- **Named Pipes:** It is mainly used for LAN connectivity and can be enabled from the SQL Server Configuration Manager.

Database Engine

It is the core of the SQL Server architecture. It is the second layer of the architecture that provides connectivity between user connections using network protocol and SQL server operating system to perform actual execution. It shows the logical architectures of the database objects such as tables, views, stored procedures, and triggers that work with physical architecture and relation engine to fulfill client requests.

The Database Engine consists of two parts:

- **Relational Engine:** It is responsible for evaluating user requests and performs execution. It decides the most efficient way to run a query. It is also named the query processor. Query processing, memory management, thread and task management, buffer management, and distributed query processing are all main tasks performed by this engine.
- **Storage Engine:** It shows the physical database architecture, as well as data storage and retrieval from storage systems and the buffer manager.

SQL OS

It was first used in **SQL 2005**. Previously, it was only considered for small and medium applications. Microsoft upgrades SQL Server in SQL 2005 to accommodate high-end enterprise database load. It's a layer that lies between the database engine and the windows operating system. Many Operating system services are handled by SQLOS, including memory and I/O management, scheduling, threading, exception handling, and synchronization.

SQL Server Services and Tools

Both data management and business intelligence (BI) tools and services are included in MS SQL Server. Let us discuss them below:

SQL Server includes the following tools and services for data management:

SQL Server Integration Services (SSIS): This tool transfers various data types from one source to another through export, import, transformation, and loading. It converts raw data into information that can be used in the future.

SQL Server Data Quality Services (DQS): It creates a knowledge-based data quality product and employs it to perform data correction, enrichment, standardization, and de-duplication. We can also use it to cleanse data with cloud-based reference data services.

SQL Server Master Data Services (MDS): It is used to manage a master set of the organization's data. It organizes the data into models, create rules for data updation, and control who updates those data.

SQL Server Data Tool (SSDT): It is a database design and development tool.

SQL Server Management Studio (SSMS): This tool allows us to manage, deploy, and monitor SQL Server databases.

SQL Server includes the following tools and services for data analysis:

SQL Server Analysis Services (SSAS): This tool is used in decision support and business analytics analytical data engine. It is designed for deeper and faster data analysis, data mining and also has machine learning capabilities. R and Python language are integrated with SQL Server for advanced analytics.

SQL Server Reporting Services (SSRS): It has decision-making capability as well as a set of tools and services for creating, deploying, and managing reports. Hadoop is integrated with this tool.

SQL Server also has the following essential components:

SQL Server: It enables us to starts, stops, pauses, and continues the MS SQL Server instance.

SQL Server Agent: It works the same as the task scheduler in the computer system. We can use this whenever we need it.

SQL Server Browser: It receives the user's request and connects to the appropriate SQL Server instance.

SQL Server Full-Text Search: Full-text search searches all document keywords that may or may not exactly match the search criteria. It enables the user to run full-text queries against character data in tables.

SQL Server VSS Writer: It is used when the SQL Server is not running to backup and restore data files.

SQL Server Instances

An instance is the installation of SQL Server. We can install several instances on a particular machine, but only one can be the default. It is an exact copy of the server files, databases, and security credentials.

SQL Server is divided into two types:

Primary Instances: We can access the primary instance in two ways. The first is by using the server name, and the second is its IP address. It is always unique.

Named Instances: We can access it by appending a backslash and instance name.

QUESTION 14

What is primary key in SQL?

ANSWER

A primary key is a field in a table which uniquely identifies each row/record in a database table. Primary keys must contain unique values. A primary key column cannot have NULL values.

A table can have only one primary key, which may consist of single or multiple fields. When multiple fields are used as a primary key, they are called a composite key.

If a table has a primary key defined on any field(s), then you cannot have two records having the same value of that field(s).

Create Primary Key

Here is the syntax to define the ID attribute as a primary key in a CUSTOMERS table.

```
CREATE TABLE CUSTOMERS (  
    ID INT NOT NULL,  
    NAME VARCHAR (20) NOT NULL,  
    AGE INT NOT NULL,  
    ADDRESS CHAR (25) ,  
    SALARY DECIMAL (18, 2),  
    PRIMARY KEY (ID)  
);
```

To create a PRIMARY KEY constraint on the "ID" column when the CUSTOMERS table already exists, use the following SQL syntax –

```
ALTER TABLE CUSTOMER ADD PRIMARY KEY (ID);
```

For defining a PRIMARY KEY constraint on multiple columns, use the SQL syntax given below.

```
CREATE TABLE CUSTOMERS (  
  ID INT NOT NULL,  
  NAME VARCHAR (20) NOT NULL,  
  AGE INT NOT NULL,  
  ADDRESS CHAR (25) ,  
  SALARY DECIMAL (18, 2),  
  PRIMARY KEY (ID, NAME)  
);
```

To create a PRIMARY KEY constraint on the "ID" and "NAMES" columns when CUSTOMERS table already exists, use the following SQL syntax.

```
ALTER TABLE CUSTOMERS  
ADD CONSTRAINT PK_CUSTID PRIMARY KEY (ID, NAME);
```

Dropping Primary Key in MySQL

The syntax for dropping the primary key in MySQL is:

```
ALTER TABLE table_name
```

```
DROP CONSTRAINT PRIMARY KEY
```

What Are the Benefits of a Primary Key in SQL?

The most significant advantages of a primary key are:

- It uniquely identifies each row of a table
- It gets a unique index for each primary key column that helps with faster access

What Are the Properties and Rules of an SQL Primary Key?

The properties of each primary key column or columns are:

- It enforces uniqueness by not accepting any duplicate values

- A primary key uniquely identifies each field
- A table can only take one primary key
- Primary columns have a maximum length of 900 bytes
- A primary key column cannot accept null values
- A single-column primary key is a simple one. The one consisting of multiple columns is called a composite primary key

QUESTION 15

What is ETL in SQL?

ANSWER

ETL stands for Extract, Transform and Load. These are three database functions that are combined into one tool to extract data from a database, modify it, and place it into another database.

ETL allows businesses to consolidate data from multiple databases and other sources into a single repository with data that has been properly formatted and qualified in preparation for analysis. This unified data repository allows for simplified access for analysis and additional processing. It also provides a single source of truth, ensuring that all enterprise data is consistent

Extraction, in which raw data is pulled from a source or multiple sources. Data could come from transactional applications, such as customer relationship management (CRM) data from Salesforce or enterprise resource planning (ERP) data from SAP, or Internet of Things (IoT) sensors that gather readings from a production line or factory floor operation, for example. To create a data warehouse, extraction typically involves combining data from these various sources into a single data set and then validating the data with invalid data flagged or removed. Extracted data may be several formats, such as relational databases, XML, JSON, and others.

Transformation, in which data is updated to match the needs of an organization and the requirements of its data storage solution. Transformation can involve standardizing (converting all data types to the same format), cleansing (resolving inconsistencies and inaccuracies), mapping (combining data elements from two or more data models), augmenting (pulling in data from other sources), and others. During this process, rules and functions are applied, and data cleansed to prevent including bad or non-matching data to the destination repository. Rules that could be applied include loading only specific columns, deduplicating, and merging, among others.

Loading, in which data is delivered and secured for sharing, making business-ready data available to other users and departments, both within the organization and externally. This process may include overwriting the destination's existing data. and up-to-date.

ETL tools automate the extraction, transforming, and loading processes, consolidating data from multiple data sources or databases. These tools may have data profiling, data cleansing, and metadata-writing capabilities. A tool should be secure, easy to use and maintain, and compatible with all components of an organization's existing data solutions.

More specifically, the process of extracting data consists of reading data from a database. The transformation happens when the data is converted —using rules, lookup tables or combining it with others— into data that meets the requirements established with the client and then loading it into a new database or data warehouse.

Using ETL ensures that the data is relevant and useful to the client, that it is accurate, high quality, and easily accessible so that the data warehouse is used efficiently and effectively by the end users.

Now more than ever, ensuring the quality and usefulness of data is extremely important. As international data protection regulations have tightened and companies like Google have responded by removing third-party cookies from Chrome, organisations need to rethink their data strategy and adapt to the new era.

SSIS stands for SQL Server Integration Services. SSIS is part of the Microsoft SQL Server data software, used for many data migration tasks. It is basically an ETL tool that is part of Microsoft's Business Intelligence Suite and is used mainly to achieve data integration.

This platform is designed to solve issues related to data integration and workflow applications. It has a storage tool for ETL.

SSIS follows the following steps to achieve the integration:

- It starts with an operational data warehouse, a database designed to integrate data from multiple sources for additional operations on the data.
- The process of extraction, transformation and loading (ETL) is carried out.
- The data warehouse captures data from various sources for useful access and use.
- Data is stored in the data warehouse to bring together and manage data from various sources to answer business questions Therefore, it helps in decision making.

In addition to ETL, SSIS enables other processes such as data cleansing, aggregation, and merging, among others. It facilitates the transfer of data from one database to another and can extract data from a wide variety of sources such as SQL Server databases, Excel files, Oracle and DB2 databases, etc.

SSIS also includes graphical tools and wizards to perform workflow functions such as sending e-

mails, FTP operations, data sources and destinations.

It is clear, then, that it would not be rigorous to talk about the difference between ETL and SSIS, since the name ETL refers to a concept, while SSIS is a Microsoft tool developed to work with the ETL concept.

Download guide

There are several potential points of failure during any ETL process. Snowflake eliminates the need for lengthy, risky, and often labor-intensive ETL processes by making data easily accessible for internal and external partners via secure data sharing and data collaboration.

That said, Snowflake supports both transformations during (extract, transform, load) or after loading (extract, load, transform). Snowflake works with a wide range of data integration tools, including Informatica, Talend, Tableau, Matillion, and others.

In data engineering, new tools and self-service pipelines eliminate traditional tasks such as manual ETL coding and data cleaning companies. With easy ETL or ELT options via Snowflake, data engineers can instead spend more time working on critical data strategy and pipeline optimization projects. In addition, with the Snowflake Cloud Platform as your data lake and data warehouse, extract, transform, load can be effectively eliminated, as no pre-transformations or pre-schemas are needed.

STATISTICS WORKSHEET-7

1. A die is thrown 1402 times. The frequencies for the outcomes 1, 2, 3, 4, 5 and 6 are given in the following table:

Outcome	1	2	3	4	5	6
Frequency	400	300	157	180	175	190

Find the probability of getting 6 as outcome:

a) 0.34

b) 0.135

c) 0.45

d) 0.78

2. A telephone directory page has 400 telephone numbers. The frequency distribution of their unit place digit (for example, in the number 25827689, the unit place digit is 9 is given in table below: First row refers to the digits Second row to their frequencies.

0	1	2	3	4	5	6	7	8	9
44	52	44	44	40	20	28	56	32	40

What will be the probability of getting a digit with unit place digit odd number that is 1, 3,5,7,9?

a) 0.67

b) 0.60

c) 0.45

d) 0.53

3. A tyre manufacturing company which keeps a record of the distance covered before a tyre needed to be replaced. The table below shows the results of 1100 cases

Distance (miles)	<14000	4000-9000	9001-14000	>14000
Frequency	20	260	375	445

If we buy a new tyre of this company, what is the probability that the tyre will last more than 9000 miles?

a) 0.67

b) 0.459

c) 0.745

d) 0.73

4. Please refer to the case and table given in the question No. 3 and determine what is the probability that if we buy a new tyre then it will last in the interval [4000-14000] miles?

a) 0.56

b) 0.577

c) 0.745

d) 0.73

5. We have a box containing cards numbered from 0 to 9. We draw a card randomly from the box. If it is told to you that the card drawn is greater than 4 what is the probability that the card is odd?

a) 0.5

b) 0.8

c) 0.6

d) 0.7

6. We have a box containing cards numbered from 1 to 8. We draw a card randomly from the box. If it is told to you that the card drawn is less than 4 what is the probability that the card is even?

a) 0.33

b) 0.40

c) 0.56

d) 0.89

7. A die is thrown twice and the sum of the numbers appearing is observed to be 7. What is the conditional probability that the number 6 has appeared at least on one of the die?

a) 0.45

b) 0.37

c) 0.33

d) 0.89

8. Consider the experiment of tossing a coin. If the coin shows tail, toss it again but if it shows head, then throw a die. Find the conditional probability of the event that 'the die shows a number greater than 4' given that 'there is at least one Head'

a) 0.1

b) 0.22

c) 0.38

d) 0.45

9. There are three persons Evan, Ross and Michelle. These people lined up randomly for a picture. What is the probability of Ross being at one of the ends of the line?

a) 0.66

b) 0.45

c) 0.23

d) 0.56

10. Let us make an assumption that each born child is equally likely to be a boy or a girl. Now suppose, if a family has two children, what is the conditional probability that both are girls given that at least one of them is a girl?

a) 0.33

b) 0.45

c) 0.56

d) 0.26

11. Consider the same case as in the question no. 10. It is given that elder child is a boy. What is the conditional probability that both children are boys?

a) 0.33

b) 0.23

c) 0.5

d) 0.76

12. We toss a coin. If we get head, we toss a coin again and if we get tail we throw a die. What is the probability of getting a number greater than 4 on die?

a) 0.166

b) 0.34

c) 0.78

d) 0.34

13. We toss a coin. If we get head, we toss a coin again and if we get tail we throw a die. What is the probability of getting an odd number on die?

a) 0.345

b) 0.79

c) 0.2

d) 0.25

14. Suppose we throw two dice together. What is the conditional probability of getting sum of two numbers found on the two die after throwing is less than 4, provided that the two numbers found on the two die are different?

a) 0.3

b) 0.56

c) 0.24

d) 0.06

15. A box contains three coins: two regular coins and one fake two-headed coin, you pick a coin at random and toss it. What is the probability that it lands heads up?

a) $\frac{1}{3}$

b) $\frac{2}{3}$

c) $\frac{1}{2}$

d) $\frac{3}{4}$

MACHINE LEARNING ASSIGNMENT - 8

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

1. What is the advantage of hierarchical clustering over K-means clustering?

A) Hierarchical clustering is computationally less expensive

B) In hierarchical clustering you don't need to assign number of clusters in beginning

C) Both are equally proficient

D) None of these

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

A) max_depth

B) n_estimators

C) min_samples_leaf

D) min_samples_splits

3. Which of the following is the least preferable resampling method in handling imbalance datasets?

A) SMOTE

B) RandomOverSampler

C) RandomUnderSampler

D) ADASYN

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?

1. Type1 is known as false positive and Type2 is known as false negative.

2. Type1 is known as false negative and Type2 is known as false positive.

3. Type1 error occurs when we reject a null hypothesis when it is actually true.

A) 1 and 2

B) 1 only

C) 1 and 3

D) 2 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur:

1. Randomly selecting the cluster centroids

2. Updating the cluster centroids iteratively

3. Assigning the cluster points to their nearest center

A) 3-1-2

B) 2-1-3

C) 3-2-1

D) 1-3-2

6. Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?

A) Decision Trees

B) Support Vector Machines

C) K-Nearest Neighbors

D) Logistic Regression

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

A) CART is used for classification, and CHAID is used for regression.

B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of two children for a node).

C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)

D) None of the above

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

8. In Ridge and Lasso regularization if you take a large value of regularization constant(λ), which of the following things may occur?

A) Ridge will lead to some of the coefficients to be very close to 0

B) Lasso will lead to some of the coefficients to be very close to 0

C) Ridge will cause some of the coefficients to become 0

D) Lasso will cause some of the coefficients to become 0.

9. Which of the following methods can be used to treat two multi-collinear features?

A) remove both features from the dataset

B) remove only one of the features

C) Use ridge regularization

D) use Lasso regularization

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

A) Overfitting

B) Multicollinearity

C) Under fitting

D) Outliers

Q10 to Q15 are subjective answer type questions, Answer them briefly.

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Feature Engineering is an essential component of the data science model development pipeline. A data scientist spends most of the time analyzing and preparing features to train a robust model. A raw dataset consists of various types of features including categorical, numerical, time-based features.

A machine learning or deep learning model understands only numerical vectors. The categorical and time-based features need to be encoded into the numerical format. There are various feature engineering strategies to encode categorical features include One-Hot Encoding, Count Vectorizer, and many more.

Time-based features include the day of month, day of week, day of year , time. Time-based features are cyclic or seasonal in nature. In this article, we will discuss why One-Hot encoding or

dummy encoding should be avoided for cyclic features, instead discuss and implement a better and elegant solution.

Why NOT One-Hot Encoding?

One-hot Encoding is a feature encoding strategy to convert categorical features into a numerical vector. For each feature value, the one-hot transformation creates a new feature demarcating the presence or absence of feature value.



weekday_name	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
Monday	0	1	0	0	0	0	0
Sunday	0	0	0	1	0	0	0
Monday	0	1	0	0	0	0	0
Thursday	0	0	0	0	1	0	0
Monday	0	1	0	0	0	0	0

One-hot encoding creates d-dimensional vectors for each instance where d is the unique number of feature values in the dataset.

For a feature having a large number of unique feature values or categories, one-hot encoding is not a great choice. There are various other techniques to encode the categorical (ordinal or nominal) features.

Time-based features such as day of month, day of week, day of year, etc. have a cyclic nature and have many feature values. One-hot encoding day of month feature results in 30 dimensionality vector, day of year results in 366 dimension vector. It's not a great choice to one-hot encode these features, as it may lead to a curse of dimensionality.

Idea:

The elegant solution to encode these cyclic features can be using mathematical formulation and trigonometry. In this article, we will encode the cyclic features using the basic formulation of trigonometry, by computing the sin and cosine of the features

The discussed trigonometry-based feature transformation can be implemented on any of the cyclical occurring features. One Hot Encoding works well with a relatively small amount of categorical values but it's not recommended to one-hot encode features having many feature values or categories.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Classification problems are quite common in the machine learning world. As we know in the classification problem we try to predict the class label by studying the input data or predictor where the target or output variable is a categorical variable in nature.

If you have already dealt with classification problems, you must have faced instances where one of the target class labels' numbers of observation is significantly lower than other class labels. This type of dataset is called an imbalanced class dataset which is very common in practical classification scenarios. Any usual approach to solving this kind of machine learning problem often yields inappropriate results.

In this article, I'll discuss the imbalanced dataset, the problem regarding its prediction, and how to deal with such data more efficiently than the traditional approach.

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. We can better understand imbalanced dataset handling with an example.

Let's assume that XYZ is a bank that issues a credit card to its customers. Now the bank is concerned that some fraudulent transactions are going on and when the bank checks their data they found that for each 2000 transaction there are only 30 Nos of fraud recorded. So, the number of fraud per 100 transactions is less than 2%, or we can say more than 98% transaction is "No Fraud" in nature. Here, the class "No Fraud" is called the majority class, and the much smaller in size "Fraud" class is called the minority class.

More such example of imbalanced data is –

- Disease diagnosis
- Customer churn prediction
- Fraud detection
- Natural disaster

Class imbalanced is generally normal in classification problems. But, in some cases, this imbalance is quite acute where the majority class's presence is much higher than the minority class.

Approach to deal with the imbalanced dataset problem

In rare cases like fraud detection or disease prediction, it is vital to identify the minority classes correctly. So model should not be biased to detect only the majority class but should give equal weight or importance towards the minority class too. Here I discuss some of the few techniques which can deal with this problem. There is no right method or wrong method in this, different techniques work well with different problems.

1. Choose Proper Evaluation Metric

The accuracy of a classifier is the total number of correct predictions by the classifier divided by the total number of predictions. This may be good enough for a well-balanced class but not ideal for the imbalanced class problem. The other metrics such as precision is the measure of how accurate the classifier's prediction of a specific class and recall is the measure of the classifier's ability to identify a class.

For an imbalanced class dataset F1 score is a more appropriate metric. It is the harmonic mean of precision and recall and the expression is –

$$F_1 = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

So, if the classifier predicts the minority class but the prediction is erroneous and false-positive increases, the precision metric will be low and so as F1 score. Also, if the classifier identifies the minority class poorly, i.e. more of this class wrongfully predicted as the majority class then false negatives will increase, so recall and F1 score will low. F1 score only increases if both the number and quality of prediction improves.

2. Resampling (Oversampling and Under sampling)

This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling. After sampling the data we can get a balanced dataset for both majority and minority classes. So, when both classes have a similar number of records present in the dataset, we can assume that the classifier will give equal importance to both classes.

. SMOTE

Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use k nearest neighbor to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.

. BalancedBaggingClassifier

When we try to use a usual classifier to classify an imbalanced dataset, the model favors the majority class due to its larger volume presence. A `BalancedBaggingClassifier` is the same as a sklearn classifier but with additional balancing. It includes an additional step to balance the training set at the time of fit for a given sampler. This classifier takes two special parameters "sampling_strategy" and "replacement". The sampling_strategy decides the type of resampling required (e.g. 'majority' – resample only the majority class, 'all' – resample all classes, etc) and replacement decides whether it is going to be a sample with replacement or not.

5. Threshold moving

In the case of our classifiers, many times classifiers actually predict the probability of class membership. We assign those prediction's probabilities to a certain class based on a threshold which is usually 0.5, i.e. if the probabilities < 0.5 it belongs to a certain class, and if not it belongs to the other class.

For imbalanced class problems, this default threshold may not work properly. We need to change the threshold to the optimum value so that it can efficiently separate two classes. We can use ROC Curves and Precision-Recall Curves to find the optimal threshold for the classifier. We can also use a grid search method or search within a set of values to identify the optimal value.

13. What is the difference between SMOTE and ADASYN sampling techniques?

SMOTE: Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE

ADASYN: ADAPtive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data. The algorithm uses Euclidean distance for KNN Algorithm. The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions. The latter generates the same number of synthetic samples for each original minority sample.

ADASYN is an improved version of Smote. What it does is same as SMOTE just with a minor improvement. After creating those sample it adds a random small values to the points thus making it more

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not? realistic.

In almost any Machine Learning project, we train different models on the dataset and select the one with the best performance.

However, there is room for improvement as we cannot say for sure that this particular model is best for the problem at hand. Hence, our aim is to improve the model in any way possible. One important factor in the performances of these models are their hyperparameters, once we set appropriate values for these hyperparameters, the performance of a model can improve significantly. In this article, we will find out how we can find optimal values for the hyperparameters of a model by using GridSearchCV.

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

GridSearchCV is a function that comes in Scikit-learn's(or SK-learn) model_selection package. So an important point here to note is that we need to have the Scikit learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

As mentioned above, we pass predefined values for hyperparameters to the GridSearchCV function. We do this by defining a dictionary in which we mention a particular hyperparameter along with the values it can take

GridSearchCV is a machine learning library for python. We have an exhaustive search over the specified parameter values for an estimator. An estimator object needs to provide basically a score function or any type of scoring must be passed. There are 2 main methods which can be implemented on GridSearchcv they are fit and predict. There are other also predict_proba, decision_function etc. But the two mentioned are frequently used. According to the type of algorithm which is been used for the dataset at hand for analysis it has its own different parameters. The user needs to give a different set of values for the important parameters. Gridsearchcv by cross-validations will find out the best value for the parameters mentioned. There are default values set for the parameters which can be also taken into consideration.

Intuition Behind GridSearchCV:

Every Data Scientist working on a model needs the best model for the final conclusive analysis. For this GridSearchCV can help build it. The program here is told to run a grid-search with cross-validations. The cross-validation followed in GridSearchCV is k-fold cross-validation approach. So basically in k-fold cross-validation, the given data is been split into k-folds depending on the need of the analyst where every fold at some of the other point of time is been used in testing. If for example $K=3$, then in the first iteration first fold is used to test the model and the rest folds are used to train the model. In the second iteration, the second fold is used to test the model and the first and the third fold is used to train the model. This is repeated unless every fold is used for testing. Evaluating like this the grid search takes into considerations all the combinations of parameters and finds the best possible model for the algorithm being used in the particular problem.

We have in turn reduced the time for searching for the best parameter values. This can be applied to other algorithms and also more set of parameters also.

GridSearchCV is a technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.

- Advantages: exhaustive search, will find the absolute best way to tune the hyperparameters based on the training set.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.

Machine Learning is a branch of Artificial Intelligence. It contains many algorithms to solve various real-world problems. Building a Machine learning model is not only the Goal of any data scientist but deploying a more generalized model is a target of every Machine learning engineer.

Regression is also one type of supervised Machine learning and in this tutorial, we will discuss various metrics for evaluating regression Models and How to implement them using the sci-kit-learn library.

- Regression
- Why we require Evaluation Metrics
- Mean Absolute Error(MAE)
- Mean Squared Error(MSE)
- RMSE
- RMSLE
- R squared
- Adjusted R Squares

- EndNot

Regression

Regression is a type of Machine learning which helps in finding the relationship between independent and dependent variable.

In simple words, Regression can be defined as a Machine learning problem where we have to predict discrete values like price, Rating, Fees, etc.

Why We require Evaluation Metrics?

Most beginners and practitioners most of the time do not bother about the model performance. The talk is about building a well-generalized model, Machine learning model cannot have 100 per cent efficiency otherwise the model is known as a biased model. which further includes the concept of overfitting and underfitting.

It is necessary to obtain the accuracy on training data, But it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.

So to build and deploy a generalized model we require to Evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result.

If one metric is perfect, there is no need for multiple metrics. To understand the benefits and disadvantages of Evaluation metrics because different evaluation metric fits on a different set of a dataset.

Now, I hope you get the importance of Evaluation metrics. let's start understanding various evaluation metrics used for regression tasks.

Dataset

For demonstrating each evaluation metric using the sci-kit-learn library we will use the placement dataset which is a simple linear dataset that looks something like this.

Mean Absolute Error(MAE)

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

To better understand, let's take an example you have input data and output data and use Linear Regression, which draws a best-fit line.

Now you have to find the MAE of your model which is basically a mistake made by the model known as an error. Now find the difference between the actual value and predicted value that is an absolute error but we have to find the mean absolute of the complete dataset.

so, sum all the errors and divide them by a total number of observations And this is MAE. And we aim to get a minimum MAE because this is a loss.

The diagram shows the formula for Mean Absolute Error (MAE):

$$MAE = \frac{1}{N} \sum |Y - \hat{Y}|$$

Annotations in the diagram:

- An arrow points from the text "Divide by total Number of Data Points" to the fraction $\frac{1}{N}$.
- An arrow points from the text "Actual Output" to the Y term in the absolute value.
- An arrow points from the text "Predicted Output" to the \hat{Y} term in the absolute value.
- An arrow points from the text "Sum Of" to the summation symbol \sum .
- An arrow points from the text "Absolute Value of residual" to the absolute value bars $|Y - \hat{Y}|$.

Advantages of MAE

- The MAE you get is in the same unit as the output variable.
- It is most Robust to outliers.

Disadvantages of MAE

- The graph of MAE is not differentiable so we have to apply various optimizers like Gradient descent which can be differentiable.

```
from sklearn.metrics import mean_absolute_error
print("MAE", mean_absolute_error(y_test, y_pred))
```

Now to overcome the disadvantage of MAE next metric came as MSE.

2) Mean Squared Error(MSE)

MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

So, above we are finding the absolute difference and here we are finding the squared difference.

What actually the MSE represents? It represents the squared distance between actual and predicted values. we perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.

$$MSE = \frac{1}{n} \sum \left(\underbrace{y - \hat{y}}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}} \right)^2$$

Advantages of MSE

The graph of MSE is differentiable, so you can easily use it as a loss function.

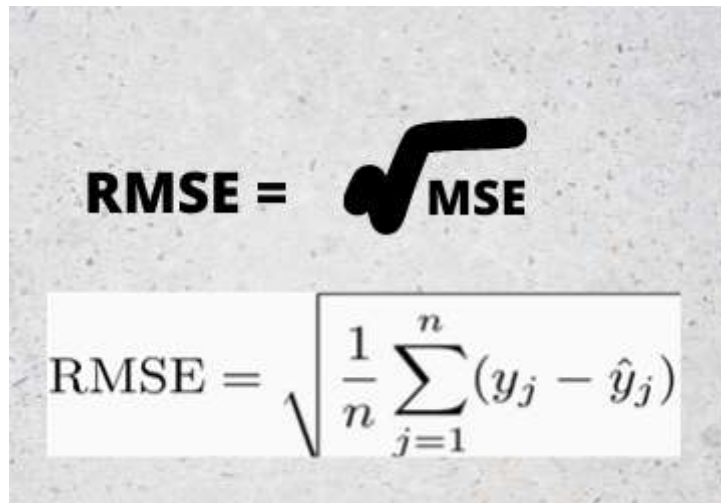
Disadvantages of MSE

- The value you get after calculating MSE is a squared unit of output. for example, the output variable is in meter(m) then after calculating MSE the output we get is in meter squared.
- If you have outliers in the dataset then it penalizes the outliers most and the calculated MSE is bigger. So, in short, It is not Robust to outliers which were an advantage in MAE.

```
from sklearn.metrics import mean_squared_error
print("MSE",mean_squared_error(y_test,y_pred))
```

3) Root Mean Squared Error(RMSE)

As RMSE is clear by the name itself, that it is a simple square root of mean squared error.



The image shows a hand-drawn diagram on a textured background. At the top, it says **RMSE =** followed by a large, bold, hand-drawn square root symbol, and then **MSE**. Below this, there is a white rectangular box containing the mathematical formula for RMSE:
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Advantages of RMSE

- The output value you get is in the same unit as the required output variable which makes interpretation of loss easy.

Disadvantages of RMSE

- It is not that robust to outliers as compared to MAE.

for performing RMSE we have to NumPy NumPy square root function over MSE.

```
print("RMSE",np.sqrt(mean_squared_error(y_test,y_pred)))
```

Most of the time people use RMSE as an evaluation metric and mostly when you are working with deep learning techniques the most preferred metric is RMSE.

4) Root Mean Squared Log Error(RMSLE)

Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.

To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.

To perform RMSLE we have to use the NumPy log function over RMSE.

```
print("RMSE",np.log(np.sqrt(mean_squared_error(y_test,y_pred))))
```

It is a very simple metric that is used by most of the datasets hosted for Machine Learning competitions.

5) R Squared (R2)

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides. The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically R2 squared calculates how much regression line is better than a mean line.

Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

$$\mathbf{R^2\ Squared = 1 - \frac{SSr}{SSm}}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

R2 Squared

Now, how will you interpret the R^2 score? suppose If the R^2 score is zero then the above regression line by mean line is equal means 1 so $1-1$ is zero. So, in this case, both lines are overlapping means model performance is worst, It is not capable to take advantage of the output column.

Now the second case is when the R^2 score is 1, it means when the division term is zero and it will happen when the regression line does not make any mistake, it is perfect. In the real world, it is not possible.

So we can conclude that as our regression line moves towards perfection, R^2 score move towards one. And the model performance improves.

The normal case is when the R^2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.

```
from sklearn.metrics import r2_score
r2 = r2_score(y_test,y_pred)
print(r2)
```

6) Adjusted R Squared

The disadvantage of the R^2 score is while adding new features in data the R^2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

But the problem is when we add an irrelevant feature in the dataset then at that time R^2 sometimes starts increasing which is incorrect.

Hence, To control this situation Adjusted R Squared came into existence.

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

Now as K increases by adding some features so the denominator will decrease, $n-1$ will remain constant. R^2 score will remain constant or will increase slightly so the complete answer will increase and when we subtract this from one then the resultant score will decrease. so this is the case when we add an irrelevant feature in the dataset.

And if we add a relevant feature then the R^2 score will increase and $1-R^2$ will decrease heavily and the denominator will also decrease so the complete term decreases, and on subtracting from one the score increases.

$n=40$

$k=2$

`adj_r2_score = 1 - ((1-r2)*(n-1)/(n-k-1))`

`print(adj_r2_score)`

Hence, this metric becomes one of the most important metrics to use during the evaluation of the model.

EndNote

STATISTICS WORKSHEET-8

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. In hypothesis testing, type II error is represented by β and the power of the test is $1-\beta$ then β is:

a. The probability of rejecting H_0 when H_1 is true

b. The probability of failing to reject H_0 when H_1 is true

c. The probability of failing to reject H_1 when H_0 is true

d. The probability of rejecting H_0 when H_1 is true

2. In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the

a. correct hypothesis

b. null hypothesis

c. alternative hypothesis

d. level of significance

3. When the null hypothesis has been true, but the sample information has resulted in the rejection of the null, a _____ has been made

a. level of significance

b. Type II error c. critical value

d. Type I error

4. For finding the p-value when the population standard deviation is unknown, if it is reasonable to assume that the population is normal, we use

a. the z distribution

b. the t distribution with $n - 1$ degrees of freedom

c. the t distribution with $n + 1$ degrees of freedom

d. none of the above

5. A Type II error is the error of

a. accepting H_0 when it is false

b. accepting H_0 when it is true

c. rejecting H_0 when it is false

d. rejecting H_0 when it is true

6. A hypothesis test in which rejection of the null hypothesis occurs for values of the point estimator in either tail of the sampling distribution is called

a. the null hypothesis

b. the alternative hypothesis

c. a one-tailed test

d. a two-tailed test

7. In hypothesis testing, the level of significance is

a. the probability of committing a Type II error

b. the probability of committing a Type I error

c. the probability of either a Type I or Type II, depending on the hypothesis to be tested
d. none of the above

8. In hypothesis testing, β is

a. the probability of committing a Type II error

b. the probability of committing a Type I error

c. the probability of either a Type I or Type II, depending on the hypothesis to be test

d. none of the above

9. When testing the following hypotheses at an α level of significance

$H_0: p = 0.7$ $H_1: p > 0.7$

The null hypothesis will be rejected if the test statistic Z is

a. $z > z_\alpha$

b. $z < z_\alpha$

c. $z < -z$

d. none of the above

10. Which of the following does not need to be known in order to compute the P-value?

a. knowledge of whether the test is one-tailed or two-tail

b. the value of the test statistic

c. the level of significance

d. All of the above are needed

11. The maximum probability of a Type I error that the decision maker will tolerate is called the

a. level of significance

b. critical value c. decision value

d. probability value

12. For t distribution, increasing the sample size, the effect will be on

a. Degrees of Freedom

b. The t-ratio

c. Standard Error of the Means

d. All of the Above

Q13 to Q15 are subjective answers type questions. Answers them in their own words briefly.

13. What is Anova in SPSS?

ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations.

ANOVA in SPSS must have a dependent variable which should be metric (measured using an interval or ratio scale). ANOVA in SPSS must also have one or more independent variables,

which should be categorical in nature. In ANOVA in SPSS, categorical independent variables are called factors. A particular combination of factor levels, or categories, is called a treatment.

In ANOVA in SPSS, there is one way ANOVA which involves only one categorical variable, or a single factor. For example, if a researcher wants to examine whether heavy, medium, light and nonusers of cereals differed in their preference for Total cereal, then the differences can be examined by the one way ANOVA in SPSS. In one way ANOVA in SPSS, a treatment is the same as the factor level.

If two or more factors are involved in ANOVA in SPSS, then it is termed as n way ANOVA. For example, if the researcher also wants to examine the preference for Total cereal by the customers who are loyal to it and those who are not, then we can use n way ANOVA in SPSS.

In ANOVA in SPSS, from the menu we choose:

"Analyze" then go to "Compare Means" and click on the "One-Way ANOVA."

Now, let us discuss in detail how the software operates ANOVA:

The first step is to identify the dependent and independent variables. The dependent variable is generally denoted by Y and the independent variable is denoted by X. X is a categorical variable having c categories. The sample size in each category of X is generally denoted as n, and the total sample size $N = n \times c$.

The next step in ANOVA in SPSS is to examine the differences among means. This involves decomposition of the total variation observed in the dependent variable. This variation in ANOVA in SPSS is measured by the sums of the squares of the mean.

The total variation in Y in ANOVA in SPSS is denoted by SS_y , which can be decomposed into two components:

$$SS_y = SS_{\text{between}} + SS_{\text{within}}$$

where the subscripts between and within refers to the categories of X in ANOVA in SPSS. SS_{between} is the portion of the sum of squares in Y related to the independent variable or factor X. Thus it is generally referred to as the sum of squares of X. SS_{within} is the variation in Y related to the variation within each category of X. It is generally referred to as the sum of squares for errors in ANOVA in SPSS.

The logic behind decomposing SS_Y is to examine the differences in group means.

The next task in ANOVA in SPSS is to measure the effects of X on Y, which is generally done by the sum of squares of X, because it is related to the variation in the means of the categories of X. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the differences among the means of Y in categories of X increases. The relative magnitude of the sum of squares of X in ANOVA in SPSS increases as the variation in Y within the categories of X decreases.

The strength of the effects of X on Y is measured with the help of η^2 in ANOVA in SPSS. The value of η^2 varies between 0 and 1. It assumes a value 0 in ANOVA in SPSS when all the category means are equal, indicating that X has no effect on Y. The value of η^2 becomes 1, when there is no variability within each category of X but there is still some variability between the categories.

The final step in ANOVA in SPSS is to calculate the mean square which is obtained by dividing the sum of squares by the corresponding degrees of freedom. The null hypothesis of equal means, which is done by an F statistic, is the ratio between the mean square related to the independent variable and the mean square related to the error.

N way ANOVA in ANOVA in SPSS involves simultaneous examination of two or more categorical independent variables, which is also computed in a similar manner.

A major advantage of ANOVA in SPSS is that the interactions between the independent variables can be examined

14. What are the assumptions of Anova?

To use the ANOVA test we made the following assumptions:

- Each group sample is drawn from a normally distributed population
- All populations have a common variance
- All samples are drawn independently of each other
- Within each sample, the observations are sampled randomly and independently of each other
- Factor effects are additive

The presence of outliers can also cause problems. In addition, we need to make sure that the F statistic is well behaved. In particular, the F statistic is relatively robust to violations of normality provided:

- The populations are symmetrical and uni-modal.
- The sample sizes for the groups are equal and greater than 10

In general, as long as the sample sizes are equal (called a balanced model) and sufficiently large, the normality assumption can be violated provided the samples are symmetrical or at least similar in shape (e.g. all are negatively skewed).

The F statistic is not so robust to violations of homogeneity of variances. A rule of thumb for balanced models is that if the ratio of the largest variance to smallest variance is less than 3 or 4, the F -test will be valid. If the sample sizes are unequal then smaller differences in variances can invalidate the F -test. Much more attention needs to be paid to unequal variances than to non-normality of data.

We now look at how to test for violations of these assumptions and how to deal with any violations when they occur.

- Testing that the population is normally distributed (see Testing for Normality and Symmetry)
- Testing for homogeneity of variances and dealing with violations (see Homogeneity of Variances)
- Testing for and dealing with outliers (see [Outliers in ANOVA](#))

15. What is the difference between one way Anova and two way Anova?

A key statistical test in research fields including biology, economics and psychology, analysis of variance (ANOVA) is very useful for analyzing datasets. It allows comparisons to be made between three or more groups of data. Here, we summarize the key differences between these

two tests, including the assumptions and hypotheses that must be made about each type of test.

There are two types of ANOVA that are commonly used, the one-way ANOVA and the two-way ANOVA. This article will explore this important statistical test and the difference between these two types of ANOVA.

one-way ANOVA?

A one-way ANOVA is a type of statistical test that compares the variance in the group means within a sample whilst considering only one independent variable or factor. It is a hypothesis-based test, meaning that it aims to evaluate multiple mutually exclusive theories about our data. Before we can generate a hypothesis, we need to have a question about our data that we want an answer to. For example, adventurous researchers studying a population of walruses might ask "Do our walruses weigh more in early or late mating season?" Here, the independent variable or factor (the two terms mean the same thing) is "month of mating season". In an ANOVA, our independent variables are organised in categorical groups. For example, if the researchers looked at walrus weight in December, January, February and March, there would be four months analyzed, and therefore four groups to the analysis.

A one-way ANOVA compares three or more than three categorical groups to establish whether there is a difference between them. Within each group there should be three or more observations (here, this means walruses), and the means of the samples are compared.

two-way ANOVA?

A two-way ANOVA is, like a one-way ANOVA, a hypothesis-based test. However, in the two-way ANOVA each sample is defined in two ways, and resultingly put into two categorical groups. Thinking again of our walruses, researchers might use a two-way ANOVA if their question is: "Are walruses heavier in early or late mating season and does that depend on the sex of the walrus?" In this example, both "month in mating season" and "sex of walrus" are factors – meaning in total, there are two factors. Once again, each factor's number of groups must be considered – for "sex" there will only two groups "male" and "female".

The two-way ANOVA therefore examines the effect of two factors (month and sex) on a dependent variable – in this case weight, and also examines whether the two factors affect each other to influence the continuous variable.

The key differences between one-way and two-way ANOVA are summarized clearly below.

1. A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A two-way ANOVA is designed to assess the interrelationship of two independent variables on a dependent variable.

2. A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.
3. In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.