WORKSHEET – 7

MACHINE LEARNING

1.D] All of the above

2.A] Random forest

3.B] The regularization will decrease

4.C] both A & B

5.C] In case of classification problem, the prediction is made by taking mode of the class tables predicted by the component trees.

6.C] Both of them

7.B] Bias will decrease, Variance increase

8.A] Model is underfitting

10. One of the biggest advantages of random forest over decision tree is the algorithm on which the former one works i.e Bagging algorithm. It means random forest replaces the data/population used to construct the tree and also the explanatory variables are bootstrapped so that partition is not done on the same important variable. I am telling this because decision trees do not follow this algorithm. It simply keeps on building tree by determining the important variable which depends on homogeneity. Bootstrapping reduces bias and variance both which makes the model more robust and accurate.

11. Feature Scaling is a method to transform the numeric features in a dataset to a standard range so that the performance of the machine learning algorithm improves. It can be achieved by normalizing or standardizing the data set values. **Vertical Scaling** and **Horizontal Scaling** are two techniques used for scaling.

12. Gradient descent is an optimization algorithm which is commonly used to train machine learning models and neural network. Training data helps these models learn over time, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates. Until the function is close to or equal to zero, the model will continue to adjust its parameters to yield the smallest possible error.

13. Accuracy is not a good metric for imbalanced dataset. This model would receive a very good accuracy score as it predicted correctly for the majority

of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class.

14. 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".

2*[(Precision*Recall) / (Precision + Recall)]

15. fit() : In the fit() method, where we use the required formula and formula and performs the calculation on the feature values of input data and fit this calculation to the transformer.

transform() : For changing the data, we probably do transform in the transform() method, where we apply the calculations that we have calculated in fit() to every data point in feature. It will only perform when we want to do some kind of transformation on the input data.

fit_transform : The fit_transform() method is basically the combination of the fit method and the transform method. This method simultaneously performs fit and transform operations on the input data and converts the data points. Using fit and transform separately when we need them both decrease the efficiency of the model. Instead, fit_transform() is used to get both work done.

SQL

1.B] Candidate keys

2.B] Primary keys cannot contain NULL values

3.C] Insert

4.C] ORDERBY

5.C] SELECT

6.C] 3NF

7.C] All of the above can be done by SQL

8.B] DML

9.B] Table

10.A] 1NF

11. SQL JOIN statement is used to combine data or rows from two or more tables based on a common field between them.

12. INNER JOIN : The INNER JOIN keyboard selects all rows from both the tables as long as the condition is satisfied. This keyboard will create the result-set by combining all rows from both the tables where the condition satisfies i.e value of the common field will be same.

LEFT JOIN : This join returns all the rows of the table on the left side of the join and matches rows for the table on the right side of the join. For the row for which there is no matching row on the right side, the result-set will contain null. LEFT JOIN is also known as LEFT OUTER JOIN.

RIGHT JOIN : RIGHT JOIN is similar to LEFT JOIN. This join return all the rows of the table on the right side of the join and matching rows for the table on the left side of the join. For the rows for which there is no matching row on the left side, the result-set will contain null. RIGHT JOIN is also known as RIGHT OUTER JOIN.

FULL JOIN : FULL JOIN creates the result-set by combining results of both LEFT JOIN and RIGHT JOIN. The result-set will contain all the rows from both tables. For the rows for which there is no matching, the result-set will contain NULL values.

13. SQL Server is a relational database management system, or RDBMS, developed and marketed by Microsoft. Similar to other RDBMS software, SQL Server is built on top of SQL, a standard programming language for interacting with relational database. SQL Server is tied to Transact-SQL, or T-SQL, the Microsoft implementation of SQL that adds a set of proprietary programming construct. SQL Server works exclusively on the windows environment for more than 20 years.

14. In SQL, a primary key is a single field or combination of fields that uniquely defines a record. None of the field that are part of the primary key can contain a NULL value. A table can have only one primary key. You use either the **CREATE TABLE** statement or the **ALTER TABLE** statement to create a primary key in SQL.

15. ETL, which stands for **Extract, Transform and Load,** is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system. ETL is often used by an organization to : Extract data from

legacy systems, Cleanse the data to improve data quality and establish consistency, Load data into a target database.

STASTISTICS

1.B] 0.135

2.D] 0.53

3.C] 0.745

4.B] 0.577

5.A] 0.5

6.A] 0.33

7.B] 0.37

8.B] 0.22

9.C] 0.23

10.C] 0.56

11.D]0.76

12.D] 0.34

13.A] 0.345

14.D]0.06

15.C]1/2

&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

WORKSHEET – 8

MACHINE LEARNING

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

2.A] max_depth

3.A]SMOTE

4.C] 1 and 3

5.D] 1-3-2

6.B] Support Vector Machines

7.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)

8.B] Lasso will lead to some of the coefficients to be very close to 0

9.D] Use Lasso regularization

10.A] Overfitting

11. One-hot encoding created d-dimensional vectors for each instance where d is the unique number of feature values in the dataset. For a features having a large number of unique feature values or categories, one-hot encoding is not a great choice. One-hot encoding in machine learning is the conversion of categorical information into a format that may be fed into machine learning algorithm to improve prediction accuracy. One-hot encoding is a common method for dealing with categorical data in machine learning.

13. SMOTE : First it finds the n-nearest neighbors in the minority class for each of the samples in the class. Then it draws a line between the neighbor an generates random points on the lines.

ADASYN : It's a improved version of SMOTE. What it does is same as SMOTE just with minor improvement. After creating those sample it adds a random small values to the points thus making it more realistic. In other words instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e they are bit scattered.

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

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

Mean Absolute Error (MAE) : MAE is very simple metrics which calculates the absolutes difference between actual and predicted values.

Mean Squared Error (MSE) : MSE is a most used and very simple metrics with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value. It represent 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.

R Squared (R2) : R2 score is a metrics that tells the performance of your model, not the loss in an absolute sense thar how many wells did your model perform. It constrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

Adjusted R Squared : The disadvantages of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decrease because it assumes that while adding more data variance of data increase. But the problem is when we add an irrelevant features in the dataset then at that time R2 sometimes starts increasing which is incorrect.

STATSTICS

1.A] The probability of rejecting H0 when H1 is true

2.B] Null Hypothesis

3.D] Type I error

4.B] The t distribution with n-1 degrees of freedom

5.C] Rejecting H0 when it is false

6.D] A two-tailed test

7.B] The probability of committing a Type I error

8.D] None of the above

9.D] None of the above

10.C] Level of significance

11.A] Level of significance

12.C] Standard Error of the Means

13. ANOVA in SPSS, is used for examining the difference in the mean values of the dependent variable associates 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 mean for two or more population. ANOVA in SPSS must have a dependent variable which should be metrics. 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, is one way ANOVA which involves only one categorical variable, or a single factor.

14. There are three primary assumption in ANOVA :

- The responses for each factor level have a normal population distribution.
- These distribution have the same variance.
- The data are independent.

15. The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two :

- **<u>One-way ANOVA</u> :** It focus on simply one independent variable and one dependent variable. However, variables rarely exist in isolation in the real world.
- **<u>One-way ANOVA</u> :** The two way ANOVA focuses on two independent variable to examine these more complex, real-life situation, thus increasing the external validity of the study.