MACHINE LEARNING WORKSHEET 4

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

1. **The value of correlation coefficient will always be:**

Ans. C) between -1 and 1

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

Ans. B) PCA

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

Ans. C) hyperplane

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

Ans. A) Logistic Regression

5. **In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be?**
   (1 kilogram = 2.205 pounds)

Ans. C) old coefficient of 'X' ÷ 2.205

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

Ans. B) increases

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

Ans. C) Random Forests are easy to interpret

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

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

Ans. B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

9. **Which of the following are applications of clustering?**

Ans. A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

C) Identifying spam or ham emails

D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

10. **Which of the following is(are) hyper parameters of a decision tree?**

Ans. B) max_features

D) min_samples_leaf

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

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

Ans. Outliers are observation that does not follow normal distribution or gaussian distribution curve. Generally, these are occurred due to false sampling, any error in collecting data or during Sampling. Although outliers are present in data, sometimes they are true values, but disturbs overall performance of Model. So, it is important to handle those outliers. IQR also known as inter quartile range, is method used to outlie those outliers from data set. In IQR, based on data quartiles are created and points in quartile are plotted in plot, called as box plot. These box plots are easy way of understanding outliers present in data set. Basically IQR, is divided into four groups, those are Q1, Q2, Q3 and Q4. IQR is difference between Q3 – Q1.

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

Ans. Bagging: It is method used to reduce variance in data

Boosting: It is method used to reduce training errors.

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

Ans. It is used to measure accuracy of Linear Regression models. It is calculated by residual mean by total mean and it is always less than and equal to R square. Formula for the same is, Adjusted R2 = 1 – [((1-R2) x (N – 1))/(N – p – 1))] Where R2 = Sample R Squared N = Total Sample Size P = Number of independent Variable

14. **What is the difference between standardisation and normalisation?**

Ans. Normalization -: It is used when features are of different scales. It is used when features are of different scales. It is affected by outliers. MinMaxSclare is used for normalization. It is useful when we don't know about the distribution. It is also known as Scaling Normalization. Scales values between (0,1) or (-1,1). The minimum and Maximum values of features are used for scaling.

Standardization-: It is used when we want to ensure zero mean and until the standard deviation. It is much less affected by outliers. StandardScaler is used for Standardization. It is useful when the feature distribution is Normal or Gaussian. It is also known as Z-score. Normalization It is not bounded to a certain range Mean and Standard Deviation is used for scaling

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

Ans. Cross-validation is a technique for evaluating a machine-learning model and testing its performance. CV is commonly used in applied ML tasks. It helps to compare and select an appropriate model for the specific predictive modeling problem. CV is easy to understand and easy to implement, and it tends to have a lower bias than other methods used to count the model's efficiency scores. All this makes cross-validation a powerful tool for selecting the best model for the specific task. Advantage - Reduces Overfitting- In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm. Disadvantage - Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.