MACHINE LEARNING

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

1. The value of correlation coefficient will always be:
   A) between 0 and 1 B) greater than -1
   C) between -1 and 1 D) between 0 and -1
2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation B) PCA
   C) Recursive feature elimination D) Ridge Regularisation
3. Which of the following is not a kernel in Support Vector Machines?
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial
4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?
   A) Logistic Regression B) Naïve Bayes Classifier
   C) Decision Tree Classifier D) Support Vector Classifier
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)
   A) 2.205 × old coefficient of 'X' B) same as old coefficient of 'X'
   C) old coefficient of 'X' ÷ 2.205 D) Cannot be determined
6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?
   A) remains same B) increases
   C) decreases D) none of the above
7. Which of the following is not an advantage of using random forest instead of decision trees?
   A) Random Forests reduce overfitting
   B) Random Forests explains more variance in data then decision trees
   C) Random Forests are easy to interpret
   D) Random Forests provide a reliable feature importance estimate

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?
   A) Principal Components are calculated using supervised learning techniques
   B) Principal Components are calculated using unsupervised learning techniques
   C) Principal Components are linear combinations of Linear Variables.
   D) All of the above
9. Which of the following are applications of clustering?
   A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
   B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.
   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?
    A) max_depth B) max_features
    C) n_estimators D) min_samples_leaf

MACHINE LEARNING

Q10 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.) An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

We can use the IQR method of identifying outliers to set up a "fence" outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. This is the method that Minitab uses to identify outliers by default.

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

Ans.) In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.

13. What is adjusted R^2 in linear regression. How is it calculated?

Ans.) Adjusted R2 and R2 both represent that how well the model fits the data points. But adjusted R2 penalizes the model for using more features. In case we increase the number of features in training data the R2 will increase but adjusted R2 will only increase if the new feature adds value to our model. Due to this reason adjusted R2 is considered as a better evaluation metric than R2 . Adjusted R 2 is always less than or equal to R2 . The formula to calculate adjusted R2 is as follows: $Radj$ 2 = [1 − (1 − R 2)(n − 1) n − k − 1] Where, n = number of data points in the dataset K = Number of features in the dataset excluding the constant term

14. What is the difference between standardisation and normalisation?

Ans.) In Normalization a dataset is scaled in such a way that all the data points lie between 0 and 1. Normalization is often called min-max scaling. Formula for Normalization is as follows: min max min new x x x x − = − Whereas, In Standardization a dataset is scaled in such a way that the mean of data points becomes 0 and standard deviation is 1. The transformed data may be positive as well as negative in standardization. The formula for standardization is as follows: i new x x x s − = Where, i x = ith data point x = sample mean s = sample standard deviation

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

Ans.) 15. Cross validation is a technique to fit a model on data set. In cross validation the data set is divided into 'k' number of sets where 'k-1' sets are used for training and 1 set is used as validation set. And this is done for all the set one by one and the final score of model is taken as average score of all the 'k' number of fits. Advantage of using Cross validation is that, there is no need of separate validation data, cross validation reduces chances of overfitting and gives a more generic model. Cross validation has a disadvantage that it takes more time to fit the model over a large dataset and the model built is more complex than the basic model.