MACHINE LEARNING

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

Answers highlighted in Green

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 Regularization B) PCA
   C) Recursive feature elimination D) Ridge Regularization
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. (a)What are outliers?

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors.

(b)Explain the Inter Quartile Range (IQR) method for outlier detection.

The interquartile range is a widely accepted method to find outliers in data. When using the interquartile range, or IQR, the full dataset is split into four equal segments, or quartiles. The distances between the quartiles is what is used to determine the IQR

Mathematically:

IQR=Q3-Q1(Where Q1 and Q3 are the 1st and 3rd Quantile respectively)

Outlier(Higher Side) = Q3+(1.5*IQR)

Outlier(Lower Side) = Q1-(1.5*IQR)

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

BAGGING	**BOOSTING**
The simplest way of combining predictions that belong to the same type	It's a means of combining predictions that belong to the different types.
Bagging tends to decrease variance instead of bias	Boosting aims to decrease bias instead of variance
It's solves over-fitting issues in a model	While Boosting can increase/worsen over-fitting problem

13. (a) What is adjusted R^2 in linear regression.

As we increase the number of independent variables in our equation, the R² increases as well. But that doesn't mean that the new independent variables have any correlation with the output variable. In other words, even with the addition of new features in our model, it is not necessary that our model will yield better results but R² value will increase. To rectify this problem, we use Adjusted R² value which penalizes excessive use of such features which do not correlate with the output data.

(b) How is it calculated?

R²adjusted = 1 – [(1-R²)(N-1)]/(N-p-1)
Where
R² = initial sample R-square
P= Number of predictors or number of features
N=Total sample size or number of records

MACHINE LEARNING

14. What is the difference between Standardisation and Normalisation?

STANDARDIZATION	NORMALIZATION
The process where data is restructured in a uniform format.	The process of arranging data in database
Values are not bounded	It is used to reduce redundancy in which values are shifted and scaled in a range of 0 and 1
Scaling is done by mean and standard deviation	Scaling is done by the highest and the lowest values
It is applied when we verify zero mean and unit standard deviation	It is applied when the features are of separate scales
Less affected by outliers	Affected by outliers
It is used when the data is Gaussian or normally distributed	It is applied when we are not sure about the data distribution

15. (a) What is cross-validation?

Cross-validation is a resampling technique used to tackle overfitting with a basic idea of dividing the training dataset into two parts i.e. train and test. On one part; you try to train the model and on the second part (i.e. the data which is unseen for the model), you make the prediction and check how well your model works on it. If the model works with good accuracy on your test data it means that the model has not overritted the training data and can be trusted with the prediction, whereas if it performs with bad accuracy then our model is not to be trusted and we need to tweak our algorithm

(b) Describe one advantage and one disadvantage of using cross-validation.

ADVANTAGES	DISADVANTAGES
Reduces overfitting as data is split into multiple folds and train the algorithm on different folds	It drastically increases the training time
Helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm	Cross Validation is computationally very expensive in terms of processing power required