MACHINE LEARNING

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
   Ans: - **(C) between -1 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**
   Ans: - 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
   Ans: - C) Hyperplane

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**
   Ans: - 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' willbe?
   (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
   Ans: - C) old coefficient of 'X' ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy ofthe model?
   A) remains same B) **increases**
   C) decreases D) none of the above
   Ans: - B) Increases

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

MACHINE LEARNING

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

Ans: - **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?

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

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?

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 sugarlevels.

Ans: A, B, C, D

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

Ans: - A, B, D

MACHINE LEARNING

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 other words,Outliers are observations which are far enough away from the mean (or from nearly all of the data points) that they are noticeably different. Simple method to detect outlier is by Interquartile Range (IQR)

-1.5XIQR to 1.5 X IQR

The interquartile range is just the width of the box in the box-and-whisker plot. That is, IQR = Q3 – Q1. The IQR can be used as a measure of how spread-out the values are.

The IQR tells how spread out the "middle" values are; it can also be used to tell when some of the other values are "too far" from the central value. These "too far away" points are called "outliers", because they "lie outside" the range in which we expect them.

The IQR is the length of the box in box-and-whisker plot. An outlier is any value that lies more than one and a half times the length of the box from either end of the box.

Example:

Sample: 4, 7, 9, 11, 12, 20 (arranged in ascending order)
Divide sample into 2 so lower half is 4, 7, 9 and upper half is 11, 12, 20
Find Median of lower and upper half which is 7 and 12 respectively
So Q1 = 7 and Q3 = 12
IQR = Q3 - Q1 which is 5 (12 -7 = 5)
Outliers: a = Q1 - (1.5*IQR) =0.5; b = Q3 = Q3 + (1.5*IQR) =19.5
any number < a or > b is an outlier.
Hence in our sample 20 is an outlier.

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

Ans: -

1. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.
2. In Bagging aim is to decrease variance, not bias also suitable for complex models and in boosting aim to decrease bias, not variance
3. Bagging is a parallel ensemble i.e each model is built independently and boosting sequential ensemble i.e., try to add new models that do well where previous models lack
4. an example of a tree-based method is random forest in bagging and an example of a tree-based method is gradient boosting in boosting.

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

Ans: -

MACHINE LEARNING

Adjusted R^2 in logistic regression: -
We use adjusted R-squared to compare the goodness-of-fit for regression models that contain differing numbers of independent variables.
Let's say we are comparing a model with five independent variables to a model with one variable and the five variable model has a higher R-squared
Is the model with five variables actually a better model, or does it just have more variables?
To determine this, just compare the adjusted R-squared values.
The adjusted R-squared adjusts for the number of terms in the model. Importantly, its value increases only when the new term improves the model fit more than expected by chance alone.
The adjusted R-squared value actually decreases when the term doesn't improve the model fit by a sufficient amount.

Adjusted R^2 Value Calculation:

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.
Every time you add an independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.
Mathematical formula of Adjusted R-squared value is given as:
Ajd rsquared = 1-{(n-1)/(n-k-1) *(RSS/TSS)

Where
n=total no of observation
k= no of features
RSS=squared sum of difference between actual observed value and predictive values
TSS =squared sum of difference between actual observed value and predictive values and mean of all values.

We can calculate the adjusted r2score using sklearn as follow:

```
import statsmodels.formula.api as sm
result = sm.ols(formula="Y ~ X1+ X2", data=df).fit()
result.rsquared, result.rsquared_adj
```

14. What is the difference between standardization and normalization?

Ans: -

Normalization	**Standardization**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.

MACHINE LEARNING

Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinmaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is an often called as Scaling Normalization	It is an often called as Z-Score Normalization.

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

Ans: -

Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:

Advantages of Cross Validation

a. **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.
b. **Hyperparameter Tuning:** Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

a. **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. For example, if you go with 5-Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.
b. Needs Expensive Computation: Cross Validation is computationally very expensive in

MACHINE LEARNING

terms of processing power required.