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
   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 : A) linear

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’ 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

Ans: A)2.205 × old coefficient of ‘X’

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

Ans: B)increases

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

Ans: a) Random Forests reduce overfitting

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?

MACHINE LEARNING

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 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
ANS : ALL

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)max_depth

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 observation which differs from an overall pattern on a sample dataset is called an outlier.

Outliers:
The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.
However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.
Significance of outliers:

- Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results.
- Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.
- Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

If a dataset has *2n / 2n+1* data points, then
Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.
IQR is the range between the first and the third quartiles namely Q1 and Q3: *IQR = Q3 – Q1*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

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

ANS :

Bagging is a homogeneous weak learners’ model that learns from each other independently in parallel and combines them for determining the model average.

Boosting is also a homogeneous weak learners’ model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

Bagging and Boosting:

Differences

As we said already,
Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.
Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

MACHINE LEARNING

In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.
Models are built independently in Bagging. New models are affected by a previously built model's performance in Boosting.
In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models.

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

ANS

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 a independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

Adjusted R squared is calculated by **dividing the residual mean square error by the total mean square error** (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

14. What is the difference between standardisation and normalisation?

ANS:

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	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.

MACHINE LEARNING

S.NO.	Normalization	Standardization
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

ANS:

Definition. Cross-Validation is **a statistical method of evaluating and comparing learning algorithms by dividing data into two segments**: one used to learn or train a model and the other used to validate the model.

ADVANTAGE :

Cross-validation is **a statistical method used to estimate the performance (or accuracy) of machine learning models**. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

DISADVANTAGE :

The disadvantage of this method is that **the training algorithm has to be rerun from scratch k times**, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.