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
   E) None of These
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**.

Answer - an outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset you're working with.

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.

What is Interquartile Range IQR?
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?**
Answer –

S.NO	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.

MACHINE LEARNING

S.NO	Bagging	Boosting
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8.	In this base, classifiers are trained parallelly.	In this base, classifiers are trained sequentially.
9	Example: The Random Forest model uses Bagging.	Example: The AdaBoost uses Boosting techniques

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

Answer – Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific effect does not improve the model.

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

Adjusted R^2 is always less than or equal to R^2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value. In the real world, adjusted R^2 lies between these values.

14. What is the difference between standardisation and normalisation?

Answer -

Normalisation	Standardisation
Scaling is done by the highest and the lowest values.	Scaling is done by mean and standard deviation.
It is applied when the features are of separate scales.	It is applied when we verify zero mean and unit standard deviation.
Scales range from 0 to 1	Not bounded
Affected by outliers	Less affected by outliers

MACHINE LEARNING

It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed
It is also known as Scaling Normalization	It is also known as Z-Score

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

Answer –

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 - One of the advantages of cross validation is the it 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 – One of the disadvantages of cross validation is increasing 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.