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)
2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation   B) PCA
   C) Recursive feature elimination   D) Ridge Regularisation
   Ans: b)
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)
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)
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)
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: 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)

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)
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

MACHINE LEARNING

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: A,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) max_depth

B) Max_features

C) 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 **a mathematical value in a set of data which is quite distinguishing from the other values**. In simple terms, outliers are values uncommonly far from the middle. Mostly, outliers have a significant impact on mean, but not on the median, or mode.

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.
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.

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

Ans: Bagging

Bagging is also known as bootstrap aggregation. It is the ensemble learning method that is generally used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement meaning that the single data points can be selected more than once.

After several data samples are generated, these weak models are trained separately and depend on the element of task regression or classification. For example, the average of those predictions yield a more efficient estimate.

Random Forest is an extension over bagging. It takes one more step to predict a random subset of records. It also creates a random selection of features instead of using all features to develop trees. When it can have several random trees, it is known as the Random Forest.

Bagging has also been leveraged with deep learning models in the finance market, automating critical functions, such as fraud detection, credit risk computations, and option pricing issues.

This research demonstrates how bagging between several machine learning techniques has been leveraged to create loan default risk. This study understands how bagging supports minimizing risk by avoiding credit card fraud within the banking and financial institutions.

Boosting

Boosting is another ensemble process to create a set of predictors. In another terms, it can fit consecutive trees, generally random samples, and at every phase, the objective is to solve net error from the previous trees.

Boosting is generally used to reduce the bias and variance in a supervised learning technique. It defines the family of an algorithm that changes weak learners (base learners) to strong learners. The weak learner is the classifiers that are correct only up to a small extent with the actual classification, while the strong learners are the classifiers that are well correlated with the actual classification.

MACHINE LEARNING

Bagging	Boosting
Objective to decrease variance, not bias.	Objective to decrease bias, not variance.
Each model is built independently.	New models are affected by the implementation of the formerly developed model.
It is the simplest way of connecting predictions that belong to a similar type.	It is a method of connecting predictions that belong to multiple types.
Bagging tries to tackle the over-fitting problem.	Boosting tries to reduce bias.
Several training data subsets are randomly drawn with replacement from the whole training dataset.	Each new subset includes the components that were misclassified by previous models.
Bagging can solve the over-fitting problem.	Boosting can boost the over-fitting problem.

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

Ans: 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.

R-squared, often written R^2, is the proportion of the variance in the response variable that can be explained by the predictor variables in a linear regression model.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all while a value of 1 indicates that the response variable can be perfectly explained without error by the predictor variables.

The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. It is calculated as:

Adjusted $R^2 = 1 – [(1-R^2)*(n-1)/(n-k-1)]$

where:

- R^2: The R^2 of the model
- n: The number of observations
- k: The number of predictor variables

MACHINE LEARNING

14. What is the difference between standardisation and normalisation?

Difference between Normalization and Standardization

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.
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:
Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.
The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Advantages of Cross Validation

1. 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.

Note: Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.

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

Disadvantages of Cross Validation

1. 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.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.