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**
3. Which of the following is not a kernel in Support Vector Machines?
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial
5. 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
6. 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**
7. 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
8. 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:

9. 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**
10. 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.
11. 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.
12. What is the primary difference between bagging and boosting algorithms?
13. What is adjusted R^2 in linear regression. How is it calculated?
14. What is the difference between standardisation and normalisation?
15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
16. Describe one advantage and one disadvantage of using cross-validation.

11. IQR method of identifying outliers to set up a “fence” outside of Q1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q1 and add this value to Q3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. This is the method that Minitab uses to identify outliers by default.

12.

Similarities	Differences
Both are ensemble methods to get N learners from 1 learner…	… but, while they are built independently for Bagging, Boosting tries to add new models that do well where previous models fail.
Both generate several training data sets by random sampling…	… but only Boosting determines weights for the data to tip the scales in favor of the most difficult cases.
Both make the final decision by averaging the N learners (or taking the majority of them)…	… but it is an equally weighted average for Bagging and a weighted average for Boosting, more weight to those with better performance on training data.
Both are good at reducing variance and provide higher stability…	… but only Boosting tries to reduce bias. On the other hand, Bagging may solve the over-fitting problem, while Boosting can increase it.

13. It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_e}{SS_{\text{tot}}/\text{df}_t}$$

14. Normalized Data Vs Standardized Data

- Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.
- Normalization scales in a range of [0,1] or [-1,1]. Standardization is not bounded by range.
- Normalization is highly affected by outliers. Standardization is slightly affected by outliers.
- Normalization is considered when the algorithms do not make assumptions about the data distribution. Standardization is used when algorithms make assumptions about the data distribution.

15. Cross validation is a model evaluation method that is better than residuals. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on ``new'' data. This is the basic idea for a whole class of model evaluation methods called cross validation.

Advantages of Cross Validation

- 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.

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

Disadvantages of Cross Validation

- 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.

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

16. Advantages of Cross Validation

- 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.

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

Disadvantages of Cross Validation

- 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.

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