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: - C

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: - A

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: - C

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

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: - C

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

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, 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

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 **an observation that lies an abnormal distance from other values in a random sample from a population**. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. However, not all outliers bad. Some outliers signify that data is significantly different from others.

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.

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 values which separate the 4 equal parts.

- Q1 represent the 25^{th} percentile of the data.
- Q2 represent the 50^{th} percentile of the data.
- Q3 represent the 75^{th} percentile of the data.

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 + 1.5IQR are outliers.

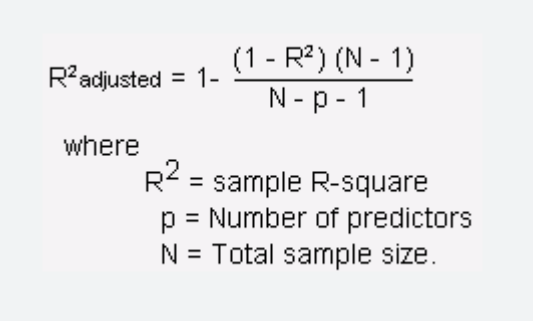
12. What is the primary difference between bagging and boosting algorithms?

Ans: -

- Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of prediction.
- Bagging decreases variance, not bias and solves overfitting issues in a model. Boosting decreases bias not variance.
- Bagging gives weight to each model, whereas in boosting technique, the new models are weighted based on their results. In boosting, new subsets of the data used for training contain observations that the previous model misclassified. Bagging uses randomly generated training data subsets.

13. What is adjusted R2 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.



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.

14. What is the difference between standardisation and normalisation?

Ans: -

- In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas, Standardisation assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis.
- **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.

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

Ans: - 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.

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.

Disadvantage 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.