ASSIGNMENT – 4

MACHINE LEARNING

In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:

Answer: C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?

Answer: C) Recursive feature elimination

3. Which of the following is not a kernel in Support Vector Machines?

Answer: C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

Answer: A) Logistic Regression

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)

Answer: C) old coefficient of 'X' ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

Answer: B) increases

7. Which of the following is not an advantage of using random forest instead of decision trees?

Answer: C) Random Forests are easy to interpret

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?

Answer: D) All of the above

9. Which of the following are applications of clustering?

Answer: All of the above

10. Which of the following is(are) hyper parameters of a decision tree?

Answer: A) max_depth B) max_features C) n_estimators

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: Outliers are data points that are significantly different from the rest of the data. They can have a significant impact on the results of statistical analyses and machine learning models, and therefore it is important to identify and handle them appropriately.

There are several methods for detecting outliers, one of which is the Inter Quartile Range (IQR) method. The IQR is a measure of the dispersion of the data, and it is calculated as the difference between the 75th percentile and the 25th percentile (also known as the first and third quartiles). The IQR is a robust measure of dispersion because it is not influenced by extreme values (i.e., outliers).

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

Answer: The primary difference between bagging and boosting algorithms is the way that they create an ensemble of models.

Bagging (short for "bootstrapped aggregation") is an ensemble learning method that involves training multiple models on different subsets of the data, and then aggregating their predictions. Bagging is used to reduce the variance of a single model, which can help to improve the stability and generalization of the model.

Boosting is an ensemble learning method that involves training multiple models sequentially, with each model trying to correct the mistakes of the previous model. Boosting is used to reduce the bias of a single model, which can help to improve the accuracy of the model.

In summary, bagging algorithms train models in parallel and combine their predictions to reduce variance, while boosting algorithms train models sequentially and combine their predictions to reduce bias.

13. What is adjusted R_2 in linear regression. How is it calculated?

Answer: In linear regression, the coefficient of determination (R2) is a measure of how well the model fits the data. It is calculated as the ratio of the sum of squared residuals (SSR) to the total sum of squares (SST), which is defined as 1 - (SSR/SST). The R2 value ranges from 0 to 1, with a value of 1 indicating a perfect fit and a value of 0 indicating that the model does not explain any of the variance in the data.

Adjusted R2 is an adjusted version of R2 that takes into account the number of independent variables in the model. It is calculated as 1 - (SSR/(SST * (n - 1)/(n - p - 1))), where n is the number of observations and p is the number of independent variables in the model. The adjusted R2 value is always less than or equal to the R2 value, and it is a more conservative estimate of the model fit because it penalizes models with a large number of independent variables.

The adjusted R2 value is often used to compare the fit of multiple regression models with different numbers of independent variables. By using the adjusted R2 value, we can more accurately compare the fit of models with different numbers of variables, because it takes into account the complexity of the model.

14. What is the difference between standardisation and normalisation?

Answer: Standardization and normalization are two techniques that are used to scale the variables of a dataset so that they have similar ranges and are comparable to each other. However, they differ in the way that the scaling is performed.

Standardization involves transforming the variables of a dataset to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the variable from each data point and dividing the result by the standard deviation. Standardization is often used when the variables of a dataset have different units of measurement or different scales, and we want to scale them so that they can be compared on the same scale.

Normalization, on the other hand, involves transforming the variables of a dataset to have a range between 0 and 1. This is done by subtracting the minimum value of the variable from each data point and dividing the result by the range (maximum value - minimum value). Normalization is often used when the variables of a dataset have a skewed distribution or large outliers, and we want to scale them so that they have a more symmetrical distribution.

In summary, standardization scales the variables of a dataset so that they have a mean of 0 and a standard deviation of 1, while normalization scales the variables to have a range between 0 and 1. Both techniques are useful for scaling the variables of a dataset so that they can be compared, but they differ in the way that the scaling is performed.

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

Answer: Cross-validation is a resampling technique that is used to evaluate the performance of a machine learning model. It involves dividing the dataset into a training set and a test set, and training the model on the training set. The model is then evaluated on the test set, and the performance is measured using a metric such as accuracy, precision, or recall.

One advantage of using cross-validation is that it allows you to estimate the generalization performance of the model, which is how well the model is expected to perform on unseen data. By evaluating the model on a hold-out test set, you can get a more accurate estimate of the model's generalization performance compared to evaluating the model on the training set.

One disadvantage of using cross-validation is that it can be computationally expensive, especially if the dataset is large. Because the model is trained and

evaluated multiple times, cross-validation requires more computing resources compared to evaluating the model on a single train-test split.

In summary, cross-validation is a resampling technique that is used to evaluate the performance of a machine learning model on unseen data. It has the advantage of providing a more accurate estimate of the model's generalization performance, but it can be computationally expensive.