**WORKSHEET -4**

**MACHINE LEARNING**

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

1.(C)

2.(D)

3.(C)

4.(A)

5.(A)

6.(B)

7.(C)

**In Q8 to Q10, more than one options are correct, Choose all the correct options:**

8.(D)

9.(A & D)

10.(A,B &D)

**Q11 to Q15 are subjective answer type questions**

11. What are outliers? Explain the Inter Quartile Range(IQR) method for outlier detection.

An outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. In other words, they’re unusual values in a dataset and are sometimes excluded from the data set

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?**

* Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.
* Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.
* In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.
* Models are built independently in Bagging. New models are affected by a previously built model’s performance in Boosting.
* Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.

13. What is adjusted R2 in logistic regression. How is it calculated?

It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes for adding independent variable that do not help in predicting the dependent variable. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

R2 adjusted = 1 – (1-R2)(N-1)/(N-p-1)

R2 – Sample R square; p – number of predictors; N – total sample size

Adjusted R-squared increases only when independent variable is significant and affects dependent variable.

**14. What is the difference between standardisation and normalisation?**

**Normalisation Standardisation**

|  |  |  |
| --- | --- | --- |
| 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. |

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

It is a method to ensure that important data is not left out during training. For example, If we train on 70% data there may be chances that some important data was present in the remaining 30% data. So, we try to eliminate this by using cross-validation; where we select some other data as training in the nex t iteration and the remaining for testing. Finally, a statistic measure like mean of all such iterations is taken.

**Advantage:**

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..

**Disadvantage:**

Cross Validation is computationally very expensive in terms of time and processing power required.