MACHINE LEARNING ASSIGNMENT – 4

1. C

2. C

3. A

4. A

5. C

6. B

7. B

8. B, C

9. A, D

10. A

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

An outlier is an individual point of data that is distant from other points in the dataset. It is an anomaly in the dataset that may be caused by a range of errors in capturing, processing or manipulating data. Outliers can skew overall data trends.

IQR is the range between the first and the third quartiles namely Q1 and Q3. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

IQR = Q3 – Q1

Q1 represents the **25th** percentile of the data.

Q2 represents the **50th** (**mean**) percentile of the data.

Q3 represents the **75th** percentile of the data.

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

Bagging: It is a homogeneous weak learners model that learns from each other independently in parallel and combines them for determining the model average.

Boosting: It is also a homogeneous weak learners model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

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

The **Adjusted R-squared** takes into account the number of independent variables used for predicting the target variable.

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.

Adjusted R2 is always less than or equal to R2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value. In the real world, adjusted R2 lies between these values.

14. What is the difference between standardisation and normalisation?

Standardisation-

It is used when we want to ensure zero mean and unit standard deviation. It is much less affected by outliers. It is useful when the feature distribution is Normal or Gaussian. It is an often called as Z-Score Normalization and it is not bounded to a certain range.

Normalization-

It is used when features are of different scales. It scales values between (0, 1) or (-1, 1). It is really affected by outliers. It is useful when we don’t know about the distribution.

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

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. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation.

Advantage- Cross-validation gives the idea about how the model will generalize to an unknown dataset

Disadvantage- with cross-validation, we need to train the model on multiple training sets.