Machine Learning Assignment - 4

1. C. between -1 and 1
2. D. Ridge Regularisation
3. A. linear
4. A. Logistic Regression
5. A. 2.205 × old coefficient of ‘X’
6. B. increases
7. C. Random Forests are easy to interpret
8. B. C.
9. A. B. C. D.
10. A. B. D.

11. 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.

    Inter Quartile Range (IQR) method for outlier detection:

    We can use the IQR method of identifying outliers to set up a “boundary” outside of Q1 and Q3. Any values that fall outside of this boundary 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.

12. Difference between bagging and boosting algorithms :

- Bagging tries to solve the over-fitting problem (Aim to decrease variance, not bias) whereas Boosting tries to reduce bias (Aim to decrease bias, not variance).
- In Bagging each model receives equal weight. While in Boosting models are weighted according to their performance.
- In Bagging each model is built independently whereas in Boosting new models are influenced by the performance of previously built models.

13. Adjusted R2 is a corrected model accuracy for linear models. It identifies the percentage of variance in the target field that is explained by the input(s).

    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. Differences between standardisation and normalisation :
    - 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 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 technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data.

    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 : Increases training time

    Cross Validation drastically increases the training time. Earlier we had to train your model only on one training set, but with Cross Validation we have to train our model on multiple training sets.