Machine Learning – 4

1. (C) Between -1 and 1
2. (D) Ridge Regularization
3. (A) Linear
4. (A) Logistic Regression
5. (C) (Old coefficient of x) divided by 2.205
6. (B) Increases
7. (C) Random Forests are easy to interpret
8. (B) and (C)
9. (A), (B), (C) and (D)
10. (B) and (D)

11. Outliers are data points which differ significantly from other data points. They can cause serious problems in statistical analysis. We may or may not remove them depending on the problem.

    There are three quartiles –
    a) Q1 (First quartile) – it represents 25th percentile of the data
    b) Q2 (Second quartile) – it represents 50th percentile of the data
    c) Q3 (Third quartile) – it represents 75th percentile of the data

    The difference between Q3 and Q1 is called Inter Quartile Range (IQR).

    $$IQR = Q3 - Q1$$

    The data points which lies below Q1-1.5*IQR or above Q3+1.5*IQR are outliers.

12. Bagging attempts to tackle the over-fitting issue (reduce variance); while boosting tries to reduce bias.

13. Adjusted R2 shows how well the data points fit a curve or line and adjusts for the number of terms in a model. That means if we add more and more useless variables to a model, adjusted R2 will decrease. Its value is always less than or equal to R2. It penalises extra (or unnecessary feature).

$$R^2 Adjusted = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

14. Normalisation (or Min-Max scaling) is used to transform features to be on a similar scale.

$$x_{new} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

    Standardization (or Z-score Normalization) is the transformation of features by subtracting from mean and dividing by standard deviation. This is called Z-score:

$$x_{new} = \frac{x - mean}{std\ deviation}$$

15. Cross Validation is a technique for assessing how the results of a statistical analysis will generalise to an independent set of data.

    Advantage: It 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: It needs expensive computation in terms of processing power required.