Machine Learning Worksheet-4

1. C
2. D
3. C
4. A
5. C
6. B
7. B
8. D
9. C, B,A,D
10. A,B,D
11. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. It can be because of typo error, miscalculation and random error.
    IQR: Interquartile range is the difference between 3rd and 1st quartiles.
    IQR=Q3-Q1
    Data points lying below Q1-1.5IQR and data points lying after Q3+1.5IQR are considered as outliers.

12.

Bagging	**Boosting**
1. Bagging is a method of merging the same type of predictions.	Boosting is a method of merging different types of predictions.
2. Bagging decreases variance, not bias, and solves over-fitting issues in a model.	Boosting decreases bias, not variance.
3. In Bagging, each model receives an equal weight.	In Boosting, models are weighed based on their performance.
4. Models are built independently in Bagging.	New models are affected by a previously built model's performance in Boosting.

12. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

    Adjusted r2=1-(SSres/dfe)/(SStot/dft)

13. **Normalization**: Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

    Standardization: Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

14. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

    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. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

    Disadvnatage:
    Increases Training Time: Cross Validation drastically increases the training time. Earlier you had to train your model only on one training set, but with Cross Validation you have to train your model on multiple training sets.

End of Document