MACHINE LEARNING ASSIGNMENT – 4

1. **C**
2. **C**
3. **A**
4. **A**
5. **A**
6. **C**
7. **A**
8. **B,C**
9. **B,C**
10. **B,D**
11. Outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In other words, they are the unusual values in a dataset. Inter Quartile Range (IQR) has the formula IQR=Q3-Q1. IQR is the range between first and third quartile range. The datapoints which falls below Q1-1.5 and above Q3+1.5 are outliers.

    The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset.

    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 whichseparate 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

    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.

12. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

    Bagging is the method of merging same type of predictions and Boosting is a method of merging different types of predictions.

In Bagging the result is obtained by averaging the responses of the N learners. Boosting assigns a second set of weights, this time for the N classifiers in order to takea weighted average of their estimates.

13. Adjusted R squared is a modified version of Rsquared that has been adjusted for the number of predictors in the model. Adjusted Rsquared value can be calculated based on value of Rsquared. Every time you add an independent variable to a model, R squared increases, even if the independent variable is insignificant.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the Rsquared increases, even if the independent variable is insignificant. It never declines.

The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. It is calculated as:

Adjusted R2 = 1 – [(1-R2)*(n-1)/(n-k-1)]

14. 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 is another scaling technique where the values are centered 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.

Normalization: rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost. Xchanged = X−XminXmax−XminXchanged=X−XminXmax−Xmin

Standardization : rescales data to have a mean (μμ) of 0 and standard deviation (σσ) of 1 (unit variance). Xchanged = X−μσ

15. Cross validation is a technique for assessing how statistical analysis generalizes to an independent dataset. It is a technique for evaluating machine learning modelsby training several models on subsets of the available input data and evaluating them on complementary subset of data.

Advantage: More accurate estimate of out-of-sample accuracy.

Disadvantage:

Cross-validation is computationally very expensive as we need to train on multiple training sets.

The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.