MACHINE LEARNING WORSHEET SOLUTIONS:

1) C
2) C
3) C
4) A
5) A
6) B
7) C
8) D
9) A & D
10) A, B & D

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

An observation which differs from an overall pattern on a sample dataset is called an outlier.
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 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.

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

In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates.

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

Adjusted R-Squared measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom.

Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

Question 14) What is the difference between standardisation and normalisation?

In normalisation Minimum and maximum value of features are used for scaling, In standarisation Mean and standard deviation is used for scaling.

Normalisation is used when features are of different scales. Standarization is used when we want to ensure zero mean and unit standard deviation.

In normalisation Scales values between [0, 1] or [-1, 1]. standarisation is not bounded to a certain range.

Normalisation is really affected by outliers. Standarisation is much less affected by outliers.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization. Scikit-Learn provides a transformer called StandardScaler for standardization.

Normalisation transformation squishes the n-dimensional data into an n-dimensional unit hypercube. Standarisation translates the data to the mean vector of original data to the origin and squishes or expands.

Normalisation is useful when we don't know about the distribution. Standarisation is useful when the feature distribution is Normal or Gaussian.

Normalisation is a often called as Scaling Normalization. Standarisation is a often called as Z-Score Normalization.

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

Cross validation defined as:

"A statistical method or a resampling procedure used to evaluate the skill of machine learning models on a limited data sample."

Advantages of Cross Validation

1. 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.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

1. 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.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.