# MACHINE LEARNING-4

**In Q1 to Q7, only one option is correct, Choose the correct option:**

**1)C, 2)B, 3)A, 4)C, 5)B, 6)B, 7)C**

**In Q8 to Q10, more than one options are correct, Choose all the correct options:**

**8)D, 9)B,D, 10)A,B**

**Q11 to Q15 are subjective answer type questions, Answer them briefly**.

**Q11) What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.**

An outlier is an observation that lies abnormally far away from other values in a dataset. Outliers can be problematic because they can affect the results of an analysis.

The interquartile range, often abbreviated IQR, is the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a dataset. It measures the spread of the middle 50% of values.

One popular method is to declare an observation to be an outlier if it has a value 1.5 times greater than the IQR or 1.5 times less than the IQR.

**Q12) What is the primary difference between bagging and boosting algorithms?**

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions. Bagging decreases variance, not bias, and solves over-fitting issues in a model**.**

**Q13) What is adjusted R2 in linear regression. How is it calculated?**

The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not. To understand adjusted R-squared, an understanding of R-squared is required.

The formula to calculate the adjusted R square of regression is below:

$$R^2 = \{(1 / N) * \Sigma [(x_i – x) * (Y_i – y)] / (\sigma_x * \sigma_y)\}^2$$

Where,

- $R^2$= adjusted R square of the **regression equation**
- N= Number of observations in the regression equation
- $X_i$= Independent variable of the regression equation
- X= Mean of the independent variable of the regression equation
- $Y_i$= Dependent variable of the regression equation
- Y= **Mean** of the dependent variable of the regression equation
- $\sigma_x$ = Standard deviation of the independent variable
- $\sigma_y$ = Standard deviation of the dependent variable.

**Q14) What is the difference between standardisation and normalisation?**

**Feature scaling** is one of the most important data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.
Tree-based algorithms are fairly insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster.

**Standardization or Z-Score Normalization** is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

Standardization can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Geometrically speaking, it translates the data to the mean vector of original data to the origin and squishes or expands the points if std is 1 respectively. We can see that we are just changing mean and standard deviation to a standard normal distribution which is still normal thus the shape of the distribution is not affected.

**Q15) What is cross-validation? Describe one advantage and one disadvantage of cross-validation.**

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

**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 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.