**Machine Learning 4**

1. C
2. B
3. A
4. A
5. A

6) B

7) A

8) B

9) B

10) D

11) 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 which separate 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.

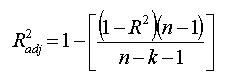
Q3 = median of n highest data points.

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 the simplest way of combining predictions that belong to the same type. Bagging aims to decrease variance, not bias.

**Boosting** is a way of combining predictions that belong to the different types. Boosting aims to decrease bias, not variance.

13) You only need R2 when working with samples. In other words, R2 isn’t necessary when you have data from an entire population.



where:

N is the number of points in your data sample.

K is the number of independent regressors, i.e. the number of variables in your model, excluding the constant.

Meaning of Adjusted R2

Both R2 and the adjusted R2 give you an idea of how many data points fall within the line of the regression equation. However, there is one main difference between R2 and the adjusted R2: R2 assumes that every single variable explains the variation in the dependent variable. The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

14) Normalization:

1. Minimum and maximum value of features are used for scaling.
2. It is used when features are of different scales.
3. Scales values between [0, 1] or [-1, 1].
4. It is really affected by outliers.

Standardization:

1. Mean and standard deviation is used for scaling.
2. It is used when we want to ensure zero mean and unit standard deviation.
3. It is not bounded to a certain range.
4. It is much less affected by outliers

15) Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

Advantages:

Checking Model Generalization: Cross-validation gives the idea about how the model will generalize to an unknown dataset

Checking Model Performance: Cross-validation helps to determine a more accurate estimate of model prediction performance

Disadvantages:

Higher Training Time: with cross-validation, we need to train the model on multiple training sets.

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