Machine Learning
Answer sheet

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

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

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

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. What is the primary difference between bagging and boosting algorithms?

Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.

Bagging tries to solve the over-fitting problem. Boosting tries to reduce bias.

13. What is adjusted R^2 in linear regression. How is it calculated?

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 an independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines.

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

14. **What is the difference between standardisation and normalisation?**

Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution. Normalization scales in a range of [0,1] or [-1,1]. Standardization is not bounded by range. Normalization is highly affected by outliers.

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

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 of Cross Validation

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.

Disadvantages of Cross Validation

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.