ASSIGNMENT -4

MACHINE LEARNING

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

1. The value of correlation coefficient will always be:

C) between -1 and 1

2. Which of the following cannot be used for dimensionality reduction?

C) Recursive Feature elimination

3. Which of the following is not a kernel in Support Vector Machines?

A) Liner

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundries

D) Support Vector Classifier

5. In a Linear Regression problem, 'X' is independent variable and 'Y' is dependent variable, where 'X' represents weight in pounds. If you convert the unit of 'X' to kilograms, then new coefficient of 'X' will be ?

C) old coefficient of 'X' ÷ 2.205

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

C) decreases

7. Which of the following is not an advantage of using random forest instead of decision trees?

A) Random Forests reduce overfitting

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

8. Which of the following are correct about Principal Components?

B) Principal Components are calculated using unsupervised learning techniques

9. Which of the following are applications of clustering?

All Are Correct

10. Which of the following is(are) hyper parameters of a decision tree?

A) max_depth . B) max_features D) min_samples_leaf

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

Ans. An outlier is simply a data point that is drastically different or distant from other data points. A set of data can have just one outlier or several. To be an outlier, a data point must not correspond with the general trend of the data set. It must be very noticeably outside the pattern. There are several types of outliers, including point outliers, contextual outliers, and collective outliers. A point outlier is when single data points fall outside the normal pattern of the distribution. A contextual outlier means the points' deviation is within the same context. Collective outliers can form their own patterns that lead to new discoveries. IQR Explanation.. 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.

Example:

Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

Step by step way to detect outlier in this dataset using **Python**:

Step 1: Import necessary libraries.

```
import numpy as np

import seaborn as sns
```

Step 2: Take the data and sort it in ascending order.

```
data = [6, 2, 3, 4, 5, 1, 50]

sort_data = np.sort(data)

sort_data
```

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

Bagging	Boosting
Each model is built independently.	Objective to decrease bias, not variance

Each model is built independently.	New models are affected by the implementation of the formerly developed model.
It is the simplest way of connecting predictions that belong to a similar type.	It is a method of connecting predictions that belong to multiple types.
Bagging tries to tackle the over-fitting problem.	Boosting tries to reduce bias.
Several training data subsets are randomly drawn with replacement from the whole training dataset.	Each new subset includes the components that were misclassified by previous models.
Bagging can solve the over-fitting problem	Boosting can boost the over-fitting problem.

13._What is adjusted R2 in linear regression. How is it calculated?

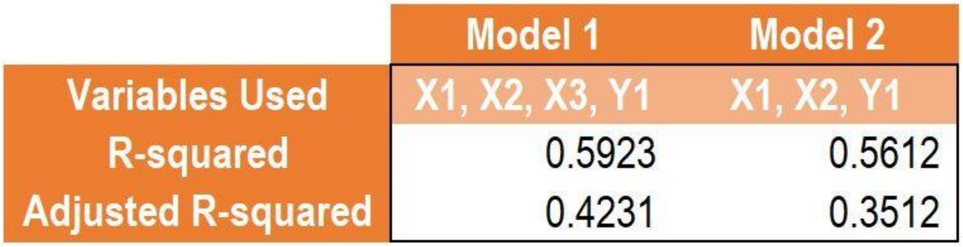
Ans. The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not. To understand adjusted R-squared, an understanding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

Example of the Adjusted R-squared

Consider two models:

- Model 1 uses input variables X1, X2, and X3 to predict Y1.
- Model 2 uses input variables X1 and X2 to predict Y1.

Which model should be used? Information regarding both models are provided below:



	Model 1	Model 2
Variables Used	X1, X2, X3, Y1	X1, X2, Y1
R-squared	0.5923	0.5612
Adjusted R-squared	0.4231	0.3512

Comparing the R-squared between Model 1 and Model 2, the R-squared predicts that Model 1 is a better model as it carries greater explanatory power (0.5923 in Model 1 vs. 0.5612 in Model 2).

Comparing the R-squared between Model 1 and Model 2, the adjusted R-squared predicts that the input variable X3 contributes to explaining output variable Y1 (0.4231 in Model 1 vs. 0.3512 in Model 2).

As such, Model 1 should be used, as the additional X3 input variable contributes to explaining the output variable Y1.

14. What is the difference between standardisation and normalisation?

Ans.

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

S.NO.	Normalization	Standardization
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Ans. 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.

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.