Machine Learning Assignment-4 Answers :-

Ans.1 :- between -1 and 1

Ans.2 :- Recursive feature elimination

Ans.3 :- hyperplane

Ans.4 :- Logistic Regression

Ans.5 :- Cannot be determined

Ans.6 :- increases

Ans.7 :- Random Forests are easy to interpret

Ans.8 :- Principal Components are calculated using unsupervised learning techniques.

Principal Components are linear combinations of Linear Variables.

Ans.9 :- Identifying spam or ham emails.

Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

Ans.10 :- max_depth

max_features

min_samples_leaf

Ans.11 :-

Outliers:-

- 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.

- However, not all outliers are bad. Some outliers signify that data is significantly different from others.

- For example, it may indicate an anomaly like bank fraud or a rare disease.

IQR method for outlier detection:-

- 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.

Ans.12 :-

primary difference between bagging and boosting algorithms :-

Bagging:-

- The simplest way of combining predictions that belong to the same type.
- Aim to decrease variance, not bias.
- Each model receives equal weight.
- Each model is built independently.
- Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.
- Bagging tries to solve the over-fitting problem.
- If the classifier is unstable (high variance), then apply bagging.
- In this base classifiers are trained parallelly.
- Example: The Random forest model uses Bagging.

Boosting:-

- A way of combining predictions that belong to the different types.
- Aim to decrease bias, not variance.
- Models are weighted according to their performance.
- New models are influenced by the performance of previously built models.

- Every new subset contains the elements that were misclassified by previous models.

- Boosting tries to reduce bias.
- If the classifier is stable and simple (high bias) the apply boosting.

- In this base classifiers are trained sequentially.
- Example: The AdaBoost uses Boosting techniques

Ans.13 :-

adjusted R2 in linear regression and how it's calculated :-

- It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable.

- It penalizes you for adding independent variable that do not help in predicting 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.

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_e}{SS_{\text{tot}}/\text{df}_t}$$

- In the above equation, dft is the degrees of freedom n– 1 of the estimate of the population variance of the dependent variable, and dfe is the degrees of freedom n – p – 1 of the estimate of the underlying population error variance.

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

- **Adjusted R Squared = 1 – [((1 – R^2) * (n – 1)) / (n – k – 1)]**

Where,

R^2 = sample R-square

p = Number of predictors

N = Total sample size

Ans.14 :-

Difference between Normalization and Standardization :-

Normalization :-

- Minimum and maximum value of features are used for scaling.
- It is used when features are of different scales.
- Scales values between [0, 1] or [-1, 1].
- It is really affected by outliers.
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
- This transformation squishes the n-dimensional data into n-dimensional unit hypercube.
- It is useful when we don't know about the distribution.
- It is a often called as Scaling Normalization.

Standardization :-

- Mean and standard deviation is used for scaling.
- It is used when we want to ensure zero mean and unit standard deviation.
- It is not bounded to a certain range.
- It is much less affected by outliers.
- Scikit-Learn provides a transformer called StandardScaler for standardization.
- It translates the data to the mean vector of original data to the origin and squishes or expands.
- It is useful when the feature distribution is Normal or Gaussian.
- It is often called as Z-Score Normalization.

Ans.15 :-

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.
- In machine learning, we couldn't fit the model on the training data and can't say that the model will work accurately for the real data.
- For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise.
- For this purpose, we use the cross-validation technique.

- The three steps involved in cross-validation are as follows :

  1. Reserve some portion of sample data-set.
  2. Using the rest data-set train the model.
  3. Test the model using the reserve portion of the data-set.

- Methods used for Cross-Validation

  There are some common methods that are used for cross-validation. These methods are given below:

  1. Validation Set Approach
  2. Leave-P-out cross-validation
  3. Leave one out cross-validation
  4. K-fold cross-validation
  5. Stratified k-fold cross-validation

Advantage and Disadvantage of using Cross-Validation :-

Advantages:-

1. More accurate estimate of out-of-sample accuracy.
2. More "efficient" use of data as every observation is used for both training and testing.

Disadvantages:-

1. Increases training time.
2. Needs expensive computation.