Machine Learning Assignment-4

Q.1. Answer- C) between -1 and 1

Q.2. Answer- D) Ridge Regularisation

Q.3. Answer- A) linear

Q.4. Answer- B) Naïve Bayes Classifier

Q.5. Answer- A) 2.205 × old coefficient of 'X'

Q.6. Answer- C) decreases

Q.7. Answer- B) Random Forests explains more variance in data then decision trees

Q.8. Answer-

B) Principal Components are calculated using unsupervised learning techniques

C) Principal Components are linear combinations of Linear Variables.

Q.9. Answer-

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index

C) Identifying spam or ham emails

Q.10. Answer-

A) max_depth

B) max_features

D) min_samples_leaf

Q.11. Answer-

Outliers- An outlier is a data point that differs significantly from other observations or an overall pattern on a sample dataset. An outlier may be due to variability in the measurement or it may indicate experimental error, the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results. Most machine learning algorithms do not work well in the presence of outlier. So, it is desirable to detect and remove outliers. Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

IQR- It 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.

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.

Q.12. Answer- The primary difference between bagging and boosting algorithms are:

Bagging:

- Bagging is a method of merging the same type of predictions.
- Bagging decreases variance, not bias, and solves over-fitting issues in a model.
- In Bagging, each model receives an equal weight.
- Models are built independently in Bagging.
- In Bagging, training data subsets are drawn randomly with a replacement for the training dataset.

Boosting:

- Boosting is a method of merging different types of predictions.
- Boosting decreases bias, not variance.
- In Boosting, models are weighed based on their performance.
- New models are affected by a previously built model's performance in Boosting.
- In Boosting, every new subset comprises the elements that were misclassified by previous models.

Q.13. Answer- The adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. When we fit linear regression models, we often calculate the R-squared value of the model. The value for R-squared can range from 0 to 1 where:

- A value of 0 indicates that the response variable cannot be explained by the predictor variables at all.
- A value of 1 indicates that the response variable can be perfectly explained by the predictor variables.

R-squared will always increase when a new predictor variable is added to the regression model. It's possible that a regression model with a large number of predictor variables has a high R-squared value, even if the model doesn't fit the data well. So, to handle this there is an alternative to R-squared known as adjusted R-squared.

Formula:

Adjusted R2 = 1 – [(1-R2) * (n-1)/(n-k-1)]

where:

- R2: The R2 of the model
- n: The number of observations
- k: The number of predictor variables

Q.14. Answer-

Normalization:

- Normalization or Min-Max Scaling is used to transform features to be on a similar scale.
- 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.
- This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.
- It is useful when we don't know about the distribution.
- It is an often called as Scaling Normalization
- Formula: X_new = (X - X_min)/(X_max - X_min)

Standardization:

- Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation.
- 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.
- 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.
- Formula: X_new = (X - mean)/Std

Q.15. Answer- 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. The three steps involved in cross-validation are as follows:

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

Methods of Cross Validation:

- Validation Set Approach
- Leave-P-out cross-validation
- Leave one out cross-validation
- K-fold cross-validation
- Stratified k-fold cross-validation

The Advantages of CV are as follows:

- CV assists in realizing the optimal tuning of hyperparameters (or model settings) that increase the overall efficiency of the ML model,
- Training data is efficiently utilized as every observation is employed for both testing and training.
- More accurate estimate of out-of-sample accuracy.
- More “efficient” use of data as every observation is used for both training and testing.

The Disadvantages of CV are as follows:

- Increases Testing and Training Time: CV significantly increases the training time required for an ML model. This is due to the numerous test cycles to be implemented along with the test preparation and examining and analyzing of the results.
- Additional computation equates to additional resources required: CV is computationally expensive, requiring surplus processing power; add the first disadvantage of extra time, then this resource requirement will add further cost to an ML model project's budget.
- For the ideal conditions, it provides the optimum output. But for the inconsistent data, it may produce a drastic result. So, it is one of the big disadvantages of cross-validation, as there is no certainty of the type of data in machine learning.
- In predictive modeling, the data evolves over a period, due to which, it may face the differences between the training set and validation sets. Such as if we create a model for the prediction of stock market values, and the data is trained on the previous 5 years stock values, but the realistic future values for the next 5 years may drastically different, so it is difficult to expect the correct output for such situations.