MACHINE LEARNING

ASSIGNMENT - 4

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

1. The value of correlation coefficient will always be:

A) between 0 and 1 B) greater than -1
C) between -1 and 1 D) between 0 and -1
ANS: C

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

A) Lasso Regularisation B) PCA
C) Recursive feature elimination D) Ridge Regularisation
ANS:

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

A) linear B) Radial Basis Function
C) hyperplane D) polynomial
ANS: C

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

A) Logistic Regression B) Naïve Bayes Classifier
C) Decision Tree Classifier D) Support Vector Classifier
ANS: A

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?

(1 kilogram = 2.205 pounds)
A) 2.205 × old coefficient of 'X' B) same as old coefficient of 'X'
C) old coefficient of 'X' ÷ 2.205 D) Cannot be determined
ANS: A

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

A) remains same B) increases
C) decreases D) none of the above
ANS: B

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

A) Random Forests reduce overfitting
B) Random Forests explains more variance in data then decision trees
C) Random Forests are easy to interpret
D) Random Forests provide a reliable feature importance estimate
ANS: B

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?

A) Principal Components are calculated using supervised learning techniques
B) Principal Components are calculated using unsupervised learning techniques
C) Principal Components are linear combinations of Linear Variables.
D) All of the above

ANS: B, C

9. Which of the following are applications of clustering?

A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index
B) Identifying loan defaulters in a bank on the basis of previous years’ data of loan accounts.
C) Identifying spam or ham emails
D) Identifying different segments of disease based on BMI, blood pressure, cholesterol, blood sugar levels.

ANS:

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

A) max_depth B) max_features
C) n_estimators D) min_samples_leaf
ANS: C

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

ANS: an **outlier** is a data point that differs significantly from other observations. 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.
IQR is used to measure variability, by dividing the datasets into quartiles. The data is sorted in ascending order and split into 4 equal parts . Q1, Q2, Q3 called first, second , third quartiles are the values which separate the 4 equal parts.

Q1 represents the 25th percentile of data
Q2 represents the 50th percentile of data
Q3 represents the 75th percentile of data

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

- **ANS:** Bagging (or Bootstrap Aggregation), is a simple and very powerful ensemble method. Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees.
- The idea behind bagging is combining the results of multiple models (for instance, all decision trees) to get a generalized result. Now, bootstrapping comes into picture.
- Bagging (or Bootstrap Aggregating) technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set.
- Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model.
- In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.
- When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. By combining the whole set at the end converts weak learners into better performing model.

13. What is adjusted R_2 in linear regression. How is it calculated?

ANS: Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific effect does not improve the model.

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. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0 indicates a model that has no predictive value. In the real world, adjusted R^2 lies between these values.

14. What is the difference between standardisation and normalisation?

ANS: Normalisation is suitable to use when the data does not follow Gaussian Distribution principles. It can be used in algorithms that do not assume data distribution, such as K-Nearest Neighbors and Neural Networks.
On the other hand, standardisation is beneficial in cases where the dataset follows the Gaussian distribution. Unlike Normalization, Standardisation is not affected by the outliers in the dataset as it does not have any bounding range.

Applying Normalization or Standardisation depends on the problem and the machine learning algorithm. There are no definite rules as to when to use Normalization or Standardisation. One can fit the normalized or standardized dataset into the model and compare the two.
It is always advisable to first fit the scaler on the training data and then transform the testing data. This would prohibit data leakage during the model testing process, and the scaling of target values is generally not required.

15. What is cross-validation? Describe one advantage and one disadvantage of using cross-validation.
ANS: Cross Validation in Machine Learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets. Below are some of the advantages and disadvantages of Cross Validation in Machine Learning:

Advantages of Cross Validation

1. 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.

Note: Chances of overfitting are less if the dataset is large. So, Cross Validation may not be required at all in the situation where we have sufficient data available.

2. Hyperparameter Tuning: Cross Validation helps in finding the optimal value of hyperparameters to increase the efficiency of the algorithm.

Disadvantages of Cross Validation

1. 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.

For example, if you go with 5 Fold Cross Validation, you need to do 5 rounds of training each on different 4/5 of available data. And this is for only one choice of hyperparameters. If you have multiple choice of parameters, then the training period will shoot too high.

2. Needs Expensive Computation: Cross Validation is computationally very expensive in terms of processing power required.