MACHINE LEARNING

WORKSHEET – 1

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

1. The value of correlation coefficient will always be: C
   A) between 0 and 1 B) greater than -1
   C) between -1 and 1 D) between 0 and -1
2. Which of the following cannot be used for dimensionality reduction? D
   A) Lasso Regularisation B) PCA
   C) Recursive feature elimination D) Ridge Regularisation
3. Which of the following is not a kernel in Support Vector Machines? C
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial
4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries? C
   A) Logistic Regression B) Naïve Bayes Classifier
   C) Decision Tree Classifier 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? B
   (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
6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model? B
   A) remains same B) increases
   C) decreases D) none of the above
7. Which of the following is not an advantage of using random forest instead of decision trees? C
   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

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? D
   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
9. Which of the following are applications of clustering? All options
   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.
10. Which of the following is(are) hyper parameters of a decision tree? A,B & D
    A) max_depth B) max_features
    C) n_estimators D) min_samples_leaf

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.

    An outlier is an observation that is unlike the other observations. It is rare, or distinct, or does not fit in some way.

    Outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile.ie: below outliers is Q1-1.5*(IQR), and above Q3 + 1.5(IQR).

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.

In Baggiing each model receives equal weight whereas in Boosting models are weighted according to their performance.

In Bagging each model is built independently whereas in Boosting new models are influenced by performance of previously built models.

In Bagging different training data subsets are randomly drawn with replacement from the entire training dataset. In Boosting every new subsets contains the elements that were misclassified by previous models.

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

If the classifier is unstable (high variance), then we should apply Bagging. If the classifier is stable and simple (high bias) then we should apply Boosting.

Bagging is extended to Random forest model while Boosting is extended to Gradient boosting.

13. What is adjusted R2 in logistic regression. How is it calculated?

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.

The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

It decreases when a predictor improves the model by less than expected by chance.The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

Calculation:

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.

$$R2 = (1-R2)(N-1)/N-p-1$$

Where R2 = sample R sq.
P = Number of predictors
N = Total number of sample.

14. What is the difference between standardisation and normalisation?

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things.

Normalization usually means to scale a variable to have a values between 0 and 1,
while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

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.

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.

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