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?

B) PCA

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

A) linear

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

A) Logistic Regression

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

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

B) increases

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

B) Random Forests explains more variance in data then decision trees

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) All of the above

9. Which of the following are applications of clustering?

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) max_depth

C) n_estimators

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 obvervation that lies an abnormal distance from values in random sample from a population this definition leaves it up to the analyst to decide what will be considered abnormal Q3 and Q1.

Each quartile is a median calculated as follows, given an even 2n or odd 2n+1, number of values third quartile.

Q3=median of the n largest values.

The QR of a set of values is calculated as the difference between upper and lower quantities.

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

Bagging:

- Various training data subsets are randomly drawn with replacement from the whole training dataset.
- Bagging attempts to tackle the over-fitting issue.
- If the classifier is unstable (high variance), then we need to apply bagging.
- Every model receives an equal weight.
- Objective to decrease variance, not bias.
- It is the easiest way of connecting predictions that belong to the same type.
- Every model is constructed independently.

Boosting:

- Each new subset contains the components that were misclassified by previous models.
- Boosting tries to reduce bias.
- If the classifier is steady and straightforward (high bias), then we need to apply boosting.
- Models are weighted by their performance.
- Objective to decrease bias, not variance.
- It is a way of connecting predictions that belong to the different types.
- New models are affected by the performance of the previously developed model.

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

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 when 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.

Calculation :

R 2 = 1 − sum squared regression (SSR) total sum of squares (SST) , = $\mathbf{1 - \sum (y_i - y_i \hat{})^2 \sum (y_i - \bar{y})^2}$. The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared.

14. What is the difference between standardisation and normalisation?

Sl No	Normalization	Standardisation
1	Minimum and maximum value of fratures 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 calledMinMaxScaler for Normalization.	Scikit-Learn provides a transformer called standardScaler for standardization.
6.	This transformation squishes the n- dimensional data into n dimensional unit hypercube.	It translate the data to mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution.	It is useful when the feature distribution is normalization or Gaussian.

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

Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

The purpose of cross–validation is to test the ability of a machine learning model to predict new data. It is also used to flag problems like overfitting or selection bias and gives insights on how the model will generalize to an independent dataset.

The 4 Types of Cross Validation in Machine Learning are:

- Holdout Method
- K-Fold Cross-Validation
- Stratified K-Fold Cross-Validation
- Leave-P-Out Cross-Validation

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.