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?

C) Recursive feature elimination

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

C) hyperplane

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

C) Decision Tree 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? (1 kilogram = 2.205 pounds)

B) same as old coefficient of ‘X’

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?

C) Random Forests are easy to interpret

8. Which of the following are correct about Principal Components?

D) All of the above

9. Which of the following are applications of clustering?

B) Identifying loan defaulters in a bank on the basis of previous years' data of loan accounts.

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

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.

Answer: An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. 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.

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

Sl.No.	Bagging Algorithms	Boosting Algorithms
1	The simplest way of combining predictions that belong to the same group	A way of combining the predictions that belong to different applications
2	It decreases variance not bias	It decreases bias not variance
3	Each model receives equal weight	The model receives weights baed on the performance
4	It decreases overfitting	It decreases the bias
5	Model is built independently	New models are influenced by the performance of the previous models.
6	The base of the classifiers are trained parallelly	The base of the classifiers are trained parallelly are trained sequencially
7	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8	Ex: Random forest	Ex: Adaboost

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

Answer: Adjusted R2 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.R2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of effects are included in the model. Adjusted R2 attempts to correct for this overestimation. Adjusted R2 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 R2 is always less than or equal to R2. 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 R2 lies between these values. The coefficient of determination, or R2R2, is a measure that provides information about the goodness of fit of a model. In the context of regression it is a statistical measure of how well the regression line approximates the actual data. It is therefore important when a statistical model is used either to predict future outcomes or in the testing of

hypotheses. There are a number of variants (see comment below); the one presented here is widely used

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - y)^2}.$$

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. As it is a percentage it will take values between 0 and

14. What is the difference between standardisation and normalisation?

Sl.No.	Standardisation	Normalisation
1	Maximum and minimum value of features 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 called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the 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 Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Answer: 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 advantages are 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. The disadvantages are Increases Training Time and computing efficiency: 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.