ASSIGNMENT - 4

MACHINE LEARNING

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

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

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?

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?

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? (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?

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?

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?

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?

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) 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.

Answer:---**Outliers:-** A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data is called outliers. EX-- the scores 25,29,3,32,85,33,27,28 both 3 and 85 are "outliers".

Different method of outliers detection like Box plot, IQR, Z score method.

IQR method:-- IQR method is used by box plot to highlight outliers. IQR stands for interquartile range, which is the difference between q3 (75th percentile) and q1 (25th percentile). The IQR method computes lower bound and upper bound to identify outliers.

*Lower Bound = q1–1.5*IQR*

*Upper Bound = q3+1.5*IQR*

Any value below the lower bound and above the upper bound are considered to be outliers.

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

Answer:--Diffrence b/w bagging and boosting

Bagging	Boosting
1.the simplest way of combining predictions that belong to the same type.	1. A way of combining predictions that belong to the different types.
2. Aim to decrease variance, not bias.	2. Aim to decrease bias, not variance.
3. Each model receives equal weight.	3.Models are weighted according to their performance.
4. Each model is built independently.	4. New models are influenced by the performance of previously built models.
5. Bagging tries to solve the over-fitting problem.	5. Boosting tries to reduce bias

6. In this base classifiers are trained parallelly.	6. In this base classifiers are trained sequentially.

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

Answer:--

Adjusted R2-- 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 ..

Before we calculate adjusted r squared, we need r square first

Using Regression outputs

R2 = Explained Variation / Total Variation
R2 = MSS / TSS
R2= (TSS – RSS) / TSS
Where:
TSS – Total Sum of Squares = Σ (Yi – Ym)2
MSS – Model Sum of Squares = Σ (Y^ – Ym)2
RSS – Residual Sum of Squares =Σ (Yi – Y^)2
Y^ is the predicted value of the model, Yi is the ith value and Ym is the mean value
Adjusted R Squared = 1 – [((1 – R^2) * (n – 1)) / (n – k – 1)]

14. What is the difference between standardisation and normalisation?

Answer:--

Standardisation	Normalisation
1. Mean and standard deviation is used for scaling.	1. Minimum and maximum value of features are used for scaling.
2. It is not bounded to a certain range.	2.Scales values between [0, 1] or [-1, 1].

3. Scikit-Learn provides a transformer called StandardScaler for standardization.	3. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.
4. It is useful when the feature distribution is Normal or Gaussian.	4. It is useful when we don’t know about the distribution
5. It is much less affected by outliers.	5. It is really affected by outliers.
6. It is a often called as Z-Score Normalization	6 It is a often called as Scaling Normalization.

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

Answer:---

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.

Advantage:-- 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.

Disadvantage:-- The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.