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

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

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?

Answer

Bagging	Boosting
Various training data subsets are randomly drawn with replacement from the whole training dataset.	Each new subset contains the components that were misclassified by previous models.
Bagging attempts to tackle the over-fitting issue.	Boosting tries to reduce bias.
If the classifier is unstable (high variance), then we need to apply bagging.	If the classifier is steady and straightforward (high bias), then we need to apply boosting.
Every model receives an equal weight.	Models are weighted by their performance.
Objective to decrease variance, not bias.	Objective to decrease bias, not variance.
It is the easiest way of connecting predictions that belong to the same type.	It is a way of connecting predictions that belong to the different types.
Every model is constructed independently.	New models are affected by the performance of the previously developed model.

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

Answer

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.

14. What is the difference between standardisation and normalisation?

Answer

Normalization (MinMaxScaler) typically means rescales the values into a range of [0,1]. Standardization (StandardScaler) typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Answer

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model or a statistical method used to estimate the performance (or accuracy) of machine learning models.

Advantage - It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

Disadvantage - In this method 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.