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

Answer: Option C – between -1 to 1

2. Which of the following cannot be used for dimensionality reduction?
   A) Lasso Regularisation
   B) PCA
   C) Recursive feature elimination
   D) Ridge Regularisation

Answer: Option 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

Answer: Option C – hyperplane

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

Answer: Option 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

Answer: Option 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

Answer: Option C – decreases

MACHINE LEARNING

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

Answer: Option C – Random Forests are easy to interpret

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

Answer: Option B – Principal Components are calculated using supervised learning techniques
Option C – Principal Components are calculated using unsupervised learning techniques

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.

   Answer: Option A,B,C,D

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

Answer: Option A – max_depth, Option D – min_samples_leaf

MACHINE LEARNING

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 a data point that lies outside the overall pattern in a distribution.

The interquartile range is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the firstquartile (Q_1). IQR = $Q_3 – Q_1$. The IQR can help determine outliers.

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

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

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.

13. What is adjusted R^2 in linear regression. How is it calculated?

Answer: 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. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size. Every time you add a independent variable to a model, the R-squared increases, even if the independent variable is insignificant.

14. What is the difference between standardisation and normalisation?

Answer:

Normalization	Standardization
Feature scaling method to bring the data into common range such as [0, 1], [-1, 1], etc.	Feature scaling method bring the data with mean 0 and unit variance
Scikit-learn provides MinMaxScaler, MaxAbsScaler and RobustScaler methods for normalization	Scikit-learn provides StandardScaler for standardization
MinMaxScaler and MaxAbsScaler are sensitive to outliers whereas RobustScaler is more robust to outliers	Standardization is less sensitive to outliers compared to MinMaxScaler and MaxAbsScaler
Useful when we don't know about the distribution of features and there are no or little outliers - MinMaxScaler: if features don't follow normal distriubtion and if there are no or less outliers - MaxAbsScaler: if the data is sparse - RobustScaler: if the data contains outliers	Useful when we know features are normally distributed (Gaussian distribution)

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

Answer: Cross-validation is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. Cross-validation is largely used in settings where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive model.

Advantage: Cross-validation is a technique that is used for the assessment of how the results of statistical analysis generalize to an independent data set. Cross-validation is largely used in settings where the target is prediction and it is necessary to estimate the accuracy of the performance of a predictive model.

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.