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:
   A) between 0 and 1 B) greater than -1
   C) between -1 and 1 D) between 0 and -1
   Answer – C) between -1 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
   Answer – C) Recursive feature elimination
3. Which of the following is not a kernel in Support Vector Machines?
   A) linear B) Radial Basis Function
   C) hyperplane D) polynomial
   Answer – 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 – 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 – A) 2.205 × old coefficient of ‘X’
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 – B) increases
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 – 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?
   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 –
   A) Principal Components are calculated using supervised learning techniques
   B) Principal Components are calculated using unsupervised learning techniques

FLIP ROBO

MACHINE LEARNING

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 –

   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

   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

    Answer - a) max_depth B) max_features

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 mathematical value in a set of data which is quite distinguishing from the other values**.**

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 –

1.While Bagging and boosting make seem similar due to the use of N learners in both techniques, They are inherently quite different. While the Bagging technique is a simple way of combining predictions of the same kind, boosting combines predictions that belong to different types.

2. n Bagging, each model is created independent of the other, But in boosting new models, the results of the previously built models are affected.

3. Bagging gives equal weight to each model, whereas in Boosting technique, the new models are weighted based on their results.

4. In boosting, new subsets of data used for training contain observations that the previous model misclassified. Bagging uses randomly generated training data subsets.

5. Bagging tends to decrease variance, not bias. In contrast, Boosting reduces bias, not variance.

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

Answer - The adjusted R-squared adjusts for the number of terms in the model. Importantly, its value increases only when the new term improves the model fit more than expected by chance alone. The adjusted R-squared value actually decreases when the term doesn't improve the model fit by a sufficient amount.

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).

14. What is the difference between standardisation and normalisation?

Answer –

Standardization – 1.Minimum and maximum value of features are used for scaling.

2. It is used when features are of different scales.

3. Scales values between [0, 1] or [-1, 1].

4. It is really affected by outliers.

5. Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

6. It is useful when we don’t know about the distribution.

7. It is a often called as Scaling Normalization.

MACHINE LEARNING

Normalisation –

1 Mean and standard deviation is used for scaling.

2. It is used when we want to ensure zero mean and unit standard deviation.

3. It is not bounded to a certain range.

4. It is much less affected by outliers.

5. Scikit-Learn provides a transformer called StandardScaler for standardization.

6. It is useful when the feature distribution is Normal or Gaussian.

7. 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 model evaluation method that is better than residuals. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on ``new" data. This is the basic idea for a whole class of model evaluation methods called cross validation.

Advantage –

Cross-validation gives us an idea about how the model will perform on an unknown dataset.

Cross-validation helps to determine a more accurate estimate of model prediction performance.

Disadvantges –

With cross-validation, we need to train the model on multiple training sets.

Cross-validation is computationally very expensive as we need to train on multiple training sets.