Name of the Assignment: Machine Learning (Worksheet-4)

Submitted by : Shalini Joshi

Designation : Data Science Intern

Date of Submission : 7th Jan,2023

Objective Type Questions:

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

Ans: 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

Ans: 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

Ans: A) 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

Ans: 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

Ans: 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?

A) remains same B) increases C) decreases D) none of the above

Ans: C) decreases

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

Ans: 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

Ans: B) Principal Components are calculated using unsupervised learning techniques & C) Principal Components are linear combinations of Linear Variables.

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.

Ans: A) Identifying developed, developing and under-developed countries on the basis of factors like GDP, poverty index, employment rate, population and living index & 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

Ans: A) max_depth, B) max_features, C) n_estimators ,D) min_samples_leaf

Subjective Type Questions:

11. What are outliers? Explain the Inter Quartile Range (IQR) method for outlier detection.

Ans: Outliers are the data points in the dataset which significantly differ from other observations. These are also known as the reason for noise in the dataset. Outliers cause problems while preparing statistical analysis. We can classify data into 4 quartiles.

1st (lower) quartile (Q1): median of the lower half of the data

2nd quartile (Q2): median of the entire data

3rd (upper) quartile (Q3): median of the upper half of the data

IQR is given by the difference of Q3 and Q1. This is the range where bulk of the data lies. The data points lower with values lower than 1.5*Q1 and higher than 1.5*Q3 are generally termed as outliers

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

Ans: Bagging and boosting are the types of method used in ensemble learning techniques. Bagging algorithm takes homogenous yet independent weak learning models and combines them parallel and learn from them.

Boosting is also a method in which the algorithm takes homogenous weak learners and learn them sequentially and adaptively to improve the models prediction making it a strong learner.

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

Ans: The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. R2 shows how well data fit a curve. The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit. Adjusted R2 will be always lesser than R2. In simple terms adjusted R2 penalizes if there is data that does not add values to the model. Adj R2 is calculated by the following formula:

$\text{Adj R2} = 1 -[(1-R^2)(n-1)/n-k-1]$

K is the number of predictors

Where n is the number of data points

14. What is the difference between standardisation and normalisation?

Ans: Normalization and Standardization are the methods used for scaling the data. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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

Ans: Cross validation is a technique used to fit different data and test different data in every iteration. This technique which involves reserving a particular sample of a dataset on which you do not train the model. Then, you test your model on this sample before finalizing it. This makes sure that the sample used for training and testing does not bias the model. For eg, if the cv is set to 5, then there are 5 train test splits done of the data, each with different data and testing with the remaining i.e. test data of very split.

**************************The End**