MACHINE LEARNING

ASSIGNMENT - 4

1 In Q1 to Q7, only one option is correct, Choose the correct option:

1. The value of correlation coefficient will always be:

C) between -1 and 1

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

D) Ridge Regularisation

3. Which of the following is not a kernel in Support Vector Machines?

C) hyperplane

4. Amongst the following, which one is least suitable for a dataset having non-linear decision boundaries?

A) Logistic Regression

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'

6. As we increase the number of estimators in ADABOOST Classifier, what happens to the accuracy of the model?

D) none of the above

7. Which of the following is not an advantage of using random forest instead of decision trees?

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?

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

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

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.

A. Outliers are extreme values that differ from most data points of a dataset.

IQR method: IQR is a range about the centre of the dataset which helps us to set boundaries around the data and hence predict the outliers

Here the Data is divided into 3 Quartiles Q1 Q2 and Q3

IQR= Q3 -Q1

Upper Limit = Q3 + (1.5*IQR)

Lower Limit = Q1 - (1.5*IQR)

Any data points above or below the said limits are considered outliers.

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

A. Bagging solves the overfitting problem by reducing variance where as boosting decreases bias.

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

A. R2 is used to evaluate the fit of a linear model. However as with the increase of independent variables the R2 value increases irrespective of the new variable having a positive relationship with the output variable.

To fix this problem Adjusted R2 is used

Adjusted R2 penalises the excessive use of features which do not correlate with the output data

Adjusted R2 = 1-(1-R2)(N-1)/N-p-1

14. What is the difference between standardisation and normalisation?

A. Standardisation rescales the dataset to have a mean at 0 and standard deviation of 1 .

This is called z distribution. we get a z-score from this distribution

Normalisation rescales the dataset so that each value (feature) fall between 0 and 1 or in other words is on the same scale.

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

A. Cross validation is used to tackle the problem of overfitting .

Cross validation is a resampling technique with a basic idea of dividing the training dataset into two parts… Train and Test. Train is used for training the model and Test is used for testing the model.

Cross Validation Techniques:

1 Train Test Split

2 K-Flod Cross Validation

3 LOOCV : leave one out cross validation