MACHINE LEARNING worksheet4

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

Q1) C) between -1 and 1

Q2) C) Recursive feature elimination

Q3) A) linear

Q4) A) Logistic Regression

Q5) C) old coefficient of 'X' ÷ 2.205

Q6) C) decreases

Q7) C) Random Forests are easy to interpret

In Q8 to Q10, more than one options are correct, Choose all the correct options:

Q8) Both B) and C) are true . B) Principal Components are calculated using unsupervised learning techniques C) Principal Components are linear combinations of Linear Variables

Q9) all are correct

Q10) A) max_depth B) max_features D) min_samples_leaf are true

Q11 to Q15 are subjective answer type questions, Answer them briefly.

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

Ans. An outliers is an object or entity that lies outside or deviates from other points in a group.

So basically outliers is point which is differs from others and creates anamoly and differences in result by providing uneven odd datapoints.They can be caused by execution or measurement error.

IQR method for detecting outliers.

IQR is used to check the variation in data set by dividing the data set into quartile.

The data set is divided into 5 quartile range a) minimum value b)25^{th} percentile c) 50^{th} percentile d) 75^{th} percentile and e) maximum percentile

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

Bagging = it's a basically a Bootstrap aggregating.it is an ensemble technique which is designed to improve the accuracy and stability of machine learning model used for statistical classification and regression problems.so we can say that it's a homogeneous weak learning method that learn from each other independently in parallel and combine them together to get the average result.

Boosting = its also a homogeneous weak learning model but works differently from bagging.

In this model learners learn sequencelly from weak learners add some weight to ractify the previous and again move forward to improve the machine learning model.it is also an ensemble technique that attempts to build a strong classifier from several weak classifier leraning and improving in every forward step in sequential manner.

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

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.

Let's have a look at the formula for adjusted R-squared to better understand its working.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

Adjusted R-squared

Here,

n represents the number of data points in our dataset

k represents the number of independent variables, and

R represents the R-squared values determined by the model.

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

On the other hand, if on adding the new independent variable we see a significant increase in R-squared value, then the Adjusted R-squared value will also increase.

14. What is the difference between standardisation and normalisation?

Normalization

a. Minimum and maximum value of features are used for scaling.

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

c. scales between [0,1] or between [-1,1]

d. It is affected by outliers.

e. Scikit-Learn provides a transformer called min max scaler for Normalization.

f. This transformation squishes the n-dimensional data into an n-dimensional unit hypercube

g. It is useful when we don't know about the distribution

h. It is a often called as Scaling Normalization

Standarization.

a. Mean and standard deviation is used for scaling

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

c. It is not bounded to a certain range

d. It is much less affected by outliers

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

f. It translates the data to the mean vector of original data to the origin and squishes or expands

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

h. It is a often called as Z-Score Normalization

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

Cross validation is basically a statistical model used for evaluating and comparing learning algorithms by dividing the data into two segments.one is used for learning the model and other one is used for evaluating the model.

Advantage

1. it is used to check the overfitting of the model

Disadvantages

1.it increases the calculating time and is costly method.