MACHINE LEARNING

WORKSHEET – 1 answers

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:

A) Lasso 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 :

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 :

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’ B) same as old coefficient of ‘X’

C) old coefficient of ‘X’ ÷ 2.205 D) Cannot be determined

Answer:

C) old coefficient of ‘X’ ÷ 2.205

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 :

D) none of the above

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 :

D) Random Forests provide a reliable feature importance estimate

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 :

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
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

C) Identifying spam or ham emails

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
D) min_samples_leaf

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

Answer :

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Inter Quartile Range(IQR) method for outlier detection:

The interquartile range or "IQR", is just the width of the box in the box-and-whisker plot. That is, IQR = q3-q1 . The IQR can be used as a measure of how spread-out the values are.

This is done using these steps: -

1. Calculate the interquartile range (IQR) for the data by formula-

IQR = Q3-Q1 (Q3 and Q1 represent one quarter and three quarters of the way through the list of all data)

2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).

3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.

4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

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

Answer :

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data.
Aim to decrease variance, not bias
Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.
Aim to decrease bias not variance

13. What is adjusted R2 in logistic 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
Calculation:

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 typically means rescales the values into a range of [0,1].

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

Answer :

In machine learning, we couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique

The advantage of this method is that the proportion of the validation or training split is not dependent on the number of folds (K-fold test)

Disadvantage as there are chances that you might miss out some observations whereas you might select some observations more than once