

MACHINE LEARNING

1. The value of correlation coefficient will always be:

Ans- (c) between -1 and +1

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

Ans- PCA

3. Which of the following is not a kernel in support vectore machine?

Ans- linear

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

Ans-Logistic Regression

5. In a Linear Regresion problem 'X' is independent variable and 'Y' is dependent variable where 'X' represents weight in pounds. If you convert the until of 'X' to kilograms,then new coefficient of 'X' will be?

Ans-

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

Ans-increases

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

Ans- Random Forests reduce overfitting

8. Which of the following is correct about principal components?

Ans- All of the above

9. Which of the following are applications of clustering?

Ans- (c)

10. Which of the following is(are) hyper parameters of decision tree?

Ans-(a) max_depth

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

Ans- An outlier is an individual point of data that is distant from other points in the dataset. It is an anomaly in the dataset that may be caused by a range of errors in capturing, processing or manipulating data. Outliers can skew overall data trends, so outlier detection methods are an important part of statistics.

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

Ans- Bagging is a method of merging the same type of

predictions. Boosting is a method of merging different type of predictions. Bagging decreases variance, not bias and solves overfitting issues in a model. Boosting decreases bias, not variance.

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

Ans-Adjusted R is corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R tends to optimistically estimate the fit of the linear regression.

14. What is the difference between standardisation and normalisation?

Ans- In Normalisation, the change in values is that they are at a standard scale without distorting the differences in the values. Whereas , Standardisation assumes that the dataset is in gaussian distribution and measures the variable at different scale making all the variables equally contribute to the analysis.

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

Ans- Cross validation in machine learning is a great technique to deal with overfitting problem in various algorithms. Instead of training our model on one training dataset, we train our model on many datasets.

Advantage- **Reduces overfitting**- In cross validation we split the dataset into multiple folds. This prevents our model from overfitting the training dataset. So, in this way the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantage- **Increase Training Time**-Cross validation drastically increases the training time. Earlier you had to train your model only on one training set.