

## MACHINE LEARNING

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

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 )

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 )

3. Which of the following is not a kernel in Support Vector Machines?
- A) linear
  - B) Radial Basis Function
  - C) hyperplane
  - D) polynomial

Ans- ( C )

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- ( A )

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)  $2.205 \times$  old coefficient of 'X'
  - B) same as old coefficient of 'X'
  - C) old coefficient of 'X'  $\div 2.205$
  - D) Cannot be determined

Ans- ( C )

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- ( B )

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 than decision trees
  - C) Random Forests are easy to interpret
  - D) Random Forests provide a reliable feature importance estimate

Ans- ( D )

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 ) and ( C )

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

## MACHINE LEARNING

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), (B), (C) and (D) All of the above

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) and (D)

**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.

Ans- Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts. Most commonly used method to detect outliers is visualization.

1. We use various visualization methods, like Box-plot, Histogram, Scatter Plot

2. Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier.

3. Data points, three or more standard deviation away from mean are considered outlier.

4. Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding.

5. Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance.

Some of the most popular methods for outlier detection are:

- Z-Score or Extreme Value Analysis (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Linear Regression Models (PCA, LMS)
- Proximity Based Models (non-parametric)
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)

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 types of predictions.

Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.

Models are built independently in Bagging. New models are affected by a previously built model's performance in Boosting.

13. What is adjusted  $R^2$  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. 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. The adjusted R-squared can be negative,

## MACHINE LEARNING

but it's usually not.

Every time you add an independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable.

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

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.

14. What is the difference between standardisation and normalisation?

Ans- 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.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

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.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

## MACHINE LEARNING

$\mu$  is the mean of the feature values and  $\sigma$  is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

Ans- Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

Advantages of cross-validation:

1. More accurate estimate of out-of-sample accuracy.
2. More “efficient” use of data as every observation is used for both training and testing.

There is a disadvantage because the cross validation process can become a lengthy one. It depends on the number of observations in the original sample and your chosen value of ‘p.’

---