MACHINE LEARNING8

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

1. What is the advantage of hierarchical clustering over K-means clustering?

D) None of these

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

A) max_depth

3. Which of the following is the least preferable resampling method in handling imbalance datasets?
B) RandomOverSampler

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?

1. Type1 is known as false positive and Type2 is known as false negative.

2. Type1 is known as false negative and Type2 is known as false positive.

3. Type1 error occurs when we reject a null hypothesis when it is actually true.

C) 1 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur:

1. Randomly selecting the cluster centroids

2. Updating the cluster centroids iteratively

3. Assigning the cluster points to their nearest center

D) 1-3-2

6. Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?

A) Decision Trees

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi

Square Automatic Interaction Detection) Trees?

C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create

multiway trees (more than two children for a node)

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

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

A) Ridge will lead to some of the coefficients to be very close to 0

D) Lasso will cause some of the coefficients to become 0.

9. Which of the following methods can be used to treat two multi-collinear features?

A) remove both features from the dataset D) use Lasso regularization

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

A) Overfitting

Q10 to Q15 are subjective answer type questions, Answer them briefly.

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in

such a case?

A11. One- hot encoding must be avoided when the categorical feature is ordinal.

In such a case Label encoding can be used

12. In case of data imbalance problem in classification, what techniques can be used to balance the

dataset? Explain them briefly.

A12. Techniques that can be used to balance imbalanced classification datasets are:

- **Down sample majority class**
  Down sampling, or under sampling, is the most straightforward technique with the least amount of impact to your pipeline. It requires that you remove observations to bring the majority class count down to a level that is more in balance, and not necessarily equal with, the minority or other classes.

- **Up sample minority class**

  Up sampling, specifically using a synthetic method, is a little more involved. It requires that you create enough observations to bring the minority class up to a level that is more in balance with the majority or other classes.

- **SMOTE**

  The SMOTE (Synthetic Minority Oversampling Technique) family of algorithms is a popular approach to up sampling. It works by using existing data from the minority class and generating synthetic observations using a k nearest-neighbors approach.

- **ADASYN**

  ADASYN (Adaptive Synthetic Sampling Approach) is closely related to SMOTE. The major differentiator with ADASYN is that the algorithm will determine how many synthetic observations are needed for each existing minority class observation, which in-turn will lead to a well balanced dataset overall.

13. What is the difference between SMOTE and ADASYN sampling techniques?

A13. The SMOTE (Synthetic Minority Oversampling Technique) family of algorithms is a popular approach to up sampling. It works by using existing data from the minority class and generating synthetic observations using a k nearest-neighbors approach.

ADASYN (Adaptive Synthetic Sampling Approach) is closely related to SMOTE. The major differentiator with ADASYN is that the algorithm will determine how many synthetic observations are needed for each existing minority class observation, which in-turn will lead to a well balanced dataset overall.

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

A14. GridSearchCV is a technique for finding the optimal parameter values from a given set of parameters in a grid. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.

A15. Metrics used to evaluate regression model are:

1) Mean Absolute Error(MAE)

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

2) Mean Squared Error(MSE)

MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

3) Root Mean Squared Error(RMSE)

As RMSE is clear by the name itself, that it is a simple square root of mean squared error

5) R Squared (R2)

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

6) Adjusted R Squared

The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.

Hence, To control this situation Adjusted R Squared came into existence.

It penalises excessive use of features which do not corelate with the output data.