MACHINE LEARNING

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

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

A) Hierarchical clustering is computationally less expensive B) In hierarchical clustering you don’t need to assign number of clusters in beginning

C) Both are equally proficient D) None of these

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

A) max_depth B) n_estimators

C) min_samples_leaf D) min_samples_split

Ques:-3 Which of the following is the least preferable resampling method in handling imbalance datasets?

A) SMOTE B) RandomOverSampler

C) RandomUnderSampler D) ADASYN

Ques:-4 Which of the following Type1 statements is/are true about “Type-1” and “Type-2” errors?

1. 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.

A) 1 and 2 B) 1 only

C) 1 and 3 D) 2 and 3

Ques:-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

A) 3-1-2 B) 2-1-3

C) 3-2-1 D) 1-3-2

Ques:-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 B) Support Vector Machines

C) K-Nearest Neighbors D) Logistic Regression

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

Square Automatic Interaction Detection) Trees?

A) CART is used for classification, and CHAID is used for regression.

B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of two children for a node).

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)

D) None of the above

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

Ques:-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

B) Lasso will lead to some of the coefficients to be very close to 0

C) Ridge will cause some of the coefficients to become 0

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

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

A) remove both features from the dataset B) remove only one of the features

C) Use ridge regularization D) use Lasso regularization

Ques:-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 B) Multicollinearity

C) Underfitting D) Outliers

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

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

such a case?

Answer:- The disadvantage of one hot encoding is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.

In such case we can use label encoding technique to encode the categorical variables.

Ques:-12 In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Answer:-There has been two different approaches to addressing imbalanced data: algorithm-level and data-level approach.

Algorithm approach: As mentioned above, ML algorithms penalize False Positives and False Negatives equally. A way to counter that is to modify the algorithm itself to boost predictive performance on minority class. This can be executed through either recognition-based learning or cost-sensitive learning.

Data approach: This consists of re-sampling the data in order to mitigate the effect caused by class imbalance. The data approach has gained popular acceptance among practitioners as it is more flexible and allows for the use of latest algorithms. The two most common techniques are over-sampling and under-sampling.

Over sampling: Over-sampling increases the number of minority class members in the training set. The advantage of oversampling is that no information from the original training set is lost, as all observations from the minority and majority classes are kept. On the other hand, it is prone to overfitting.

Under sampling: Under-sampling, on contrary to over-sampling, aims to reduce the number of majority samples to balance the class distribution. Since it is removing observations from the original data set, it might discard useful information.

EditedNearestNeighbours under-sampling technique (E2_ENN): The ENN method was proposed by Wilson (1972), in which a majority instance is removed if its class label does not agree with its K nearest neighbors. The ENN method tends to omit the noisy and borderline instances, which will therefore enhance the accuracy of decision boundary.

NearMiss 3 under-sampling technique (E3_NM): NearMiss-3 belongs to the NearMiss family, which conducts under-sampling on the majority class according to their distance. NearMiss-3 in particular removes majority samples with the largest distance from minority samples' K nearest neighbors.

SMOTE over-sampling technique (E4_SMT): SMOTE first considers the K nearest neighbors of the minority instances. It then constructs feature space vectors between these K neighbors, generating new synthetic data points on the lines.

ADASYN over-sampling technique (E5_ADS): Very similar to SMOTE, ADYSYN also creates synthetic data points with feature space vectors. However, for the new data points to be realistic, ADYSYN adds a small error to the data points to allow for some variance. This is because observations are not perfectly correlated in real life.

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

--> The key difference between ADASYN and SMOTE is that the ADASYN uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions. The SMOTE generates the same number of synthetic samples for each original minority sample.

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

--> GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.

One of the drawbacks of grid search is that when it comes to dimensionality, it suffers when evaluating the number of hyperparameters grows exponentially. However, there is no guarantee that the search will produce the perfect solution, as it usually finds one by aliasing around the right set.

However we can use Random search instead, Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. It is similar to grid search, and yet it has proven to yield better results comparatively. The drawback of random search is that it yields high variance during computing. Since the selection of parameters is completely random; and since no intelligence is used to sample these combinations, luck plays its part.

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

Answer:- There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square
2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
3. Mean Absolute Error(MAE)

R Square/Adjusted R Square:- R Square measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square.

R Square is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value is between 0 to 1 and bigger value indicates a better fit between prediction and actual value.

R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalise additional independent variables added to the model and adjust the metric to prevent overfitting issue.

Mean Square Error(MSE)/Root Mean Square Error(RMSE):- While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret much insights from one single result but it gives you an real number to compare against other model results and help you select the best regression model.

Root Mean Square Error(RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and make it easier for interpretation.

Mean Absolute Error(MAE):- Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of absolute value of error.

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalisation to big prediction error by square it while MAE treats all errors the same.