MACHINE LEARNING ASSIGNMENT

only one option is correct, Choose the correct option:

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

B) In hierarchical clustering you don't need to assign number of clusters in beginning

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?

C) RandomUnderSampler

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?

B) Support Vector Machines

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

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).

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

MACHINE LEARNING ASSIGNMENT - 8

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

C) Use ridge regularization

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?

Ans: One-hot encoding should be avoided when dealing with high cardinality categorical features, as it can result in a large number of dummy variables, causing memory issues and increasing computational cost. In such cases, count or frequency encoding can be used, where each category is replaced with its count or frequency of occurrence in the dataset. However, it is important to evaluate the performance of different encoding techniques based on the specific problem at hand

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

Ans:

There are several techniques that can be used to balance a dataset in the case of data imbalance problem in classification. Some of the commonly used techniques are:

Random undersampling: Randomly removing samples from the majority class to balance the number of samples in both classes.Random oversampling: Randomly duplicating samples from the minority class to balance the number of samples in both classes.Synthetic Minority Over-sampling Technique (SMOTE): Generating synthetic samples from the minority class by interpolating between existing samples to increase the number of minority class samples.

Adaptive Synthetic Sampling (ADASYN): Similar to SMOTE, but generates more synthetic samples for minority samples that are harder to learn by the model.Class-weighted loss function: Assigning higher weights to the minority class samples in the loss function during training to give more importance to the minority class.

Ensemble methods: Using ensemble methods such as bagging, boosting, or stacking to combine multiple models and balance the dataset through the diversity of models.

These techniques can be used individually or in combination to balance the dataset and improve the performance of the classification model. It is important to carefully evaluate the performance of the model on the balanced dataset and choose the appropriate technique(s) based on the specific problem at hand.

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

Ans: SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are both techniques used to address the data imbalance problem in classification. Both techniques generate synthetic samples from the minority class to balance the number of samples in both classes.The key difference between SMOTE and ADASYN is in how they generate synthetic samples. SMOTE generates synthetic samples by linearly interpolating between existing minority class samples, whereas ADASYN generates synthetic samples by using a density distribution-based approach.

In more detail, the SMOTE algorithm selects a minority class sample and identifies its k nearest minority class neighbors. It then selects a random number between 0 and 1 and generates a new synthetic sample by interpolating between the selected minority class sample and one of its k nearest minority class neighbors. This process is repeated until the desired number of synthetic samples is generated.

ADASYN, on the other hand, adjusts the density distribution of the minority class samples based on their level of difficulty in learning by the model. Specifically, ADASYN generates more synthetic samples for minority class samples that are harder to learn by the model. This is done by computing a density distribution around each minority class sample and generating synthetic samples in regions with lower density. This helps to balance the dataset while also providing more focus on samples that are difficult to classify.In summary, both SMOTE and ADASYN are effective techniques for addressing the data imbalance problem in classification, but they differ in how they generate synthetic samples. SMOTE uses linear interpolation between existing minority class samples, while ADASYN uses a density distribution-based approach that focuses more on difficult-to-classify samples.

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

Ans: The purpose of using GridSearchCV (Grid Search Cross-Validation) is to find the best hyperparameters for a machine learning model. Hyperparameters are parameters that are set prior to training the model and cannot be learned from the data, such as the regularization parameter in linear regression or the number of trees in a random forest. GridSearchCV is a technique that exhaustively searches over a predefined hyperparameter space to find the combination of hyperparameters that gives the best performance.

GridSearchCV can be used with any machine learning algorithm and is especially useful when the number of hyperparameters is small. It is preferable to use GridSearchCV when working with small to medium-sized datasets where the model training time is reasonable. However, for very large datasets, GridSearchCV may take a long time to complete, as it requires training and evaluating the model for every combination of hyperparameters. In such cases, it may be more efficient to use other hyperparameter tuning techniques such as random search or Bayesian optimization, which sample from the hyperparameter space more efficiently.

In summary, GridSearchCV is a useful technique for finding the best hyperparameters for a machine learning model, but its efficiency depends on the size of the dataset and the number of hyperparameters being tuned. It may not be preferable to use in case of large datasets, but there are other hyperparameter tuning techniques that can be used in such cases.

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

Ans: Mean Squared Error (MSE): MSE measures the average of the squared differences between the predicted and actual values. It gives more weight to large errors and is sensitive to outliers.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and is a widely used metric in regression. It is more interpretable than MSE and has the same units as the target variable.

Mean Absolute Error (MAE): MAE measures the average of the absolute differences between the predicted and actual values. It is less sensitive to outliers than MSE.

R-squared (R^2): R^2 measures the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit.

Adjusted R-squared: Adjusted R^2 is a modification of R^2 that takes into account the number of predictors in the model. It penalizes the addition of unnecessary predictors that do not improve the fit.

Mean Absolute Percentage Error (MAPE): MAPE measures the average of the absolute percentage differences between the predicted and actual values. It is commonly used in forecasting and is more interpretable than other metrics.

Coefficient of Determination (COD): COD measures how well the regression line approximates the actual data points. It ranges from 0 to 1, with higher values indicating a better fit.

Explained Variance Score (EVS): EVS measures the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit.

These evaluation metrics help to assess the performance of a regression model and can be used to compare different models and select the best one. It is important to choose the appropriate evaluation metric based on the specific problem at hand and interpret the results in the context of the data