ASSIGNMENT – 8

MACHINE LEARNING

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

Ans: 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?

Ans: A) max_ depth

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

Ans: B) Random Over Sampler

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

Ans: C) 1 and 3

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

Ans: A) 3-1-2

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?

Ans: B) Support Vector Machines

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

Ans: 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?

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

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

Ans: B) remove only one of the features

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?

Ans: A) Overfitting

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

Ans: One-hot encoding is a popular technique for converting categorical variables into numerical data that can be used in machine learning models. However, there are some situations where it may not be the best choice.

One such situation is when dealing with high cardinality categorical variables. High cardinality variables are categorical variables with a large number of unique values. One-hot encoding these variables can result in a large number of new features, which can lead to the curse of dimensionality and overfitting.

In these situations, an alternative encoding technique called target encoding or mean encoding can be used. In target encoding, each category is replaced with the average value of the target variable for that category. This technique can work well when there are a large number of categories and when the categories have a strong relationship with the target variable.

Another situation where one-hot encoding may not be appropriate is when dealing with ordinal variables, where the categories have a natural order or ranking. In this case, ordinal encoding can be used, where the categories are assigned, numerical values based on their order or rank. This can help to preserve the natural ordering of the categories, which can be important in some applications.

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

Ans: Data imbalance is a common problem in classification, where the number of examples in one class is significantly higher than the number of examples in another class. This can lead to biased models that perform poorly on the underrepresented class. There are several techniques that can be used to balance the dataset:

1. **Random under sampling: In this technique, some examples are randomly removed from the majority class until the dataset is balanced. This can be effective when there is a large amount of data and the majority class has a significant number of examples.**
2. **Random oversampling: In this technique, some examples are randomly duplicated from the minority class until the dataset is balanced. This can be effective when the minority class has a small number of examples and there is no risk of overfitting.**
3. **Synthetic minority oversampling technique (SMOTE): In this technique, new examples are generated for the minority class by interpolating between existing examples. This can be effective when the minority class has a small number of examples and there is a risk of overfitting with random oversampling.**
4. **Class weights: In this technique, a weight is assigned to each class based on its frequency in the dataset. This can be used to balance the dataset during training by giving more weight to the minority class.**
5. **Cost-sensitive learning: In this technique, the misclassification cost of each class is taken into account during training. This can be used to balance the dataset by making the classifier more sensitive to the minority class.**

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

Ans: SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling) are both techniques used to address the problem of data imbalance in classification tasks, and they are both based on generating synthetic

examples for the minority class. However, there are some differences between these techniques:

1. **Sampling strategy: SMOTE generates synthetic examples by interpolating between existing examples in the minority class, while ADASYN generates synthetic examples by focusing on the regions where the density of the minority class is low and generating examples near the decision boundary.**
2. **Sampling density: ADASYN is designed to generate more synthetic examples in regions where the density of the minority class is low, which can help to improve the generalization performance of the classifier. In contrast, SMOTE generates the same number of synthetic examples for all minority class examples.**
3. **Degree of randomness: SMOTE generates synthetic examples by randomly selecting existing examples and interpolating between them, while ADASYN uses a weighted distribution to generate examples based on the density of the minority class.**
4. **Implementation: SMOTE is a relatively simple technique and is widely used in many machine learning libraries, while ADASYN is a more recent technique and may not be available in all libraries.**
5. **In summary, while both SMOTE and ADASYN are effective techniques for addressing the problem of data imbalance, they have some differences in their approach and implementation. Researchers and practitioners should consider their specific needs and goals when selecting one of these techniques.**

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

Ans: GridSearchCV is a technique used for hyperparameter tuning in machine learning models. It involves defining a grid of possible hyperparameter values and systematically searching through this grid to find the optimal combination of hyperparameters that results in the best performance on a validation set. The purpose of using GridSearchCV is to automate the process of hyperparameter tuning and to help the model achieve the best possible performance.

In general, GridSearchCV can be used with datasets of any size. However, it can become computationally expensive and time-consuming for large datasets or when the number of hyperparameters to search is large. This is because GridSearchCV performs an exhaustive search over the entire grid of hyperparameter values, which can quickly become impractical for very large grids or datasets.

To address this issue, there are some alternatives to GridSearchCV that can be used for hyperparameter tuning with large datasets. One approach is to use a randomized search, which randomly samples from the hyperparameter grid rather than performing an exhaustive search. This can be more efficient than GridSearchCV and can still result in good performance.

Another approach is to use a Bayesian optimization method, which uses a probabilistic model to optimize the hyperparameters. This can be more efficient than both GridSearchCV and randomized search and can perform well even with very large datasets.

In summary, GridSearchCV is a useful technique for hyperparameter tuning in machine learning models, but it may not be the best choice for very large datasets or when the number of hyperparameters to search is large. In these cases, alternative methods such as randomized search or Bayesian optimization may be more efficient and effective.

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

Ans: Evaluation metrics are used to assess the performance of a regression model. Here are some commonly used evaluation metrics for regression models:

1. **Mean Squared Error (MSE): MSE is the average of the squared differences between the predicted and actual values. It measures the average squared distance between the predicted and actual values and penalizes large errors. The lower the MSE, the better the model.**
2. **Root Mean Squared Error (RMSE): RMSE is the square root of MSE. It is also a measure of the average distance between the predicted and actual values,**

but it is in the same units as the target variable. The lower the RMSE, the better the model.

3. **Mean Absolute Error (MAE): MAE is the average of the absolute differences between the predicted and actual values. It measures the average absolute distance between the predicted and actual values and is less sensitive to outliers than MSE. The lower the MAE, the better the model.**
4. **R-squared (R2): R2 measures the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit. A value of 1 indicates that the model perfectly predicts the target variable, while a value of 0 indicates that the model does not explain any of the variance in the target variable.**
5. **Adjusted R-squared (R2 adjusted): R2 adjusted is similar to R2, but it adjusts for the number of predictor variables in the model. It penalizes models that have too many predictor variables and can help to avoid overfitting.**
6. **Mean Absolute Percentage Error (MAPE): MAPE is the average of the absolute percentage differences between the predicted and actual values. It measures the average percentage distance between the predicted and actual values and is commonly used when the target variable has a significant range. However, it can be sensitive to small values and can be problematic when the actual values are close to zero.**
7. **Coefficient of determination (CD): Coefficient of determination measures how well the regression model fits the observed data. It is a ratio of the explained variation to the total variation of the dependent variable. It gives a value between 0 and 1, where 0 means that the model doesn't explain the variability of the dependent variable at all and 1 means that the model perfectly explains the variability of the dependent variable.**