**ML – Assignment – 8**

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

**Ans: - B**

1. **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\_splits

**Ans:- A**

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

A) SMOTE

B) RandomOverSampler

C) RandomUnderSampler

D) ADASYN

**Ans:-C**

1. **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.

A) 1 and 2

B) 1 only

C) 1 and 3

D) 2 and 3

**Ans:-C**

1. **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

**Ans:- D**

1. **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

**Ans:-C**

1. **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

**Ans:- B**

1. **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.

**Ans:- A&D**

1. **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

**Ans:-C&D**

1. **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

**Ans:- A&B**

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

One-Hot encoding can be avoided in situations where the no.of categorical variables is very large, or if the categorical variable has high cardinality, i.e many unique categories. One-hot-encoding can result in a large no. of features and lead to the curse of dimensionality, where the model becomes too complex and may overfit the training data.

In such cases, an alternative encoding technique that can be used is target encoding or likelihood encoding. Target encoding involves replacing each category in categorical variable with the average target value of the corresponding category. This encoding technique can be useful when there is a strong relationship between the category and the target variable, and it can help reduce the dimensionality of the dataset.

Likelihood encoding is similar to target encoding, but it takes into account the prior probability of the target variable, in addition to the frequency of each category. This can be useful in situations where the target variable is imbalanced or if there are some rare categories in the categorical variable.

It is important to note that the choice of encoding technique depends on the specific problem and the nature of the data, and it is important to evaluate the performance of different encoding techniques using cross-validation or other appropriate methods.

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

Data imbalance is common problem in classification tasks where the no.of examples in one class is significantly smaller than the other class. In such cases, the classification model tends to perform poorly on the minority class, and its overall accuracy can be misleading. To address this problem, various techniques can be used to balance the dataset. Here are some of the commonly used techniques:

**Random undersampling**: This technique involves randomly removing examples from the majority class to balance the no. of examples in both classes. This technique can be useful when the majority class has a very high no. of examples.

**Random Oversampling:** This technique involves randomly duplicating examples from the minority class to increase the no of examples in that class. This technique can be useful when the minority class has a very low no. of examples.

**Synthetic minority over-sampling Technique(SMOTE):** This technique involves generating synthetic examples from the minority class based on the existing examples. This technique can help balance the dataset without losing valuable information.

**Class weighting:** This technique involves assigning higher weights to the minority class during model training to increase its importance. This techniques can be useful when the dataset is too large to be balanced using the other techniques.

**Data Augmentation:** This technique involves generating new examples by applying transformations to the existing examples, such as rotating or flipping images. This technique can help increase the no. of examples in both classes and improve the performance of the model.

It’s important to note that the choice of technique may depend on the specifics of the problem, and a combination of these techniques maybe used to achieve the best results.

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

SMOTE(synthetic minority over-sampling technique) and ADASYN(adaptive synthetic) are two commonly used oversampling techniques used in imbalanced classification problems. Both techniques generate synthetic examples for the minority class to balance the class distribution.

The main difference between SMOTE and ADASYN adjusts the importance of the minority class samples based on their level of difficulty in learning, where as SMOTE does not consider the difficulty of learning. In other words, ADASYN generates more synthetic examples for minority class samples that are harder to learn.

**Here are some more details on each technique**:-

**SMOTE :- SMOTE** generates synthetic examples for the minority class by interpolating between existing minority class samples. Specifically, SMOTE selects a minority class sample and find its K nearest neighbors in the feature space. Then, for each of these K neighbors, SMOTE generates a new synthetic example by interpolating between the minority class sample and its neighbor. The interpolation is performed by choosing a random point along the line segment connecting the two samples. This process is repeated for each minority class sample until the desired level of balance is achieved.

**ADASYN:-**

ADASYN also generates synthetic examples for the minority class, but it does so in an way that depends on the level of difficulty of learning each minority class sample. Specifically, ADASYN first calculates the difference between the number of minority class samples and the no of majority class samples and the no. of its K nearest neighbors that belong to the majority class. The higher the weight, the harder it is to learn the minority class sample. ADASYN then generates synthetic examples for each minority class sample in proportion to its weight. This means that harder-to-learn minority class samples will be over sampled more than easier-to-learn minority class samples.

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

GridSearchCV is a technique used in machine learning to tune hyperparameters of a model in order to optimize its performance. Hyperparameters are values set by the user that determine how a machine learning model is trained, such as the learning rate or the regularization strength. GridSearchCV searches over a specified set of hyperparameters for the model and returns the set of hyperparameters that performs the best according to a given evaluation metric.

GridSearchCV is a widely used technique in machine learning, especially when there are many hyperparameters to be tuned. It is particularly useful when there are multiple hyperparameters to be tuned and their effects on the model’s performance are not known in advance. It is also useful when there are several models to be compared with each other.

However, when the dataset is large, the time required to perform a grid search can be come very long, especially when many hyperparameters are being tuned. In this case, other optimization techniques such as Randomized SearchCV or Bayesian optimization may be more appropriate, as they can find the optimal hyperparameters more efficiently.

In Summary, GridSearchCV is a powerful technique for hyperparameter tuning in machine learning, but its usefulness depends on the specific context of the problem, including the size of the dataset, the no. of hyperparameters to be tuned and the computational resources available.

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

There are several evaluation metrics used to evaluate a regression model. Some of the commonly used ones are:

**Mean squared error(MSE):** - It is measures the average difference between the predicted and actual values. A lower value of MSE indicates better model performance.

**Root Mean squared error(RMSE):-** It is square root of the MSE and represents the average difference between predicted and actual values. As with MSE, a lower value of RMSE indicates better model performance.

**Mean absolute error(MAE):-** It is the average absolute difference between predicted and actual values. A lower value of MAE indicates better model performance.

**R-squared (R2):-** It represent the proportion of variance in the dependent variable that can be explained by the independent variable(s). A higher value of R2 indicates better model performance.

**Mean absolute percentage error(MAPE):-** It measures the percentage difference between predicted and actual values. A lower value of MAPE indicates better model performance.

**Root mean squared percentage error(RMSPE):-** It is the square root of the mean squared percentage error and represents the percentage difference between predicted and actual values. A lower value of RMSPE indicates better model performance.

**Co-efficient of Determination(CD):-** It measures the proportion of the variation in the dependent variable that is explained by the independent variable(s). A higher of CD indicates better model performance.

Each of these metrics has its own strengths and weakness, and the choice of which one to use depends on the specific context and goals of the regression model. In general, it is a good practice to use multiple metrics to get a more comprehensive understanding of the model’s performance.