**HRITIK YADAV**

**INTERNSHIP=34**

**(WORKSHEET=1)**

**MACHINE LEARNING**

**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?**

**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**

**ANSWER:-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 B) n\_estimators**

**C) min\_samples\_leaf D) min\_samples\_splits**

**ANSWER:-A) max\_depth**

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

**A) SMOTE B) RandomOverSampler**

**C) RandomUnderSampler D) ADASYN**

**ANSWER:-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.**

**A) 1 and 2 B) 1 only**

**C) 1 and 3 D) 2 and 3**

**ANSWER:-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**

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

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

**ANSWER:-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 B) Support Vector Machines**

**C) K-Nearest Neighbors D) Logistic Regression**

**ANSWER:-C) K-Nearest Neighbors**

**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**

**ANSWER:-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**

**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.**

**ANSWER:-B) Lasso 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**

**B) remove only one of the features**

**C) Use ridge regularization**

**D) use Lasso regularization**

**ANSWER:-B) remove only one of the features**

**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 B) Multicollinearity**

**C) Underfitting D) Outliers**

**ANSWER:-A) Overfitting**

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

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

**ANSWER:-**One-hot encoding must be avoided in situations where the number of unique values in a categorical feature is very high, such as in the case of a feature with a large number of unique words or categories. This can lead to a large number of new features after one-hot encoding, which can cause computational problems and make the model too complex.

In such situations, we can use the Count encoding or Target encoding technique instead of one-hot encoding.

Count encoding replaces each category in a feature with the count of observations in the training data that have that category. This technique reduces the dimensionality of the data and provides some information about the frequency of each category.

Target encoding, also known as mean encoding, replaces each category in a feature with the mean of the target variable for that category. This technique is useful when there is a strong correlation between the target variable and the categorical feature.

Both count encoding and target encoding are useful when dealing with high-cardinality categorical features and can provide valuable information for the model without increasing its complexity.

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

**them briefly.**

**ANSWER:-**Imbalanced datasets in classification problems occur when the number of observations in one class is much larger or smaller than the number of observations in the other class. This can cause problems for the model, such as biased prediction accuracy towards the majority class. To balance the dataset, we can use the following techniques:

1.Random undersampling: Random undersampling involves randomly selecting a subset of observations from the majority class to balance the dataset. This technique can be effective when the dataset is not too large, and there is a large overlap between the minority and majority classes.

2.Random oversampling: Random oversampling involves randomly duplicating observations from the minority class to balance the dataset. This technique can be effective when the dataset is not too large, and there is a small overlap between the minority and majority classes.

3.Synthetic Minority Over-sampling Technique (SMOTE): SMOTE is a technique that generates new synthetic observations for the minority class by interpolating between existing observations. This technique can be effective when the dataset is imbalanced, and there is a clear separation between the minority and majority classes.

4.Ensemble methods: Ensemble methods such as Bagging, Boosting, and Stacking can be used to balance the dataset by combining multiple classifiers trained on balanced subsets of the dataset.

5.Cost-sensitive learning: Cost-sensitive learning involves modifying the loss function to account for the imbalance in the dataset. This technique can be effective when the cost of misclassifying the minority class is much higher than that of the majority class.

6.Anomaly detection: Anomaly detection techniques can be used to detect and remove outliers in the majority class, which can help balance the dataset and improve the model's performance on the minority class.

Overall, the choice of technique depends on the specific problem and dataset, and a combination of techniques may be required to balance the dataset effectively.

窗体底端

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

**ANSWER:-**Both SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are techniques used to address the problem of imbalanced datasets in classification problems. The main difference between SMOTE and ADASYN is how they generate synthetic samples for the minority class.

SMOTE generates synthetic samples by interpolating between existing samples of the minority class. Specifically, for each minority sample, SMOTE selects k nearest neighbors from the minority class and generates new samples along the line segments joining the minority sample to its k neighbors.

ADASYN, on the other hand, adaptively generates synthetic samples by focusing more on samples that are difficult to classify. Specifically, ADASYN starts by randomly selecting a minority sample and computes its k nearest neighbors from the minority and majority classes. It then generates new samples for the minority class along the line segments joining the minority sample to its k neighbors, but with a higher density of synthetic samples near the minority samples that are more difficult to classify.

In summary, SMOTE generates synthetic samples by interpolating between existing samples of the minority class, whereas ADASYN adaptively generates synthetic samples by focusing more on the minority samples that are difficult to classify. ADASYN is considered an extension of SMOTE and has been shown to outperform SMOTE in some cases, particularly when the degree of class imbalance is very high.

窗体底端

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

**ANSWER:-**GridSearchCV is a method used to tune hyperparameters in a machine learning model. It involves specifying a grid of hyperparameter values to search over and using cross-validation to evaluate the performance of each combination of hyperparameters. The purpose of GridSearchCV is to find the set of hyperparameters that results in the best performance of the model on a validation set.

GridSearchCV is often used to fine-tune hyperparameters of complex models, such as Support Vector Machines, Random Forests, or Neural Networks. The advantage of using GridSearchCV is that it systematically searches the hyperparameter space and helps avoid the risk of missing the best set of hyperparameters. It also allows for efficient comparison of multiple models and selection of the best one.

However, GridSearchCV can be computationally expensive, particularly for large datasets and models with many hyperparameters. The number of possible combinations of hyperparameters can grow exponentially, making GridSearchCV impractical for large datasets. In such cases, other methods such as RandomizedSearchCV, which randomly samples from the hyperparameter space, or Bayesian optimization, which uses probability distributions to guide the search, may be more efficient.

In summary, GridSearchCV is a powerful method for fine-tuning hyperparameters in machine learning models. However, it may not be the best choice for large datasets or models with many hyperparameters, in which case other methods such as RandomizedSearchCV or Bayesian optimization may be more efficient.

窗体底端

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

**ANSWER:-**Here are some of the evaluation metrics commonly used to evaluate a regression model:

1.Mean Squared Error (MSE): MSE is the average of the squared differences between predicted and actual values. It measures the average squared deviation of the predicted values from the true values. It is the most commonly used metric for regression models.

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

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

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, where 0 indicates that the model does not explain any variance, and 1 indicates that the model explains all the variance in the target variable. R2 is a commonly used metric for evaluating regression models, but it has limitations, such as being biased towards complex models and not indicating the quality of predictions.

5.Adjusted R-squared: Adjusted R2 is a modified version of R2 that adjusts for the number of predictor variables in the model. It penalizes overfitting and is a better metric than R2 for comparing models with different numbers of predictor variables.

6.Mean Percentage Error (MPE): MPE is the average of the percentage differences between predicted and actual values. It measures the average percentage deviation of the predicted values from the true values. MPE is useful for evaluating models in which the magnitude of the errors is important

7.Mean Absolute Percentage Error (MAPE): MAPE is the average of the absolute percentage differences between predicted and actual values. It measures the average absolute percentage deviation of the predicted values from the true values. MAPE is useful for evaluating models in which the magnitude of the errors is important.

In summary, there are several evaluation metrics for regression models, each with its own strengths and weaknesses. The choice of metric depends on the specific problem and the goals of the analysis.

窗体底端