MACHINE LEARNING WORK SHEET SET 01

Assignment 08

Q1 to Q7 have only one correct answer. Choose the correct option to answer your question.

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: A) SMOTE

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: A) Ridge 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: 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 B) Multicollinearity

C) Underfitting D) Outliers

Answer: 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?

Answer:

One-hot encoding is a technique that converts categorical variables into binary vectors, where each element represents the presence or absence of a category. One-hot encoding must be avoided when:

- The number of categories is very large, which can lead to increased dimensionality and sparsity of the data
- The categories have some relationship or order among them, which cannot be captured by one-hot encoding
- The computational cost and memory usage are too high for the model

Some alternative encoding techniques that can be used in such cases are:

- Label encoding, which assigns a numerical value to each category
- Ordinal encoding, which preserves the order of the categories
- Frequency encoding, which replaces the categories with their frequency of occurrence
- Embedding encoding, which maps the categories to low-dimensional vectors that capture semantic information

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

Answer:

Data imbalance problem in classification is when there is an unequal distribution of classes in the training dataset. This can cause algorithms to be biased towards the majority class and ignore the minority class. Some techniques that can be used to balance the dataset are:

- **Oversampling**: Creating copies of the minority class examples until it reaches a desired proportion. This can help increase the sensitivity of the model to the minority class, but it can also introduce overfitting.
- **Under-sampling**: Reducing the number of majority class examples until it reaches a desired proportion. This can help reduce the computational cost and overfitting, but it can also discard useful information from the majority class.
- **Synthetic Minority Oversampling Technique (SMOTE)**: Generating new examples for the minority class by interpolating between existing ones. This can help create more diversity and avoid overfitting, but it can also introduce noise and outliers.
- **Cost-Sensitive Learning**: Assigning different weights or costs to different classes based on their importance or frequency. This can help penalize misclassification errors more for the

minority class than for the majority class, but it can also require careful tuning of parameters.

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

Answer:

SMOTE and ADASYN are both oversampling techniques that generate synthetic samples for the minority class by interpolating between existing ones. The main difference is that

- ADASYN uses a density distribution to decide how many synthetic samples to generate for each minority sample, while SMOTE uses a uniform weight for all minority samples. This means that ADASYN can adaptively change the weights of different minority samples based on their difficulty to be classified, while SMOTE treats all minority samples equally.

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 popular method for hyperparameter tuning in machine learning. The purpose of using GridSearchCV is to search over a defined parameter space for the best hyperparameters that can optimize a given model's performance.

It works by exhaustively searching through a grid of hyperparameters, which are specified in advance, and evaluating the model's performance for each combination of hyperparameters. It then returns the combination of hyperparameters that resulted in the best performance.

GridSearchCV can be especially useful for large datasets as it can help to optimize the model's performance without the need for manual hyperparameter tuning. However, it can also be computationally expensive and time-consuming, particularly when dealing with large datasets or complex models.

In some cases, it may be more appropriate to use other methods for hyperparameter tuning, such as randomized search or Bayesian optimization, which can be more efficient and require less computation. It ultimately depends on the specific problem, the size of the dataset, and the resources available.

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

Answer:

Here are some commonly used evaluation metrics for regression models:

1. **Mean Squared Error (MSE):** MSE measures the average squared difference between the predicted and actual values. The MSE value will be higher for larger differences between predicted and actual values, which indicates a poor model fit.

2. **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, and it gives a value in the same units as the target variable. RMSE is preferred over MSE as it is more interpretable and easier to understand.
3. **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted and actual values. It is less sensitive to outliers as compared to MSE.
4. **R-squared (R2):** R-squared is a statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variables. The value of R2 ranges from 0 to 1, where 0 means the model does not explain any of the variance and 1 means the model perfectly explains all the variance.
5. **Adjusted R-squared (Adj-R2):** Adjusted R-squared is a modified version of R-squared that adjusts for the number of independent variables in the model. It penalizes the addition of irrelevant variables to the model, making it a more reliable metric for evaluating model performance.
6. **Mean Squared Logarithmic Error (MSLE):** MSLE measures the average logarithmic difference between the predicted and actual values. It is particularly useful when dealing with exponential growth, as errors in predicting small values are penalized more heavily than errors in predicting large values.
7. **Median Absolute Error (MedAE):** MedAE is a robust measure of error that is less sensitive to outliers than MAE. It measures the median absolute difference between the predicted and actual values.

Each of these evaluation metrics can provide different insights into the performance of a regression model, and the choice of which one to use will depend on the specific requirements of the problem at hand.