Machine Learning Assignment_8

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

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

QUES 2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

Answer: A) max_depth

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

Answer. D) ADASYN.

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

Answer. B) 1 only.

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

Answer. D) 1-3-2.

QUES 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?

Answer. B) Support Vector Machines.

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

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).

QUES 8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

Answer. C) Ridge will cause some of the coefficients to become 0

D) Lasso will cause some of the coefficients to become 0.

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

Answer. C) Use ridge regularization.

D) Use Lasso regularization.

QUES 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?

Answer. A) Overfitting.

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

Answer. One-hot encoding creates d-dimensional vectors for each instance where d is the unique number of feature values in the dataset. For a feature having a large number of unique feature values or categories, one-hot encoding is not a great choice.

Binary encoding might be a good alternative to one-hot encoding because it creates fewer columns when encoding categorical variables. Ordinal encoding is a good choice if the order of the categorical variables matters.

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

Answer. We can use seven different techniques to handle data imbalance problem in classification:

1. Use the right evaluation metrics.
2. Resample the training set.
3. Use K-fold Cross-Validation in the Right Way.
4. Ensemble Different Resampled Datasets.
5. Resample with Different Ratios.

6. Cluster the abundant class.
7. Design Your Models.

Resampling (Oversampling and Undersampling)
When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling.

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

Answer. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

Synthetic Minority Oversampling Technique-SMOTE

ADASYN is a generalized form of the SMOTE algorithm. This algorithm also aims to oversample the minority class by generating synthetic instances for it. But the difference here is it considers the density distribution, r_i which decides the no. of synthetic instances generated for samples which difficult to learn. Due to this, it helps in adaptively changing the decision boundaries based on the samples difficult to learn.

Adaptive Synthetic Sampling Approach

QUES 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 library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible.

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

Answer. There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square.

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE).

3. Mean Absolute Error(MAE).

Adjusted R Square. Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error (which is the sample variance of the target field). The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2.

R Square: R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. The R^2 tells us the percentage of variance in the outcome that is explained by the predictor variables (i.e., the information we do know). A perfect R^2 of 1.00 means that our predictor variables explain 100% of the variance in the outcome we are trying to predict.

Mean Square Error(MSE): Mean squared error (MSE) measures error in statistical models by using the average squared difference between observed and predicted values.

Root Mean Square Error (RMSE) is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

Mean Absolute Error(MAE): The mean absolute error (MAE) characterizes the alteration among the original and predictable values and is mined as the dataset's total alteration mean.