MACHINE LEARNING WORKSHEET 8

**In Q1 to Q7, only one option is correct, Choose the correct option:**

Q1. 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) In hierarchical clustering you don’t need to assign number of clusters in beginning

Q2. 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) max\_depth

Q3. Which of the following is the least preferable resampling method in handling imbalance datasets?

A) SMOTE B) RandomOverSampler

C) RandomUnderSampler D) ADASYN

Ans. a) SMOTE

Q4. 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) 1and 3

Q5. 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) 1-3-2

Q6. 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.b) Support Vector Machines

Q7. 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. 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:**

Q8. 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 and b options are correct**

Q 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

Ans. a) and d)

Q10. 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) Overfitting**

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

Q11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Ans. We should not use the One Hot Encoding method when :

• When the categorical features present in the dataset are ordinal i.e for the data being like Junior, Senior, Executive, Owner.

•  When the number of categories in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption.

Q12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Ans. . Choose Proper Evaluation Metric

The accuracy of a classifier is the total number of correct predictions by the classifier divided by the total number of predictions. This may be good enough for a well-balanced class but not ideal for the imbalanced class problem. The other metrics such as precision is the measure of how accurate the classifier’s prediction of a specific class and recall is the measure of the classifier’s ability to identify a class.

2.Resampling (Oversampling and Undersampling)

This technique is used to upsample or downsample the minority or majority class. 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.

3. . SMOTE

Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don’t add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use k nearest neighbor to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.

4.  BalancedBaggingClassifier

When we try to use a usual classifier to classify an imbalanced dataset, the model favors the majority class due to its larger volume presence. A BalancedBaggingClassifier is the same as a sklearn classifier but with additional balancing.

Q13. What is the difference between SMOTE and ADASYN sampling techniques?

Ans. The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions.

Q14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

Ans. GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

GridSearchCV is a function that comes in Scikit-learn’s(or SK-learn) model\_selection package.So an important point here to note is that we need to have the Scikit learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

Q15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.

Ans. 1) Root Mean Squared Error(RMSE)

As RMSE is clear by the name itself, that it is a simple square root of mean squared error.

2) Root Mean Squared Log Error(RMSLE)

Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.

To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.

3) R Squared (R2)

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides. The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically R2 squared calculates how must regression line is better than a mean line

4) Adjusted R Squared

The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.

Hence, To control this situation Adjusted R Squared came into existence.