**MACHINE LEARNING**

**WORKSHEET – 5**

**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) entropy

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:---**

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 centre

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 Neighbours D) Logistic Regression

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

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 regularisation if you take a large value of regularisation 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

B) Lasso will lead to some of the coefficients to be very close to 0

9. Which of the following methods can be used when there are correlated features in the dataset?

A) remove both features from the dataset

B) remove only one of the features

C) Use ridge regularisation D) use Lasso regularisation

**ANSWER:---** A) remove both features from the dataset

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 has the advantage that the result is binary rather than ordinal and that everything sits in an orthogonal vector space. The disadvantage is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.

Label encoding technique can be used in such a case.

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

**ANSWER:---** When faced with imbalanced data sets there is no one stop solution to improve the accuracy of the prediction model.  One may need to try out multiple methods to figure out the best-suited sampling techniques for the dataset. In most cases, synthetic techniques like SMOTE and MSMOTE will outperform the conventional oversampling and undersampling methods.

For better results, one can use synthetic sampling methods like SMOTE and MSMOTE along with advanced boosting methods like Gradient boosting and XG Boost.

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

**ANSWER:--- SMOTE:** Synthetic Minority Over sampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space. The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. SMOTE algorithm is a pioneer algorithm and many other algorithms are derived from SMOTE.

**ADASYN**:  ADAptive SYNthetic (ADASYN) is based on the idea of adaptively generating minority data samples according to their distributions using K nearest neighbor. The algorithm adaptively updates the distribution and there are no assumptions made for the underlying distribution of the data.  The algorithm uses Euclidean distance for KNN Algorithm. 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.

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

**ANSWER:---** Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

No, it is not preferable to use GridSearchCV in case of large datasets.

Grid Search can be thought of as an exhaustive search for selecting a model. In Grid Search, the data scientist sets up a grid of hyperparameter values and for each combination, trains a model and scores on the testing data. In this approach, every combination of hyperparameter values is tried which can be very inefficient.

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

## ANSWER:--- R Square,Adjusted R Square, MSE, RMSE

R-squared (R2) 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](https://www.investopedia.com/terms/r/regression.asp)model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable.

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it’s usually not.  It is always lower than the R-squared.

MSE is the average of the squared error that is used as the loss function for least squares regression: It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points. RMSE is the square root of MSE.