Worksheet Set-8 Machine Learning

1. What is the advantage of hierarchical clustering over K-means clustering?
   Ans. D) None of these
2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
   Ans. A) max_depth
3. Which of the following is the least preferable resampling method in handling imbalance datasets?
   Ans. C) RandomUnderSampler
4. Which of the following statements is/are true about “Type-1” and “Type-2” errors?
   Ans. D) 2 and 3
5. Arrange the steps of k-means algorithm in the order in which they occur:
   Ans. 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?
   Ans. B) Support Vector Machines
7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
   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)
8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?
   Ans. 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?
   Ans. 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?
    Ans. A) Overfitting & B) Multicollinearity
11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?
    Ans. The disadvantage of one hot encoding is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.
    In such case we can use label encoding technique to encode the categorical variables.
12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.
    Ans. There has been two different approaches to addressing imbalanced data: algorithm-level and data-level approach.

    Algorithm approach: As mentioned above, ML algorithms penalize False Positives and False Negatives equally. A way to counter that is to modify the algorithm itself to

boost predictive performance on minority class. This can be executed through either recognition-based learning or cost-sensitive learning.

Data approach: This consists of re-sampling the data in order to mitigate the effect caused by class imbalance. The data approach has gained popular acceptance among practitioners as it is more flexible and allows for the use of latest algorithms. The two most common techniques are over-sampling and under-sampling.

Over sampling: Over-sampling increases the number of minority class members in the training set. The advantage of over-sampling is that no information from the original training set is lost, as all observations from the minority and majority classes are kept. On the other hand, it is prone to overfitting.

Under sampling: Under-sampling, on contrary to over-sampling, aims to reduce the number of majority samples to balance the class distribution. Since it is removing observations from the original data set, it might discard useful information.

EditedNearestNeighbours under-sampling technique (E2_ENN): The ENN method was proposed by Wilson (1972), in which a majority instance is removed if its class label does not agree with its K nearest neighbors. The ENN method tends to omit the noisy and borderline instances, which will therefore enhance the accuracy of decision boundary.

Near Miss 3 under-sampling technique (E3_NM): NearMiss-3 belongs to the NearMiss family, which conducts under-sampling on the majority class according to their distance. NearMiss-3 in particular removes majority samples with the largest distance from minority samples' K nearest neighbors.

SMOTE over-sampling technique (E4_SMT): SMOTE first considers the K nearest neighbors of the minority instances. It then constructs feature space vectors between these K neighbors, generating new synthetic data points on the lines.

ADASYN over-sampling technique (E5_ADS): Very similar to SMOTE, ADYSYN also creates synthetic data points with feature space vectors. However, for the new data points to be realistic, ADYSYN adds a small error to the data points to allow for some variance. This is because observations are not perfectly correlated in real life.

13. What is the difference between SMOTE and ADASYN sampling techniques?
Ans. The key difference between ADASYN and SMOTE is that the ADASYN 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. The SMOTE generates the same number of synthetic samples for each original minority sample.

Worksheet Set-8 Machine Learning

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?
    Ans. GridSearchCV tries all the combinations of the values passed in the dictionary and evaluates the model for each combination using the Cross-Validation method. Hence after using this function, we get accuracy/loss for every combination of hyperparameters and we can choose the one with the best performance.
    One of the drawbacks of grid search is that when it comes to dimensionality, it suffers when evaluating the number of hyperparameters grows exponentially. However, there is no guarantee that the search will produce the perfect solution, as it usually finds one by aliasing around the right set.
    However, we can use Random search instead, Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. It is similar to grid search, and yet it has proven to yield better results comparatively. The drawback of random search is that it yields high variance during computing. Since the selection of parameters is completely random; and since no intelligence is used to sample these combinations, luck plays its part.
15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief
    Ans. There are 3 main metrics for model evaluation in regression:

    R Square/Adjusted R Square

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

    Mean Absolute Error(MAE)

    R Square/Adjusted R Square:- R Square measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square.

    R Square is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value is between 0 to 1 and bigger value indicates a better fit between prediction and actual value.

    R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalise additional independent variables added to the model and adjust the metric to prevent overfitting issue.

Worksheet Set-8 Machine Learning

Mean Square Error(MSE)/Root Mean Square Error(RMSE):- While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret much insights from one single result but it gives you an real number to compare against other model results and help you select the best regression model.

Root Mean Square Error (RMSE) is the square root of MSE. It is used more commonly than MSE because firstly sometimes MSE value can be too big to compare easily. Secondly, MSE is calculated by the square of error, and thus square root brings it back to the same level of prediction error and make it easier for interpretation.

Mean Absolute Error (MAE):- Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of absolute value of error.

Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalization to big prediction error by square it while MAE treats all errors the same.

Worksheet Set-8 Statistics

1. In hypothesis testing, type II error is represented by β and the power of the test is 1−β then β is:

Ans. b. The probability of failing to reject H0 when H1 is true

2. In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the

Ans. b. null hypothesis

3. When the null hypothesis has been true, but the sample information has resulted in the rejection of the null, a

_____ has been made

Ans. d. Type I error

4.For finding the p-value when the population standard deviation is unknown, if it is reasonable to assume that the

population is normal, we use

Ans. b. the t distribution with n - 1 degrees of freedom

5. A Type II error is the error of

Ans. a. accepting Ho when it is false

6. A hypothesis test in which rejection of the null hypothesis occurs for values of the point estimator in either tail of

the sampling distribution is called

Ans. d. a two-tailed test

7. In hypothesis testing, the level of significance is

Ans. b. the probability of committing a Type I error

8. In hypothesis testing, b is

Ans. a. the probability of committing a Type II error

9. When testing the following hypotheses at an α level of significance

H0: p = 0.7

H1: p > 0.7

The null hypothesis will be rejected if the test statistic Z is

Ans. a. $z > z\alpha$

10. Which of the following does not need to be known in order to compute the P-value?

Worksheet Set-8 Statistics

Ans. c. the level of significance

11. The maximum probability of a Type I error that the decision maker will tolerate is called the

Ans. a. level of significance

12. For t distribution, increasing the sample size, the effect will be on

Ans. d. All of above

13. What is Anova in SPSS?

Ans. Analysis of Variance, i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations.

ANOVA in SPSS must have a dependent variable which should be metric (measured using an interval or ratio scale). ANOVA in SPSS must also have one or more independent variables, which should be categorical in nature. In ANOVA in SPSS, categorical independent variables are called factors. A particular combination of factor levels, or categories, is called a treatment.

14. What are the assumptions of Anova?

Ans. To use the ANOVA test we made the following assumptions:

- Each group sample is drawn from a normally distributed population
- All populations have a common variance
- All samples are drawn independently of each other
- Within each sample, the observations are sampled randomly and independently of each other
- Factor effects are additive

15. What is the difference between one-way Anova and two-way Anova?

Ans. A hypothesis test that enables us to test the equality of three or more means simultaneously using variance is called One way ANOVA. A statistical technique in which the interrelationship between factors, influencing variable can be studied for effective decision making, is called Two-way ANOVA.

There is only one factor or independent variable in one way ANOVA whereas in the case of two-way ANOVA there are two independent variables.

One-way ANOVA compares three or more levels (conditions) of one factor. On the other hand, two-way ANOVA compares the effect of multiple levels of two factors.

Worksheet Set-8 Statistics

In one-way ANOVA, the number of observations need not be same in each group whereas it should be same in the case of two-way ANOVA.

One-way ANOVA need to satisfy only two principles of design of experiments, i.e., replication and randomization. As opposed to Two-way ANOVA, which meets all three principles of design of experiments which are replication, randomization, and local control.