ASSIGNMENT_NO_8

MACHINE LEARNING

Question No	Answer
1	B) In hierarchical clustering you don't need to assign number of clusters in beginning
2	A) max_depth
3	C) RandomUnderSampler
4	C) 1 and 3
5	D) 1-3-2
6	B) Support Vector Machines
7	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	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	B) remove only one of the features C) Use ridge regularization
10	A) Overfitting D) Outliers

Q11 Answer: -

In which situation One-hot encoding must be avoided...?

- 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.

Let us understand first what is Categorical data, it is non-numeric and often can be characterized into categories or groups. A simple example is colour; red, blue, and yellow are all distinct colours. Another example could be age groups or other interval-type data. Like 1–25 years old, 25–50 years old, and so on. The data used represents numbers, but the intervals themselves are categorical.

Let us understand Which encoding technique can be used in such a case?
Label encoding is probably the most basic type of categorical feature encoding method after one-hot encoding. Label encoding does not add any extra columns to the data but instead assigns a number to each unique value in a feature. Let us use the colours example again. Instead of adding a column for red, another one for blue, and one more for yellow, we just assign each value a number. Red is 1, blue is 2, and yellow is 3. We saved a lot of room and do not add more columns to our data, resulting in a much cleaner look for the data. The numbers assigned for

red, blue, and yellow are arbitrary and their labels have no actual meaning, but they are simple to deal with.

Q12 Answer: -

- Cost-Sensitive Learning Technique: - The Cost-Sensitive Learning (CSL) takes the misclassification costs into consideration by minimising the total cost. The goal of this technique is mainly to pursue a high accuracy of classifying examples into a set of known classes. It is playing as one of the important roles in the machine learning algorithms including the real-world data mining applications. In this technique, the costs of false positive (FP), false negative (FN), true positive (TP), and true negative (TN) can be represented in a cost matrix as shown below where C (I, J) represents the misclassification cost of classifying an instance and also "i" the predicted class and "j" is the actual class. Here is an example of cost matrix for binary classification.

Advantages Cost-Sensitive Learning Technique: -

- This technique avoids pre-selection of parameters and auto-adjust the decision hyperplane.

- Ensemble Learning Techniques: - The ensemble-based method is another technique which is used to deal with imbalanced data sets, and the ensemble technique is combined the result or performance of several classifiers to improve the performance of single classifier. This method modifies the generalisation ability of individual classifiers by assembling various classifiers. It mainly combines the outputs of multiple base learners. There are various approaches in ensemble learning such as Bagging, Boosting, etc. Bagging or Bootstrap Aggregating tries to implement similar learners on a smaller dataset and then takes a mean of all the predictions. The Boosting (Adaboost) is an iterative technique that rectifies the weight of an observation depending on the last classification. This method decreases the bias error and builds strong predictive models.

Advantages Ensemble Learning Techniques: -

- This is a more stable model.

- The prediction is better.

- Combined Class Methods: - In this type of method, various methods are fused together to get a better result to handle imbalance data. For instance, like SMOTE can be fused with other methods like MSMOTE (Modified SMOTE), SMOTEENN (SMOTE with

Edited Nearest Neighbours), SMOTE-TL, SMOTE-EL, etc. to eliminate noise in the imbalanced data sets. However, the MSMOTE is the modified version of SMOTE which classifies the samples of minority classes into three groups such as security samples, latent nose samples, and border samples.

Advantages Combined Class Methods: -

- No loss of useful information.
- Good generalisation.

❖ Random Oversampling Imbalanced Datasets

Random oversampling involves randomly duplicating examples from the minority class and adding them to the training dataset. Examples from the training dataset are selected randomly with replacement. This means that examples from the minority class can be chosen and added to the new "more balanced" training dataset multiple times; they are selected from the original training dataset, added to the new training dataset, and then returned or "replaced" in the original dataset, allowing them to be selected again.

This technique can be effective for those machine learning algorithms that are affected by a skewed distribution and where multiple duplicate examples for a given class can influence the fit of the model.

❖ Random Under sampling Imbalanced Datasets

Random under sampling involves randomly selecting examples from the majority class to delete from the training dataset. This has the effect of reducing the number of examples in the majority class in the transformed version of the training dataset. This process can be repeated until the desired class distribution is achieved, such as an equal number of examples for each class. This approach may be more suitable for those datasets where there is a class imbalance although a sufficient number of examples in the minority class, such a useful model can be fit. A limitation of under sampling is that examples from the majority class are deleted that may be useful, important, or perhaps critical to fitting a robust decision boundary.

Q13 Answer: -

Sr.No.	SMOTE	ADASYN
1	SMOTE stands for Synthetic Minority Over-sampling Technique.	ADASYN stands for the adaptive synthetic sampling approach, or ADASYN algorithm
2	First it finds the n-nearest neighbours in the minority class for each of the samples in the class.	ADASYN algorithm, builds on the methodology of SMOTE, by shifting the importance of the classification boundary to those minority classes which are difficult.
3	Then it draws a line between the neighbour's an generates random points on the lines. then draws a line to each of them. Then create samples on the lines with class == minority class.	ADASYN uses a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn.

Q14 Answer: -

GridSearchCV is a technique for finding the optimal parameter values from a given set of parameters in a grid. It is essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made. GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. GridSearchCV is also known as GridSearch cross-validation, an internal cross-validation technique is used to calculate the score for each combination of parameters on the grid.

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it is not practically feasible. One can shift to Random Search CV where the algorithm will

randomly choose the combination of parameters. The performance of Random Search CV is somewhat equal or less than Grid Search CV but the time complexity decreases and it becomes practically feasible to apply Grid Search of a large-size dataset.

Q15 Answer: -

Once we build our regression model, how can we measure the goodness of fit? We have various regression evaluation metrics to measure how well our model fits the data.

- MEAN SQUARED ERROR: The first metric we are going to see is the mean squared error. It calculates the average of the square of the errors between the actual and the predicted values. Lower the value, better the regression model. Here yi denotes the true score for the ith data point, and ŷi indicates the predicted value and n is the number of data points.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- ROOT MEAN SQUARED ERROR: RMSE is the most popular metric to measure the error of a regression model. This metric is calculated as the square root of the average squared distance between the actual and the predicted values. Taking the square root of the mean squared error will give you RMSE. Since we are, taking the square root it reverts the unit of measurement to its original scale. It can be used to compare models only whose errors are measured in the same units.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- MEAN ABSOLUTE ERROR: - It is calculated as the mean of the absolute difference between the actual and the predicted values. Where ŷi is the predicted value of the ith sample, and yi is the corresponding actual value, and N is the number of samples. Both RMSE and MAE are scale dependent and can be used to compare models only if they are measured in the same units. To compare models with different units, we can use metrics like MAPE or RAE. t is calculated as the mean of the absolute difference between the actual and the predicted values.

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i|$$

- MEAN ABSOLUTE PERCENTAGE ERROR(MAPE): - MAPE measures the error in percentage terms. MAPE is calculated as the absolute difference between the actual and predicted values divide over every observation. It is multiplied by 100 to make it a percentage error. Where n is the size of the sample, ŷt is the value predicted by the model, and yt is the actual value.

$$MAPE = \frac{1}{\mathrm{n}}\sum_{t=1}^{\mathrm{n}}\frac{|\hat{y}_t - \mathrm{y}_t|}{\mathrm{y}_t} \times 100$$

- RELATIVE ABSOLUTE ERROR: - RAE is defined as the ratio between the sum of absolute errors and the sum of absolute deviations. pi is the predicted value, and ai is the actual value, and a_bar is the mean of actual values.

$$RAE = \frac{\sum_{i=1}^{n}|p_i - a_i|}{\sum_{i=1}^{n}|\overline{a} - a_i|}$$

- R-SQUARE: - R-square, also known as the coefficient of determination, is one of the commonly used regression evaluation metrics. It measures the proportion of variance of the dependent variable explained by the independent variable. If the R-squared value is 0.90, then we can say that the independent variables have explained 90% of the variance in the dependent variable. It ranges from 0 to 1, where 0 indicates that the fit is poor. It is determined as the ratio of the sum of squares and the total sum of squares. However, the problem with r-square is that the value spuriously increases as a greater number of independent variables are added.

$$R^2 = 1 - \frac{SSE}{SST}$$

where SSE is the sum of squared errors and computed as,

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

and SST (total sum of squares) is given as,

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

- ADJUSTED R-SQUARE: - To counter the problem which is faced by r-square, adjusted r-square penalizes adding more independent variables which do not increase the explanatory power of the regression model. The value of adjusted r-square is always less than or equal to the value of r-square. It ranges from 0 to 1, the closer the value is to 1, the better it is. n = the sample size k = the number of independent variables

$$R^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-k-1}\right]$$

ASSIGNMENT_NO_8

STATISTICS WORKSHEET-8

Question No	Answer
1	b. The probability of failing to reject H0 when H1 is true
2	b. null hypothesis
3	d. Type I error
4	a. the z distribution
5	a. accepting Ho when it is false
6	d. a two-tailed test
7	d. none of the above
8	c. the probability of either a Type I or Type II, depending on the hypothesis to be test
9	b. $z < z\alpha$
10	c. the level of significance
11	a. level of significance
12	d. All of the Above

Q13 Answer: -

An ANOVA test is a statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using a variance. Another Key part of ANOVA is that it splits the independent variable into two or more groups. For example, one or more groups might be expected to influence the dependent variable, while the other group is used as a control group and is not expected to influence the dependent variable.

Q14 Answer: -

The assumptions of the ANOVA test are the same as the general assumptions for any parametric test,

- An ANOVA can only be conducted if there is no relationship between the subjects in each sample. This means that subjects in the first group cannot also be in the second group (e.g., independent samples/between groups).
- The different groups/levels must have equal sample sizes.
- An ANOVA can only be conducted if the dependent variable is normally distributed so that the middle scores are the most frequent and the extreme scores are the least frequent.
- Population variances must be equal (i.e., homoscedastic). Homogeneity of variance means that the deviation of scores (measured by the range or standard deviation, for example) is similar between populations.

Q15 Answer: -

Sr.No.	One-way anova test	Two-way anova test
1	A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.	A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.
2	The independent variable divides cases into two or more mutually exclusive levels, categories, or groups.	The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.
3	An example of a one-way ANOVA includes testing a therapeutic intervention (CBT,	An example of factorial ANOVAs includes testing the effects of social contact

	medication, placebo) on the incidence of depression in a clinical sample.	(high, medium, low), job status (employed, self-employed, unemployed, retired), and family history (no family history, some family history) on the incidence of depression in a population.