MACHINE LEARNING

1. D
2. A
3. D
4. C
5. D
6. D
7. C
8. C
9. A
10. A
11. **In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?**
    The disadvantage is that for high cardinality, the feature space can really blow up quickly and you start fighting with the curse of dimensionality.
12. **In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly**
    Imbalance data distribution is an important part of machine learning workflow. An imbalanced dataset means instances of one of the two classes is higher than the other, in another way, the number of observations is not the same for all the classes in a classification dataset. This problem is faced not only in the binary class data but also in the multi-class data.
    In this article, we list some important techniques that will help you to deal with your imbalanced data.

1. Oversampling

This technique is used to modify the unequal data classes to create balanced datasets. When the quantity of data is insufficient, the oversampling method tries to balance by incrementing the size of rare samples.

A primary technique used in oversampling is SMOTE (Synthetic Minority Over-sampling TEchnique). In this technique, the minority class is over-sampled by producing synthetic examples rather than by over-sampling with replacement and for each minority class observation, it calculates the k nearest neighbours (k-NN). But this technique is limited to an assumption that local space between any two positive instances belongs to the minority class, which may not always true in the case when the training data is not linearly separable.

Depending upon the amount of oversampling required, neighbours from k-NN are randomly chosen.

Advantages

- No loss of information
- Mitigate overfitting caused by oversampling.

2.Undersampling

Unlike oversampling, this technique balances the imbalance dataset by reducing the size of the class which is in abundance. There are various methods for classification problems such as cluster centroids and Tomek links. The cluster centroid methods replace the cluster of samples by the cluster centroid of a K-means algorithm and the Tomek link method removes unwanted overlap between classes until all minimally distanced nearest neighbours are of the same class.

Advantages

- Run-time can be improved by decreasing the amount of training dataset.
- Helps in solving the memory problems

3.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

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

4.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

- This is a more stable model
- The prediction is better

5. 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

- No loss of useful information
- Good generalization

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

14. **What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?**
    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.
    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. So, in the end, you can select the best parameters from the listed hyperparameters

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

There are 3 main metrics for model evaluation in regression:

R Square/Adjusted R Square: R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

- Mean Square Error(MSE)/Root Mean Square Error(RMSE): **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.
- Mean Absolute Error(MAE): The mean absolute error (MAE) is the simplest regression error metric to understand. We'll calculate the residual for every data point, taking only the absolute value of each so that negative and positive residuals do not cancel out. We then take the average of all these residuals.

PYTHON – WORKSHEET 1

1. C
2. C
3. C
4. A
5. D
6. C
7. A
8. C
9. B &C
10. A & B
11. **Write a python program to find the factorial of a number.**

```
Num=int(input("Enter a number: "))
Factorial=1
For i in range(1,Num+1):
Factorial=factorial*i
Print("The factorial of the", num , "is", factorial)
```

12. **Write a python program to find whether a number is prime or composite**

```
num = int(input("Enter a number: "))
if num >1:
    for i in range(2,number):
    if (num%i)=0:
     print(num, " is not prime")
     break
else:
   (print(num, "is prime)
```

13. **Write a python program to check whether a given string is palindrome or not**

```
String= input(("Enter a string:"))
if (string==string[::-1]):
    print("The string is palindrome")
Else("The string is not palindrome")
```

14. **Write a Python program to get the third side of right-angled triangle from two given sides**

```
from math import sqrt
print("Input lengths of shorter triangle sides:")
a = float(input("a: "))
b = float(input("b: "))

c = sqrt(a**2 + b**2)
print("The length of the hypotenuse is", c )
```

15. Write a python program to print the frequency of each of the characters present in a given string

```
input_string = input((“Enter a string:”))
frequencies = {}
for char in input_string:
if char in frequencies:
    frequencies[char] += 1
 else:
    frequencies[char] = 1
```

STATISTICS WORKSHEET-8

1. B
2. **B**
3. **A**
4. **A**
5. **A**
6. **D**
7. **A**
8. **A**
9. **D**
10. **D**
11. **A**
12. **D**

13. What is Anova in SPSS?

Statistical Analysis. 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.

14. What are the assumptions of Anova?

The factorial **ANOVA** has a several **assumptions** that need to be fulfilled – (1) interval data of the dependent variable, (2) normality, (3) homoscedasticity, and (4) no multicollinearity.

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

The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

- **One-way ANOVA**: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.
- **Two-way ANOVA**: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master's), and race finishing times in a marathon.

All ANOVAs are designed to test for differences among three or more groups. If you are only testing for a difference between two groups, use a t-test instead.