Experiment - 05 (Group (A))

Title: Data Analytics - II.

* **Aim:** 1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset.
2. Compute confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, precision, Recall on the given dataset.

- Objective of the assignment - Students should be able to data analysis using logistic regression using python for any open source dataset.
- Pre-requisite - 1. Basic of python programming.
2. Concept of regression.

THEORY :

I) Logistic Regression -

- Classification techniques are an essential part of machine learning and data mining application.
- Approximately 70% of problems in data science are classification problems.
- Another method for classification is multinomial classification, which handles the issues where multiple classes are present in target variable.

Page No.
Date

For example- iris dataset is an example of multinomial classification. Other examples are classifying article/blog/document catagories.

- Logistic regression can be used for various classification problems such as spam detection.
- Diabetes prediction if a given customer will purchase a particular product or will they churn another competitor, whether the user will click on given advertisement link or not.

- Logistic regression is a statistical method for predicting binary classes.
- The outcome or target variable is dichotomous in nature.
- Dichotomous means there are only two possible classes. For example it can be used for cancer detection problems.
- It computes the probability of an event occuring.
- It is a special case of linear regression where the target variable is categorical in nature.
- It uses a log of odds as the dependent variable. Logistic regression predicts the probability of occurence of a binary event utilizing a logit function.

II] Linear Regression Equation -

$$y = \beta 0 + \beta 1 x1 + \beta_2 x2 + \cdots + \beta n xn.$$

Where y is dependent variable and x_1, x_2 x_n are explanatory variables.

Sigmoid Function - $p = 1/1 + e^{-y}.$

Apply sigmoid function on linear regression :

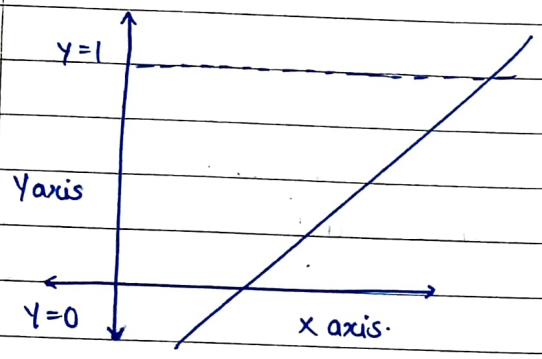
$$p = 1/1 + e^{-(\beta 0 + \beta 1 x1 + \beta_2 x2 + \cdots + \beta n x0)}$$

3] Differente between linear and logistic regression -

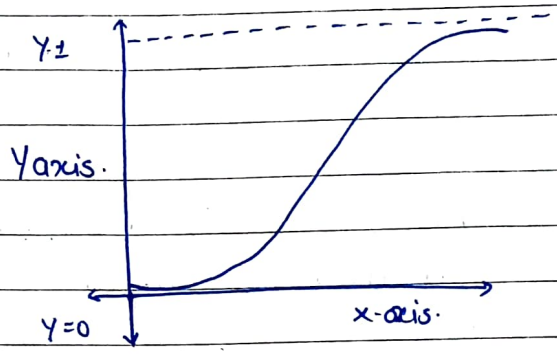
- Linear regression gives you a continuous output, but logistic regression provides a constant output
  An example of continuous output is house price and stock price.

- Examples of the discrete output is predicting whether a patient has cancer or not, predicting whether a patient has cancer or not, predicting whether customer will churn.

- Linear regression is estimated using Ordinary Least Squares (OLS) while logistic regression is estimated using Maximum likelihood estimation (MLE).



Linear regression



Logistic regression.

* Sigmoid Function –

1. The sigmoid function, also called logistic function gives an 's' shaped curve that can take any real-valued number and map it into a value between 0 and 1.

2. If the curve goes to positive infinity, y will predicted be 1. and if curve goes to negative infinity, y predicted will become 0.

3. For example: if output is 0.75 we can say that there is 75 percent chance that patient will suffer from cancer.

$$f(x) = \frac{1}{1 + e^{(-x)}}.$$

4] Types of Logistic regression –

1. Binary Logistic Regression - The target variable has only two possible outcomes such as Spam or not spam, cancer or no cancer.

2. Multinomial Logistic regression - The target variable has three or more nominal categories such as predicting the type of wine.

3. Ordinal Logistic regression - the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.

5] Confusion Matrix Evaluation Metrics.

Contingency table or confusion matrix is often used to measure the performance of classifiers. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.

The following table shows the confusion matrix for a two class classifier.:

n	predicted	
actual	TP	FN
	FP	TN

P
N

Confusion Matrix.

- Here each row indicates the actual classes recorded in the test data set and the each column indicates the classes as predicted by the classifier.
- Numbers on the descending diagonal indicate correct predictions, while the ascending diagonal concerns predictions errors.

* Some important measures derived from confusion matrix are:

- Number of positive (Pos): Total number instances which are labelled as positive in a given dataset.
- Number of negative (Neg): Total number instances which are labelled as negative in a given dataset.
- Number of true positive (TP): Number of instances which are actually labelled as positive and predicted class by classifier is also positive.
- Number of True negative (TN): Number of instances which are actually labelled as negative and the predicted class is also negative
- Number of False Positive (FP): Number of instances which are actually labelled as negative and predicted class by classifier is positive.
- Number of False Negative (FN): Number of instances which are actually labelled as positive and the class predicted by the classifier is negative.

- **Accuracy**: Accuracy is calculated as the number of correctly classified instances divided by total number of instances.
The ideal value of accuracy is 1, and worst is 0. It is also calculated as the sum of true positive and true negative (TP + TN) divided by the total number of instance.

$$acc = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{Pos + Neg}$$

- **Error Rate**: Error rate is calculated as the number of incorrectly classified instances divided by total number of instances.
The ideal value of accuracy is 0 and the worst is 1. It is also calculated as the sum of false positive and false negative divided by total number of instances.

$$err = \frac{FP + FN}{TP + FP + TN + FN} = \frac{FP + FN}{Pos + Neg} \quad \text{or}$$

$$err = 1 - acc.$$

- **Precision**: It is calculated as number of correctly classified positive instances divided by the total number of instances which are predicted positive.
It is also called confidence value. The ideal value is 1 whereas worst is 0.

$$precision = \frac{TP}{TP + FP}$$

- **Recall**: It is calculated as the number of correctly classified positive instances divided by the total number of positive instances. It is also called recall or sensitivity. The ideal value of sensitivity is 1 worst is 0.

$$\text{Recall} = \frac{TP}{TP + FN}$$

6] Stepwise Algorithm for logistic regression on social media Adv. dataset.

Step 1: Import required libraries.

```
import numpy as np.
import pandas as pd.
import matplotlib.pyplot as plt.
from sklearn.model_selection import train-test-split
from sklearn.preprocessing import StandardScaler.
from sklearn.linearmodel import LogisticRegression.
```

Step 2: Load the dataset -

```
data = pd.read-csv ("Social-media.csv");
```

Step 3: Perform EDA.

- head
- tail
- info
- describe
- describe (include = "all")
- shape
- size
- ndim
- columns.

Step 4: Handle null values if any.

Step 5: Split data into independent variables (x), (y).

```
x = data.drop(['Purchased'], axis=1)
y = data['Purchased']
```

Step 6: Split dataset into training and testing sets.

```
xtrain, xtest, ytrain, ytest = train_test_split(x, y,
test_size = 0.2, random_state = 0).
```

Step 7: Train the logistic regression model.

```
logreg := LogisticRegression()
logreg.fit(xtrain, ytrain).
```

Step 8: Make predictions.

```
ytrain_pred = logreg.predict(xtrain).
ytest_pred = logreg.predict(xtest).
```

Step 9: Evaluate Model performance.

```
cm = confusion_matrix(ytest, ytest_pred).
print("Confusion Matrix:", cm)
```

Step 10: Calculate evaluation parameters.

```
accuracy = accuracy_score(ytest, ytest_pred).
error_rate = 1 - accuracy.
precision = precision_score(ytest, ytest_pred.)
recall = recall_score(ytest, ytest_pred)
report = classification_report(ytest, ytest_pred).
f1 = f1_score(ytest, ytest_pred).
```

Step 11: Now print all above values.
If neccesary plot them.

* **Conclusion:** In this way we have done data analysis using logistic regression for social Media Adv. and evaluate the performance of model.

* * * * * * *