2020_1125_Logistic_Regression

November 27, 2020

1 Classifying Pulsars from the High Time Resolution Universe Survey (HTRU2) - Logistic Regression

1.1 Overview & Citation

In this code notebook, we attempt to classify pulsars from the High Time Resolution Universe Survey, South (HTRU2) dataset using logistic regression. The dataset was retrieved from the UC Irvine Machine Learning Repository at the following link: https://archive.ics.uci.edu/ml/datasets/HTRU2#.

The dataset was donated to the UCI Repository by Dr. Robert Lyon of The University of Manchester, United Kingdom. The two papers requested for citation in the description are listed below:

- R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656
- R. J. Lyon, HTRU2, DOI: 10.6084/m9.figshare.3080389.v1.

1.2 Import the Relevant Libraries

[34]:
```
# Data Manipulation
import pandas as pd
import numpy as np

# Modeling & Evaluation
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix,classification_report
```

1.3 Import & Check the Data

[3]:
```
df = pd.read_csv('2020_1125_Pulsar_Data.csv')
pulsar_data = df.copy()
```

[4]:
```
pulsar_data.head()
```

[4]:

	IP_Mean	IP_StdDev	IP_Kurtosis	IP_Skewness	DM_Mean	DM_StdDev	\
0	140.562500	55.683782	-0.234571	-0.699648	3.199833	19.110426	
1	102.507812	58.882430	0.465318	-0.515088	1.677258	14.860146	
2	103.015625	39.341649	0.323328	1.051164	3.121237	21.744669	
3	136.750000	57.178449	-0.068415	-0.636238	3.642977	20.959280	
4	88.726562	40.672225	0.600866	1.123492	1.178930	11.468720	

	DM_Kurtosis	DM_Skewness	Class
0	7.975532	74.242225	0
1	10.576487	127.393580	0
2	7.735822	63.171909	0
3	6.896499	53.593661	0
4	14.269573	252.567306	0

1.4 Train Test Split

[6]:
```
X = pulsar_data.drop('Class',axis=1)
y = pulsar_data['Class']
```

[7]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,␣
↪random_state=42)
```

[10]:
```
logmodel = LogisticRegression(max_iter=200) # Max iterations = 200 since␣
↪default 100 does not work
logmodel.fit(X_train,y_train)
```

[10]:
```
LogisticRegression(max_iter=200)
```

[12]:
```
y_pred = logmodel.predict(X_test)
```

1.5 Model Evaluation

[29]:
```
confusion = confusion_matrix(y_test,y_pred)
print(f'CONFUSION MATRIX:
↪\n\n{confusion[0][0]}\t{confusion[0][1]}\n{confusion[1][0]}\t{confusion[1][1]}')
```

```
CONFUSION MATRIX:

4049    21
77      328
```

[28]:
```
print(f'CLASSIFICATION REPORT:\n\n{classification_report(y_test,y_pred)}')
```

```
CLASSIFICATION REPORT:

              precision    recall  f1-score   support
```

```
              0       0.98      0.99      0.99      4070
              1       0.94      0.81      0.87       405

    accuracy                            0.98      4475
   macro avg       0.96      0.90      0.93      4475
weighted avg       0.98      0.98      0.98      4475
```

The dataset contains a total of 1,639 actual pulsars out of 16,259 instances in the dataset (approximately 10%). This means that we have an unbalanced classification problem, and accuracy is not a good metric. Therefore, the most important metrics for predicting a pulsar with this model are: * Precision = 0.94 * Recall = 0.81 * F1-Score = 0.87

Let's save this data to a .csv file for future comparison with the other classification models:

[33]:
```
with open("2020_1125_Logistic_Results.csv","w") as file:
    file.write('Model,Accuracy,Precision,Recall,F1-Score\n')
    file.write('Logistic,0.98,0.94,0.81,0.87\n')
```