TECHNOLOGICAL UNIVERSITY OF THE PHILIPPINES – MANILA
College of Engineering
Electronics Engineering Department
Ayala Blvd., Ermita, Manila, 1000, Philippines
Tel. No. +632-301-3001 | Fax No. +632-521-4063

Programming Assignment 3

LOGISTIC REGRESSION

In this assignment, you will train a model using Logistic Regression. Go to https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 and download the download the Wisconsin Breast Cancer dataset. The dataset contains 699 instances of breast tumors with the following attributes:

1. Sample code number id number
2. Clump Thickness 1 - 10
3. Uniformity of Cell Size 1 – 10
4. Uniformity of Cell Shape 1 - 10
5. Marginal Adhesion 1 - 10
6. Single Epithelial Cell Size 1 - 10
7. Bare Nuclei 1 - 10
8. Bland Chromatin 1 - 10
9. Normal Nucleoli 1 - 10
10. Mitoses 1 - 10
11. Class: (2 for benign, 4 for malignant)

From the raw data set, remove rows with missing values, remove the column "Sample code number" and replace the "Class" values into 0's and 1's (0 for benign, 1 for malignant). You may choose to do this in Python (Pandas) or manually in spreadsheet application. The goal is to make a classifier for the tumor status.

General Guidelines

1. Split the samples into 70% Training and 30% Testing at random.
2. Use `stratify=y` in the *test_train_split* function. Build a pipeline using the Standard scaler and logistic regression.
3. Use the default penalty settings of Logistic Regression.
4. After fitting the data, what is the model's training and testing accuracy? Which features are most important?
5. Generate a confusion matrix, then calculate the other metrics: F1-score, Precision, Recall, and False alarm rate.
6. Plot the ROC curve and report the AUC. For this item, make a result for both the training and testing data, separately.

Guide Questions

You are expected to answer the following questions using your analysis:

1. What steps are required to remove the "Sample code number" column?
2. How can the "Class" values be converted into binary (0 and 1) form in Python?
3. What is the importance of splitting the dataset into training and testing sets?
4. Why is it necessary to use the stratify parameter in the train_test_split function?
5. What is logistic regression, and why is it suitable for this dataset?
6. How can the training and testing accuracy of the logistic regression model be calculated?

7. Which features in the dataset are most influential in determining the tumor status, and how can their importance be assessed?
8. How is a confusion matrix generated, and what does it represent?
9. How are precision, recall, F1-score, and false alarm rate calculated from the confusion matrix?
10. Why are these metrics important for evaluating the performance of a classifier?
11. What is an ROC curve, and how is it plotted for a logistic regression model?
12. How is the AUC (Area Under the Curve) calculated, and what does it signify about the model's performance?
13. How do the training and testing ROC curves compare, and what insights can be derived from this comparison?
14. What challenges did you encounter during the preprocessing or model training phases, and how did you address them?
15. If the model's performance is not satisfactory, what adjustments could be made to improve it?

Requirements

- Ensure that your code is clean, well-commented, and organized.
- Use Python libraries such as `numpy` and `pandas` for data manipulation and `matplotlib` or `seaborn` for visualization.

Submission

1. Submit your work as a Jupyter Notebook (`.ipynb`) file.
2. Upload your Jupyter Notebook to your GitHub repository. Ensure the notebook is well-documented with markdown cells explaining each step and the corresponding results.
3. Provide the link to your GitHub repository for grading.