Data Analytics 1

Assignment 4

Classification

Release : 19th October 2023
Deadline : 30th October 2023 (11:55 pm)

In this assignment, you are required to create a multiclass classifier to classify Penguin species, based on some given attributes which is a mix of both numerical and categorical attributes. Use the data provided in train.csv to train your classifier for the column 'Species'(target variable).

Note: You are allowed to use sklearn or other libraries for implementing the metrics and individual classifiers like SVM(task 1) and decision trees(task 2) but you are **not allowed** to direct functions like sklearn.multiclass.OneVsOneClassifier or sklearn.multiclass.OneVsRestClassifier (for task 1) and sklearn.ensemble.RandomForestClassifier (for task 2)

Tasks:

1. Build two multi-class classifiers one-vs-one , one-vs-all using SVM classifier and train it on the given data(penguin_train.csv). [correctness: 30 marks, accuracy score on test.csv: 20 marks] - Consider factors such as data cleaning, data skew, handling numerical vs categorical variables.
    a. Running `teamId_classifier_ovo.py <path to test file>(will be in same format)` should output a ovo.csv file with the predicted labels by one-vs-one classifier with column names as "predicted" (all in lower case) , which will then be checked with the actual labels to determine your model's accuracy score
    b. Running `teamId_classifier_ova.py <path to test file>(will be in same format)` should output a ova.csv file with the predicted labels by one-vs-all classifier with column names as "predicted" (all in lower case), which will then be checked with the actual labels to determine your model's accuracy score

2. Build a random forest classifier using n Decision trees for the above given dataset (try different values of n) [30 marks].
    a. For this task just print the output for the predictions of test dataset in the notebook itself and print the accuracy and other metrics also in the notebook file.

3. Report and Analysis [20 marks]:
    a. **Theoretical question** :There are very less samples in the above dataset , How do you deal with that? (5 marks)
    b. Plot the confusion matrix,and include the precision, recall, f1-score metrics in the report .
    c. Compare the results obtained for one-vs-one and one-vs-all(which according to you performs better for the above dataset)

    Include all the above plots, analysis and answer for theoretical question in team_id_report.pdf

Submission instructions:

- Submit a teamId_assignment4.zip file containing the following
    - teamId_classifier_ovo.py
    - teamId_classifier_ova.py
    - teamId_random_forest.ipynb
    - teamId_report.pdf
    - And any other implementation files if needed