Data Analytics 1
Assignment 4
Classification

Release: 7 October 2024
Deadline: 21 October 2024 (11:55 pm)

In this assignment, you are required to create a multiclass classifier to classify Customers, based on some given attributes, a mix of numerical and categorical attributes. Use the data provided in train.csv to train your classifier for the column 'Segmentation' (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 use direct functions like sklearn.multiclass.OneVsOneClassifier or sklearn.multiclass.OneVsRestClassifier (for task 1) and sklearn.ensemble.RandomForestClassifier (for task 2).

Tasks

Task 1 [50 marks]

Build two multi-class classifiers one-vs-one, one-vs-all using SVM classifier and train it on the given data (Customer_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.

Task 2 [30 marks]

Build a random forest classifier using n Decision trees for the above given dataset (try different values of n).

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.

Task 3: Report and Analysis [20 marks]

a. How does class imbalance affect multiclass classification, and what strategies can be employed to mitigate its impact, especially with small datasets? (5 marks)

b. How can the choice of hyperparameters make the random forest classifier and SVM classifiers more prone to over or under fitting? (5 marks)

c. Plot the confusion matrix and include the precision, recall, f1-score metrics in the report. (5 marks)

d. Compare the results obtained for one-vs-one and one-vs-all (which according to you performs better for the above dataset). (5 marks)

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