Lab work 1: Multi Class Classification

1. **Goal**

   Study and assess the performance of the two different multi class classification approaches discussed in the lecture. You will be using a kaggle data source, the so called "wine quality" [1], a public usable data source. It provides more than 1000 samples with 11 attributes and its "Y" which designates the quality of a red wine.

   The original Kaggle data source was modified towards having now three targets Y, i.e., the wine quality or classes:
   Y=0 poor quality for {Y=2,3,4},
   Y=1 medium quality for {Y=5,6}
   Y=2 premium quality for {Y=7,8}

[1] https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

2. **After completion you have learned**

- The importance of balanced data sources
- Applying the multi class classification using scikit learn libs
- Display further assessment parameter apart from the Accuracy

3. **Tasks**

   3.1 Use original data source File
   Use "Wine_Test_02.csv"

a) Plot histogram of each attribute regarding Y=0, Y=1 and Y=2, and display the number of samples (Y) for each quality classes.
   What can you say regarding the quality (Y) classes distribution?
   What is your conclusion regarding the **expected performance** of the classifier?

b) Perform one run of modeling and test. Compare the obtained **test accuracy** by using:
   1. **One versus All Classifier**
      And
   2. **One versus One Classifier**

   Use estimator=GaussianProcessClassifier().
   You do not need to apply CV this time, one run only is sufficient.

c) Print or plot the confusion matrix
   Discuss the entries of the matrix!

d) Print Classification report for both multi class classifier solutions, i.e.,

	precision	recall	f1-score
Class 0	xy	ab	cd
Class 1			
Class 2			

e) Plot the ROC curve for all three classes (optional, but nice to have)

3.2 Use filtered data source file

Use 'Wine_Test_02_6_8_red.csv'

Remark: for this file a data source sampling has been performed.

Re-do all tasks a) – e) as stated in 3.1), this time with the new source file.

4. Submission/presentation

Each person submits via moodle three days before presentation:

a) Source code

b) One pdf page of a report containing:
- Histogram of source data
- Test confusion matrix and test accuracy, precision and recall with both data source for both **One versus All Classifier** and **One versus One Classifier.**
- A brief comment of the obtained results
- A plot of the ROC curves

c) The running code is presented and explained during the practical course sessions.

5. Remarks

Use a train/test split of 80/20

For the OneVsRestClassifier and the OneVsOneClassifier use "estimator=GaussianProcessClassifier()".

Use the proper classes of the scikit learn libs

Literature

https://scikit-learn.org/stable/modules/multiclass.html#ovo-classification
https://scikit-learn.org/stable/modules/multiclass.html
https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsOneClassifier.html

RJ/04/2024