Lab work 2: Multi Class Classification using k-nearest neighbor (knn)

1. **Goal**

   Study and assess the performance of the knn classifier for a Multiclass Problem. The goal is to compare two methods for finding the optimal knn parameters focusing upon the accuracy only. You will be using the known 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**

- Applying the knn Classifier for multi class classification using **scikit learn libs**
- Applying two methods to find the optimal Hyperparameter for the Knn Classifier.
- Assessing the employed methods regarding the accuracy.
- Interpret the ROC curve.

3. Tasks

Use 'Wine_Test_02_6_8_red.csv'

Set train/test = 0,8

You may select anyone of the **distance metrics algorithms** and **Algorithm type** (Brut force, etc.). Please state this in the report.

3.1 k value determination by minimizing the error

Here the first task is to determine the optimal k value and then see the achieved accuracy.

Carry out five runs use **random selection of the samples** within the train/test split code assignment.

a) For **each run plot the mean error of "predicted vs known and k"** and **search for the minimum error** the optimal k value using the test data, see Fig. 1.

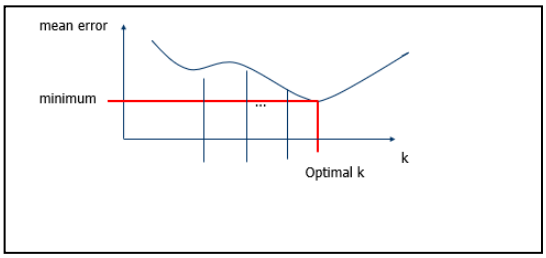

Fig. 1 mean error rate of test data

b) After five runs chose a solution to determine the final k value ?

c) With the **determined final k value from b)** perform now five **runs**, again use **random selection of sample** within the train/test split assignment and list the obtained accuracy for each run.

Calculate the mean value of the accuracy.

d) Plot the ROC curves for all three classes for **one run only.**

Remark:

For multi-class datasets, the ROC curves are plotted by dissolving the confusion matrix into one-vs-all matrices getting a confusion matrix for binary classification according to the sklearn documentation.

Discuss the obtained curves.

3.2 k value determination by employing gridsearch

Within this task you employ the gridsearch for finding the optimal k as well as the distance metric algorithm.

Parametrize for the gridsearch,

- provide a list of distance metric algorithms to be considered.
- Provide a list of Algorithm types (e.g. brut force, etc.) to be considered
- Provide a range of k values to be considered.
- Select other sklearn parameter properly.
-

a) perform **five runs and list for each run**
- selected metric algorithm
- selected Algorithm type (e.g. brut force, etc.)
- selected k value
- the achieved accuracy

b) Calculate the mean value of the accuracy.

4. Submission/presentation

Each person submits via moodle **three days before presentation**:
Source code

pdf report containing for
Add Task 3.1)
4.1) curve showing k versus the error, s. fig. 1
4.2) solution for determine the final k value
4.3) the final k value
4.4) list of the obtained accuracy for each run with the final k value
4.5) mean of achieved accuracy
4.6) ROC
Add Task 3.2)
4.7) list of k value, distance metric algorithms, Algorithm types and accuracy **for each run**
4.8) mean of achieved accuracy
4.9) comment on accuracy results obtained of **Labwork 1 ('Wine_Test_02_6_8_red.csv'.)**

3 Remarks

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