C23 Computation for Transportation Engineering (Hilary Term 2023)

Assignment 2. [about: classification, performance metrics, plotly].

Istructions: Please submit your solution (Jupyter notebook) by April 15th.

The objective of this assignment is to solve a classification task using different classifiers and to compare them according to different evaluation metrics.

Use the breast cancer dataset provided during the lectures.

1. *Consider the following classifiers : (i) SVM, (ii) Logistic Regression, (iii) NeuralNetwork class (NN) studied during the course (set 200 hidden nodes in the hidden layer, epoch num=500, learning rate=0.005). Split the dataset into training (75%) and testing (25%) and for each of the three classifiers, provide on both datasets:*
    a. *The confusion matrix (interactive plot)*
    b. *Accuracy, Precision, Recall, F1-score*

For points b., it is asked to create one interactive plot per evaluation metric, illustrating the respective performance for the three classifiers on the same axes. Note that depending on how you decide to compute the results, the custom NeuralNetwork class may require some modifications in order to perform the requested calculations (although you can also use appropriate functions from scikit-learn).

Observe if overfitting issues are present.

[4 marks]

2. *Calculate the values of Accuracy, Precision, Recall and F1-score :*
    a. *For SVM and LR, using a training set size ranging from 50% to 90% with 10% increments*
    b. *For NN, splitting the data into training, validation and test sets as follows :*
        i. *First, obtain two sets such that Training set : 70%, Test set : 30%*
        ii. *Then, within the Training set use 80% for Training and 20% for validation, to determine the best of value of the learning rate parameter (use for instance sklearn.metrics.roc_auc_score value). Investigate at least four different values.*

[4 marks]

3. *For Logistic Regression (LR) and for NeuralNetwork (NN), vary the decision threshold between 0.05 and 0.95 with 0.05 increments and compute on the test set (use 75%-25% configuration as in 1.):*
    a. *The ROC curve (one interactive plot showing both LR and NN curve)*
    b. *The associated Area Under the Curve.*
    c. *Accuracy, Precision, Recall, F1-score as a function of the threshold. (two separate plots, one with the four curves for LR and one with the four curves for NN).*
    d. *Precision vs Recall curves (one interactive plot showing both LR and NN curve)*

[4 marks]