Supervised Learning and Validation

Aim
The aim of this lab is to learn three popular ML algorithms (Gaussian Naive Bayes, Decision Trees, and Random Forests) and validation techniques for regression and classification models. This lab is divided into three parts.

To pass the lab, write down your answers to the questions in Part 1, 2 and 3. During the lab demonstration, explain your answers to the lab assistant. The lab assistant will ask you some general question about the parts of the lab and how you solved them.

Part 1
Read the tutorial Gaussian Naive Bayes Classification available on the course website. This tutorial explains NB classification and how you can implement it without using sklearn library. The objective is to dive into the algorithm and understand its parts. Execute the Python scripts GaussNB.py and look how the probabilities are implemented and used.

To pass this part explain the concepts and Gaussian NB algorithm to the lab assistant.

Part 2
Random forest algorithms have been succesfully used in solving facial recognition problems. A very well-known face recognition issue is to complete unknown parts of the face from the known parts. Please read the code in the following link carefully:
`https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_multioutput_face_completion.html`

1. Explain what the program does.
2. What is your interpretation of the final plot? Which algorithm has better performance in building the unknown parts of the face?
3. Download the code from the link above and modify it by adding the results of the following algorithms to the final plot:
   (a) Regression decision tree with max depth of 10 and max number of features 50
   (b) Regression decision tree with max depth of 20 and max number of features 50
   (c) Regression decision tree with max depth of 20 and max number of features 25
   (d) Random forest with max depth of 10 and max number of features 50
   (e) Random forest with max depth of 20 and max number of features 50
   (f) Random forest with max depth of 20 and max number of features 25

   How do you interpret the results?

4. How could performance of random forest be improved? (Hint: have a look at the example of using Haar-like feature in face detection here: `https://realpython.com/traditional-face-detection-python/`)

Part 3

Start by reading the Validation Metrics tutorial (available on the course website) which presents validation metrics for regression and classification models.

Go through the three questions below and explain your solution to the lab assistant.

1. In the script of the Regression section (the one before the section RFE), we apply cross validation with 10 folds. Note that the script does not make any change to the dataset. Modify the script in oder to reshuffle the rows of the data set to randomize the cross validation folds before applying the cross validation.

   Run again the script but on the reshuffled data set and re-calculate the MSE and R^2 scores. Do you obtain a better performance?

   To know more about reshuffling You can read the part about reshuffling in Section 3.1 "Cross-validation: evaluating estimator performance" of scikit-learn[1].

2. What happens if you do reshuffling and RFE? do you get better results than only reshuffling?

3. In the section Car Evaluation Quality, we performed the evaluation metrics for linear support vector machine, naive bayes, logistic regression and k nearest neighbours. As you can see at page 18, they do have a poor performance.

   Find out if there are ML algorithms that perform better on the data_cars.csv data set. You may test decision trees and random forest as well as other type of SVM.

[1] `https://scikit-learn.org/stable/modules/cross_validation.html`