University of Ottawa

School of Electrical Engineering and Computer Science

CSI5155 - Fall 2022

Assignment 2: Evaluation of Learning

TOTAL MARKS 100

The aim of this assignment is threefold.
First, we will explore the impact of resampling on model construction and model quality.
Second, we will continue to explore more supervised learning algorithms.
Third, we will focus on the evaluation of the results of learning.

Instruction:

1. This is an individual assignment.
2. Submit your assignment using BrightSpace, before the due date. We cannot accept late submission.
3. Submit a report on BrightSpace that contains all your results. For the implementation, you should either upload your code on BrightSpace or provide a link to a GitHub repository. Note that, if you choose to use GitHub, the date and time of last change to your repository should be **before** the assignment deadline.
4. Use Scikit-Learn to complete the assignment.

Context - In assignment 1, we used the Drug Consumption dataset from the UCI Machine Learning Repository to construct binary classification models to explore an individual's risk of drug consumption and misuse. We constructed models using four (4) different learners, namely a single decision tree (DT), a random forest (RF) learner, a support vector machine (SVM), and a k-nearest neighbor (k-NN) classifier, using the hold-out method of evaluation.

Topic: Supervised learning and Evaluation of Learning

For assignment 2, you should select the dataset that obtained the highest overall accuracy in assignment 1, when using the holdout method. We refer to this dataset as **dataset D**.

In all your evaluations, use the tenfold cross validation approach. This implies that you will need to rerun the four algorithms against **dataset D**, prior to completing the following tasks.

Answer the following questions.

1. Implement one (1) *over-sampling* method to convert **dataset D** to a balanced **dataset DB1**. [**10 marks**]

2. Retrain the four (4) classification algorithms (DT, k-NN, SVM, and RF) using **dataset DB1**. [**5 marks]**

3. Implement one (1) *under-sampling* method to convert **dataset D** to a balanced **dataset DB2**. [**10 marks**]

4. Retrain the four (4) classification algorithms (DT, k-NN, SVM, and RF) against **dataset DB2**. [**5 marks]**

5. Use the multi-layer perceptron (MLP) algorithm and the gradient boosting (GB) ensemble to construct models against **datasets D, DB1,** and **DB2**. You should aim to produce the highest possible accuracies for the algorithms, through parameter tuning [**20 marks**]

Steps 1 to 5, as listed above, will result in three different sets of experiments:
(A) - models built against the original **dataset D**, using ten-fold cross validation,
(B) - models built against the over-sampled **dataset DB1**, and
(C) - models built against the under-sampled **dataset DB2**.

6. Next, apply the six (6) algorithms to the following two (2) datasets. You should aim to produce the highest possible accuracies for the algorithms, through parameter tuning **[20 marks]**

   - https://archive.ics.uci.edu/ml/datasets/Labor+Relations, a dataset used to predict whether labor negotiations will be successful or not.
   - https://www.kaggle.com/ronitf/heart-disease-uci, a dataset that used to predict whether a patient has heart disease, denoted by the binary feature "target".

7. Create a table to show the accuracies of the six (6) algorithms against the five (5) datasets, namely the three (3) different versions of the drug consumption dataset (**datasets D, DB1** and **DB2**), as well as the **labor-relations** and **heart-disease** datasets. Show the steps you followed to determine whether there are any statistically significant differences between the results using Friedman's test, when $\alpha = 0.05$. If you find a significant difference, then show how you used the Nemenyi post hoc test to determine the critical differences. **[10 marks**]

8. Write a 300 to 400 words summary discussing the lessons you learned during this assignment. Your answer should focus on the results obtained when comparing the different sampling methods, while using the various algorithms. **[20 marks]**