University of Ottawa

School of Electrical Engineering and Computer Science

CSI5155 Fall 2020

Homework Assignment 2

Topics: Ensembles, Feature engineering, Class imbalance, Evaluation of learning

TOTAL MARKS 100

Instruction:

1. Submit your assignment on Brightspace.
2. No late assignments will be accepted.
3. This is an individual assignment.
4. Use Scikit-Learn to complete the assignment.

Question 1: Class imbalance, ensembles, and feature selection [80 = 8 x 10 marks]

For this question, we are again using the Portuguese Bank Marketing dataset from the UCI Machine Learning Repository. (The direct link is located at https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. You should again use the **bank-additional-full.csv** file with all examples (41188) and 20 inputs for this assignment.) The aim of this binary learning task is to predict whether a client will purchase a product from the bank, i.e. the output variable (desired target) is feature 21, with classes 'yes' and 'no'.

Recall that this data set is imbalanced. In this assignment, our aim is to improve the performance of the algorithms used in homework 1, apply ensemble-based methods and assess the significance of the results.

Complete the following tasks.

1. Rebalance the data set using three different approaches.
    a. Oversampling of the minority class.
    b. Under-sampling of the majority class.
    c. Balanced sampling, i.e. combining oversampling and under-sampling.

2. Apply the four algorithms (a support vector machine (SVM), a decision tree, a Naïve Bayesian learner, and the k-nearest neighbor (k-NN)), as used in homework 1 to the three resampled data sets. Report your results when using tenfold cross validation.
3. In addition, construct models using the random forests (RFs) and extreme learning trees algorithms, again using tenfold cross validation.
4. Choose the sampling method that produces the best results and motivate your answer.
5. Create a table showing the accuracies of the six algorithms against **each one of the ten folds** when trained using the sampling technique you selected in question A.4, similar to Example 12.4 in the textbook.
   Hint: Also refer to Section 12.3, and notably Example 12.6, in the textbook as well as the slides.
6. Determine whether there is a statistically significant difference in the accuracies obtained by the six algorithms against this dataset.
7. Next, apply two different feature selection techniques to the data you chose in question A.4.
8. Retrain the "best two" algorithms, as selected during question A.6, to determine whether feature selection led to improvements in accuracies. Motivate your answer by showing the average accuracies obtained during tenfold cross validation, before and after feature selection.

Part B: Comparison of algorithms - multiple datasets [20 = 10 + 10]

Consider the following three benchmarking datasets, together with the Portuguese Bank Marketing dataset.

- https://archive.ics.uci.edu/ml/datasets/Labor+Relations
- https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records
- https://archive.ics.uci.edu/ml/datasets/Iris

1. Apply only the SVM, k-NN and RFs algorithms to the three new data sets, again using tenfold cross validation, to obtain the average accuracies over the ten folds.
2. Create a table showing the average accuracies of the three algorithms against all four data sets. (For the Portuguese Bank Marketing dataset, report the most accurate results you obtained in Part A.) Use Friedman's test to determine whether there is a statistically significant difference in the accuracies obtained in question B.1, calculate the critical difference (CD) and draw the Nemenyi diagram.
   Hint: Refer to Section 12.3, and notably Example 12.8 and Figure 12.1 (top), of the textbook. Also, see the slides that contain the Friedman test values.