University of Ottawa

School of Electrical Engineering and Computer Science

Homework 2

Instructions:

1. Submit your assignment using the uOttawa Virtual Campus, before the due date.
2. No late assignments will be accepted.
3. This is an individual assignment. Recall that assignments are awarded a participation mark and that you are required to answer all the questions, in order to obtain full marks.

Part A: Seismic bumps data revisited [60 = 10 marks x 6]

In the first part of this assignment, you will reconsider the Seismic Bumps data set used in homework 1. Recall that this data set is imbalanced. Our aim is to improve the performance of the algorithms used in homework 1.

1. Rebalance the data set using three different approaches.
    a. Oversampling
    b. Under-sampling
    c. Balanced sampling, i.e. combining oversampling and under-sampling
2. Apply the four algorithms you used in homework 1 to the three resampled data sets, using tenfold cross validation.
3. Select the sampling method that produces the best results; motivate your answer.
4. Create a table showing the accuracies of the four algorithms against each one of the ten folds when trained against the sampling technique you selected in question 3. Determine whether there is a statistically significant difference in the accuracies obtained by the four algorithms.

5. Apply two different feature selection techniques to the data set you selected in question 3.
6. Retrain the four algorithms to determine whether feature selection leads to an overall improvement in accuracies. Motivate your answer.

Part B: Comparison of algorithms - multiple datasets [30 = 10 marks x 3]

Consider the following three benchmarking datasets, together with the Seismic Bumps data.

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

1. Apply the four algorithms you used in homework 1 to the three new data sets, using tenfold cross validation.
2. Create a table showing the accuracies of the four algorithms against the four data sets.
3. Determine whether there is a statistically significant difference in the accuracies obtained in question 2. If the difference is statistically significant, you should calculate the critical difference (CD) and draw the Nemenyi diagram.