ComS 474/574 Introduction to Machine Learning
Lab 2 (100 points)
Decision Tree Learning

Note

- For help with laboratory assignments, please contact TA.
- You may perform the experiments using a machine learning package, such as Weka (https://www.cs.waikato.ac.nz/ml/weka/).

In this lab assignment, you will experiment with the Decision Tree classifier.

1 Dataset

We will use the following two data sets from the UC Irvine Machine Learning Repository:

- Congressional Voting Records Data Set
  (https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records).
  The dataset has missing values.
- Breast Cancer Wisconsin (Original) Data Set
  (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)).
  The dataset has numeric attributes and missing values.

2 Tasks

1. Learn Decision Tree classifiers on the two data sets (for example, using J48 in Weka). Visualize the tree constructed by the decision tree algorithm. Are there some interesting rules that make sense based on what you understand about the data?
2. Perform the following 5-fold cross-validation experiments with the Congressional Voting Records Data Set to study the stability of decision tree learning algorithm over the variability of data samples.
   (a) Randomly split the dataset into 5 data sets of (roughly) equal size $D_1, D_2, \ldots, D_5$.
   (b) For $i = 1, 2, \ldots, 5$, each time use D_i as test data and the rest as training data to learn a decision tree and measure its accuracy p_i.
   (c) Visualize the five trees constructed. Do the five trees differ with each other and with the tree constructed using all the data?

(d) Compute the average of $p_1, \ldots, p_5$ as the 5-fold cross-validation estimation of the accuracy of the Decision Tree classifier. Report 95% confidence interval.

3. Report the accuracy of the Decision Tree classifier on the Breast Cancer Wisconsin (Original) Data Set using 5-fold cross validation. Report 95% confidence interval.

4. Both datasets have missing values. C4.5 applies method of fractional instances during training and testing. A different method for dealing with missing values is to preprocess the data and fill them in with the most common values or average values (Weka has a missing values filter to do this). Now preprocess data by filling in missing values, and then compute the accuracy of the Decision Tree classifier on the two data sets using 5-fold cross validation. How does filling missing values affect the performance of the classifiers?

3 What to turn in

Turn in via Canvas the following:

- A lab report (in pdf file) with your experimental results.
- Readme file with instructions on how to reproduce your experiments. You should specify the parameters of every experiment in a way such that they can be replicated by the TA.
- Any source code that you may have written (in a zip file).