

# CSCI 6506 Assignment 2: GP sandbox

Malcolm Heywood

Due date: Oct 13

## 1 Goals

The goal of this assignment is to have you get your fingers dirty with a model of Genetic Programming under the context of supervised learning and the classification task. This is not expected to represent state-of-the art, or a particularly good example of how GP may ‘outshine’ other classification methods; SVM approaches are pretty difficult to better in any ‘bake off’.

What I want to see is you gain some intuition regarding what the various tradeoffs are for parameter selection in GP.

I strongly recommend that you start earlier than latter! This code is not optimized for speed (there are various course project opportunities if you are interested in tinkering with the SBB code:)

## 2 The ask...

1. Download v1 of SBB from:
  - <http://web.cs.dal.ca/~mheywood/Code/SBB/>
2. Assume the initial parameterization from the 2008 paper included in the distribution and familiarize yourself with the ‘readme’ file.
3. The distribution comes with two data sets ‘ANN’ and ‘Iris’. I suggest that you concentrate on Iris to minimize the cost of performing runs!
4. With respect to the training partition, re-standardize your data (ANN or Iris) such that the attribute data falls in the interval:

- (a)  $[-1.0, 1.0]$
- (b)  $[-10.0, 10.0]$

The end result should be two data sets with training/ test partitions over these intervals. Needless to say, you will need to design a suitable mapping based on the min/ max of the training attribute instances. Document this! The same parameterization for this mapping will need applying to the test data.

5. Make 10 runs of SBB under the same parameterization, but different seed number, for each of the standardizations. Is there a preference with regards to the resulting classifier performance?
6. What impact does population size and generation count have on the model outcome? Suggest that you divide and multiply the current value by ten, but also make sure that you keep the total evaluation count a constant!
7. Does changing the probabilities associated with the variation operators have any impact on the solution quality? Can you establish any rules of thumb?
8. How important is the selection of total number of registers, maximum number of instructions and maximum team size? [suggest that you experiment with smaller register and instruction limits, but consider larger team sizes]
9. What properties from your result files might suggest when team size or maximum instruction limits represent significant limitations?
10. Take your final recommended parameterization and apply it to the second data set included in the SBB distribution. Are the solutions still effective or is your parameterization only 'good' for the data set you made the initial experiments on?

### 3 Reporting

- Provide a written account of your experiences for questions 5 through 10.
- You are free to make use of graphing / statistical tests of your choice.
- Are there any particular strengths / weaknesses of the approach you can identify?
- The training overhead of GP is often cited as a deterrent to its wider utility. Are there any practices you foresee as being appropriate for speeding SBB up without compromising the classification accuracy?