CS 550 -- Machine Learning
Homework #3
Due: 10:40 (class time), December 18, 2018

Design and implement a genetic algorithm based approach for cost sensitive learning, in which the misclassification cost is considered together with the cost of feature extraction. In this question, you can design different algorithms. Although you are not required to follow the steps below, they may help you design your algorithm.

- Select a classification algorithm to start with. It is suggested to select a simple one.
- Use a bit string representation to indicate what features are selected (e.g., the bit string 100101 may indicate that the 1st, 4th, and 6th features are selected and the remaining ones are not).
- Write an appropriate fitness function that guides your algorithm. You may want to write a function that includes both the misclassification cost and the cost of extracting the selected features.

You will conduct your experiments on the "Thyroid data set", which is taken from the UCI repository and available on the course web page. The details of this data set are given as follows:

- This data set contains separate training ("ann-train.data") and test ("ann-test.data") sets.
- The training set contains 3772 instances and the test set contains 3428 instances.
- There are a total of 3 classes.
- In the data files, each line corresponds to an instance that has 21 features (15 binary and 6 continuous features) and 1 class label.
- The 21st feature is defined using the 19th and 20th features. This means that you do not need to pay for this feature if the 19th and 20th features have already been extracted. Otherwise, you have to pay for the cost of the unextracted feature(s).
- The cost of using each feature is given in another file ("ann-thyroid.cost"). It does not include the cost of the 21st feature because it is a combination of the other features.

In this assignment, you may use any programming language that you would like. Your report should include the following.

- The detailed explanation of the algorithm you will design. You should include the pseudocode and/or the flowchart of your algorithm.
- The classification algorithm, representation, and fitness function you will use.
- The parameters of the algorithm and their selected values.
- The features selected by your genetic algorithm for the Thyroid data set. Also report the total cost of these selected features.
- The training and test set accuracies obtained on the Thyroid data set.

Please note that due to the size of the data set and the number of the classifiers you will use (since you need to use a different classifier for each subset of features), the runs of this assignment can take a considerable amount of time. Please do not leave this assignment to the last minute; make sure to give yourselves enough time to finish it before the deadline.

Submit the hardcopy of your report but do not submit the printout of your source code. Your report should be a maximum of 4 pages. Email the source code of your implementation; the subject line of your email should CS 550: HW3.