CS 550 -- Machine Learning
Homework #3
Due: 17:30, January 3, 2020

This homework asks you to implement two algorithms for cost sensitive learning, in which the misclassification cost is considered together with the cost of feature extraction. You will test both these algorithms 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 is a total of three classes.
- In the data files, each line corresponds to an instance that has 21 features (15 binary and 6 continuous features) and one 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 19th and 20th features.

The two algorithms that you will implement use (1) forward selection and (2) genetic algorithm, respectively. With respect to these algorithms, you are asked to select the "best" subset of features on the training instances.

PART 1: First make your selections that you use in your implementations. In particular;

- Select a classification algorithm that you use at all steps of forward selection as well as of genetic algorithm. For example, if you choose to use a decision tree classifier, you need to use it at all steps of both of the algorithms. Of course, features on which you construct a decision tree change from one step to another. Here you can use any built-in classifier from any library (i.e., you do not need to code your own classifier). However, since you need to train this classifier for many different feature subsets, I recommend you to select a simple classifier for which training is fast.
- Define a joint cost function that combines the misclassification cost with the cost of feature extraction (for example, you may use a linear combination of these two types of costs). You need to use this function to select features one by one in the forward selection algorithm and as the fitness function that guides the genetic algorithm. You need to use exactly the same function definition for both of the algorithms. That is, for example, if you use a linear function to combine these costs, you have to use the same coefficients for both of the algorithms.
- Define a stopping condition. Use a similar (if possible, the same) stopping condition for both of the algorithms.
- *In your report;*
  a) give the name of the classifier and list its parameters, if any,
  b) give the definition of the joint cost function, and
  c) explain your stopping condition(s).

PART 2: Design and implement a forward selection algorithm that determines the "best" feature subset. In your report;

- Give the implementation details that are not covered in Part 1, if any.
- Indicate what feature is selected at each step. Report the training set accuracy obtained after this feature selection. Here report the overall accuracy as well as the class-based accuracies (which means you need to report four accuracies at each step). Also report the value of the joint cost function for this feature selection.
- Then, report the training and test set accuracies obtained when all selected features are used. Likewise, report the overall accuracy as well as the class-based accuracies for the training and test sets, separately.
- Report the total cost of the selected features (just for a single instance).

PART 3: Design and implement a genetic algorithm that determines the "best" feature subset. In your report;

- Explain how you represent a hypothesis. As also mentioned in class, for example, you may use a bit string representation to indicate what features are selected (e.g., the bit string 100101 may indicate that the 1^{st}, 4^{th}, and 6^{th} features are selected and the remaining ones are not).
- List the parameters of the genetic algorithm and give their selected values.
- Give the implementation details that are not covered in Part 1 as well as up to that point, if any.
- Indicate the features selected by your genetic algorithm at the end.
- Then, report the training and test set accuracies obtained when all selected features are used. Likewise, report the overall accuracy as well as the class-based accuracies for the training and test sets, separately.
- Report the total cost of the selected features (just for a single instance).

PART 4: Compare your findings in Part 2 and Part 3 very briefly (at most in five sentences).

In this assignment, you may use any programming language you want. 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 must be CS 550: HW3.