# Introduction to Evolutionary Computation (22753) Course Work II (10%): Symbolic Regression with Genetic Programming

## Due: December 11, 2009: 12pm

We have provided a file containing comma-separated values describing the energy consumption of a building. Each line in the file contains four values (comma separated): the measured energy consumption of a building $c$, the time of day ($T$), the external temperature ($t$) and the solar radiation ($s$).

For this assignment, you need to write a genetic programming based symbolic regression, which can derive a symbolic expression $E$ that describes an approximation for $c$ as a function of $T$, $t$ and $s$: $c' = f(T, t, s)$. You should use Root Mean Square (RMS) between the predicted energy comsumption $c'$ and the measured energy comsumption $c$ as error function for $N$ samples:

$$e = \sqrt{\frac{\sum_{i=0}^{N} (c_i - E(T_i, t_i, s_i))^2}{N}} \tag{1}$$

You are also required to document your work and the submitted report should include at least the following:

1. . A description and justification of the design of the EA, including comments on your choice of
   - function set
   - terminal set
   - the method and parameters used to generate the initial population.
2. A copy of your working program in Java/C/C++, matlab. You should **not** use any of the exiting GA/GP libraries. If you want to use any libraries, please ask first.
3. The parameter values used for the EA in your experiments (e.g., crossover rate). Justify your choices.
4. The results from your program. You should carry out 10 independent runs (trials) of your algorithm with identical settings. Each trial should have an upper limit of at least 100,000 fitness evaluations. The results should be reported in both a table and 2 figures: The table shows the 10 independent results, the mean and the standard deviation. One figure shows the evolutionary process of all trials, the other shows the mean and standard deviation of all trials.
5. For the best result you can find (either one of the 10 runs above, or from a seperate run): the fitness achieved, a print-out of the genome (the symbolic expression), a csv file with the actual and predicted energy consumption $c'_i$ for all samples, and a plot of the predicted and actual energy consumption for the first 300 samples.

6. A signed declaration that the work is entirely written and done by yourself, unless it is noted otherwise. Submission to be done online, as per instructions shown on the course web site. No late submissions.

Hint: an RMS of 30 is relatively easy to achieve. Can you get an RMS less than 15?