Derivation of Least-Squares Estimate of Coefficients in Crossover and Mutation for GSGP

April 2, 2015

The crossover and mutation operators used in this package offer a number of ways in which the various required coefficients can be established. In the case of both crossover and mutation, least squares estimates of the coefficients can be performed to arrive at optimal (with respect to the training data) coefficients for the current operation. This document outlines the derivation of these methods, as used in the framework (see `gsgp.c`).

Throughout all of this document, a squared error function between a model F and the response variable y is assumed:

$$E = \sum_{i=1}^{N} (y_i - F_i)^2 , \tag{1}$$

where N is the number of instances in the data set being used to estimate the coefficients, and y_i and F_i denote (respectively) the response value and model prediction for an instance i.

1 Derivation of Mutation Coefficient: ms

Assuming that a mutant individual M is created by:

$$M = B + msR \tag{2}$$

where B is the individual from which the mutation is based, and R is a newly created random model (e.g., a parse tree). The error function is therefore:

$$E = \sum_{i=1}^{N} (y_i - B_i - msR_i)^2 \tag{3}$$

where ms is the mutation scale coefficient. An estimate of ms can be found when:

$$\frac{dE}{dms} = -2 \sum_{i=1}^{N} [R_i (y_i - B_i - msR_I)] = 0, \tag{4}$$

which, when rearranged, becomes:

$$ms = \frac{\sum_{i=1}^{N} R_i y_y - \sum_{i=1}^{N} R_i B_i}{\sum_{i=1}^{N} R_i^2} \tag{5}$$

2 Derivation of Crossover Coefficient: p

Assuming that an offspring individual O is created by:

$$O = pM + (1-p)\,F \tag{6}$$

where M and F are the parents (mother and father, respectively) from which the crossover is based. Rearranged, this becomes:

$$O = F + p\,(M - F)\,, \tag{7}$$

where p is the crossover coefficient. This has the same overall structure as the equation for mutation, albeit working on different base solutions. The error function is therefore:

$$E = \sum_{i=1}^{N} \left(y_i - F_i - p\,(M_i - F_i)\right)^2, \tag{8}$$

and we can reuse the approach in the previous section for estimating ms. Substituting the equivalent parts into 5, we get:

$$p = \frac{\sum_{i=1}^{N}(M_i - F_i)y_y - \sum_{i=1}^{N}(M_i - F_i)F_i}{\sum_{i=1}^{N}(M_i - F_i)^2} \tag{9}$$

3 Derivation of Separate Crossover Coefficients for Each Parent: p and q

Assuming that an offspring individual O is created by:

$$O = pM + qF \tag{10}$$

where M and F are the parents (mother and father, respectively) from which the crossover is based, and p and q are the coefficients attached to each parent. The error function becomes:

$$E = \sum_{i=1}^{N} \left(y_i - pM_i - qF_i\right)^2. \tag{11}$$

An estimate of p can be found when:

$$\frac{\partial E}{\partial p} = -2\sum_{i}^{N} \left[M_i\left(y_i - pM_i - qF_i\right)\right] = 0 \tag{12}$$

and, likewise, an estimate of q when:

$$\frac{\partial E}{\partial q} = -2\sum_{i}^{N}\left[F_i\left(y_i - pM_i - qF_i\right)\right] = 0. \tag{13}$$

Rearranging 12 and 13, we get:

$$p = \frac{\sum_i^N M_i y_i - q\sum_i^N M_i F_i}{\sum_i^N M_i^2} \tag{14}$$

and:

$$q = \frac{\sum_i^N F_i y_i - p\sum_i^N M_i F_i}{\sum_i^N F_i^2}. \tag{15}$$

Finally, substituting 14 in place of p, and then rearranging 15, we get:

$$q = \frac{\sum_i^N F_i y_i \times \sum_i^N M_i^2 - \sum_i^N M_i y_i \times \sum_i^N M_i F_i}{\sum_i^N M_i^2 \times \sum_i^N F_i^2 - \left(\sum_i^N M_i F_i\right)^2}. \tag{16}$$