Symbolic Regression Competition

Objective

The objective is to generate a robust symbolic regression model that relates an expensive but noisy lab data of a chemical composition (output) to 57 cheap process measurements, such as temperatures, pressures, and flows (inputs). The selected equation has to include the most sensitive inputs relative to the output, i.e. some form of variables selection is recommended. If accepted by process engineers, the proposed symbolic regression solution could be implemented in a chemical process monitoring system.

Data Set Description

The data set includes 57 measurements of process variables, which are potentially related to the composition. However, not all of them are highly correlated to the output. The data file includes two worksheets with training and test data. The input columns are named x1 to x57 and the output column is named as y. The training worksheet includes 747 training data points and the test data sheet includes 319 test data points.

Rules

Entrants must submit a report with a description of the model development process, including: GP parameters used, variable selection method used, symbolic regression expression build in the Excel spreadsheet with the data, and key statistical performance metrics, such as R2 and RMSE on training and test data. Submissions will be reviewed by a committee of industrial experts. Each model will be scored by statistical performance on test data, its simplicity, and interpretability.

Important dates

Submission deadline: March 8, 2010
Conference: April 7 – 9, 2010