Linear regression of the outputs of a device and optimizing the linear model using linear programming

Dristy Parveg
Mohsin Khan
Finland

March 20, 2017

It has been observed that the device has two outputs $y_1, y_2 \in \mathbb{R}$ which are functions of 5 variables $x_1, x_2, x_3, x_4, x_5 \in \mathbb{R}$. We are interested to know the values of the variables x_1, x_2, x_3, x_4, x_5 that produce the outputs y_1, y_2 which are close to optimal values.

Let $f_i : \mathbb{R} \to \mathbb{R}$ be the function that takes the variable x_i as input while the other input variables x_j for $j \in [5] \setminus i$ are kept at constant values and $f_i(x_i) = y_1$. We define the function $g_i : \mathbb{R} \to \mathbb{R}$ in the same way so that $g_i(x_i) = y_2$. From the result of the experiments, presented in previous section, it is evident that f_i and g_i are almost, if not exactly, linear functions.

We are interested to find models $f : \mathbb{R}^5 \to \mathbb{R}$, $g : \mathbb{R}^5 \to \mathbb{R}$ where $f(x_1, x_2, x_3, x_4, x_5) = y_1$ and $g(x_1, x_2, x_3, x_4, x_5) = y_2$. We make an educated guess that both f and g are linear. We do not provide an analytical argument to justify the linearity of f and g. Instead, we use linear regression on the experimental results. There are 20 experiments. We randomly choose 15 experiments and use them as training data to obtain the linear models

$$\begin{aligned} f &= a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 \\ g &= b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 \end{aligned}$$

If $x_1, x_2, x_3, x_4, x_5, y_1, y_2$ be an experimental result then $\frac{|f(x_1,x_2,x_3,x_4,x_5)-y_1|}{y_1} \times 100$ and $\frac{|f(x_1,x_2,x_3,x_4,x_5)-y_2|}{y_2} \times 100$ are the error percentages of the obtained models f and g. We compute the training error of the models using the 15 experimental results and find that the average training error is around ?? percent. We use the rest of the 5 experimental results to compute the test error of the obtained linear models. We find the error percentage to be below ?? percent in both of the cases of f and g.

We reckon, this error percentages indicate that the underlying models are linear enough and the obtained approximated linear models can be used to find the approximately optimal values of y_1 and y_2 and also the corresponding inputs $x_1, ..., x_5$ that produces outputs y_1, y_2. As the underlying models are not not exactly linear and our obtained models are only a linear approximation, the optimal values obtained from the model will have some errors. We expect these errors to be similar to the test errors of the models. In this paper we look forward to find the inputs for which y_1 is maximized and y_2 is minimized. In other words we look forward to maximize the expression

$$\begin{aligned}
y_1 - y_2 &= f(x_1, ..., x_5) - g(x_1, ..., x_5) \\
&= (a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5) - \\
&\quad (b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5) \\
&= (a_1 - b_1)x_1 + (a_2 - b_2)x_2 + (a_3 - b_3)x_3 + (a_4 - b_4)x_4 + (a_5 - b_5)x_5 \\
&= \sum_{i \in [5]} (a_i - b_i)x_i
\end{aligned}$$

We model this maximization problem as a linear programming problem as follows:

$$\begin{aligned}
\text{maximize} &: \sum_{i \in [5]} (a_i - b_i)x_i \\
\text{with constraints} &: 0 < x_i \leq c_i \text{ for } i \in [5]
\end{aligned}$$

This is a fairly small instance of a linear programming problem. There exist efficient algorithms, namely, simplex, ellipsoid, etc. for solving linear programming problems. In practice, being oblivious to the algorithm used, this linear programming problem can easily be solved by any standard linear programming solver like lp_solve. Let $x_1', x_2', x_3', x_4', x_5'$ be the solution of the above linear programming problem.

We run an experiment using these input values and get output $y_1^{'}, y_2^{'}$. It can be expected that $\frac{|f(x_1,\ldots,x_5)-y_1^{'}|}{y_1^{'}} \times 100$, $\frac{|f(x_1,\ldots,x_5)-y_2^{'}|}{y_2^{'}} \times 100$ are approximately equal to the test error of f and g respectively.