Optimal Variable Subset Selection Problem in Regression Analysis is NP-Complete

Paolo Victor T. Redondo  
School of Statistics  
University of the Philippines Diliman

Combinatorial and optimization problems are classified into different complexity classes, e.g., whether an algorithm that efficiently solve the problem exists or a hypothesized solution to the problem can be quickly verified. The optimal selection of subset variables in regression analysis is shown to belong to a complexity class called NP-hard (Welch, 1982) in which solutions to the problems in the same class may not be easily (in terms of computing speed) proven optimal. Variable selection in regression analysis based on correlations is shown to be NP-hard, i.e., a complexity class of problems with easily verifiable solutions.

Keywords: optimal variable selection, regression analysis, np-completeness

1. Introduction

Despite the technological advancements in computing power, most existing combinatorial and optimization problems have not yet been found an efficient solution. This gives rise to one of the seven Millennium Problems selected by the Clay Mathematics Institute in 2000; the P versus NP Problem (de Figueiredo, 2012). Suppose a decision problem, i.e., a question answerable by “yes” or “no”, is easily solvable. Such problem belongs to the complexity class P (polynomial time) where the time it takes to find the solution is simply a polynomial function of the size of the inputs. Consider the basic multiplication of two -digit numbers. Regardless of how large n is, the algorithm to efficiently solve the problem only requires \( n^2 \) single-digit multiplications (a polynomial in ). This is considerably “quick” compared to some exponential functions, say \( 3^n \). Other examples are string matching, number sorting and finding the maximum and minimum value in an array. On the other hand, the complexity class NP (non-deterministic polynomial time) includes all problems that are solvable in nondeterministic polynomial time,