On classical and modern variable selection in regression

Daniel Nevo

Assistant Professor
Department of Statistics and Operations Research
School of Mathematical Sciences
Tel Aviv University

Abstract:

A central challenge in both classical and modern data analysis is how to choose which variables should be included in a regression model. The 20-th century has taught us to use F-test, AIC, BIC and other methods. Towards the end of that century, regularization-based methods were developed with increased intensity in the past 15 years, due to the emergence of big data. The decision on which variables should be included in the model ultimately depends on the goals of the analysis. In particular, even if one is willing to assume a “true model” exists, the task of finding it may be too hard without strong and unverifiable assumptions, especially in the high-dimensional setting. In this talk, I will review the goals of model/variable selection as I see them. I will then present one possible approach, which is replacing the goal of finding a single set of variables with finding a minimal class of models. The minimal class of models includes candidate models, each serves as a potential good model for prediction of the outcome. If time permits, I will demonstrate how the minimal class of models can be utilized for practical data analysis.