The MDL model choice for linear regression

Erkki P. Liski

University of Tampere, Tampere, Finland

Abstract

In this talk, we discuss the principle of *Minimum Description Length (MDL)* for problems of statistical modeling. By viewing models as a means of providing statistical descriptions of observed data, the comparison between competing models is based on *the stochastic complexity (SC)* of each description. *The Normalized Maximum Likelihood (NML)* form of the SC (Rissanen 1996) contains a component that may be interpreted as the parametric complexity of the model class. Once the SC for the data, relative to a class of suggested models, is calculated, it serves as a criterion for selecting the optimal model with the smallest SC. This is the MDL principle (Rissanen 1978, 1983) for model choice.

If the parametric complexity of a model family is unbounded, then one must deviate from the clean definition of the SC. The most important example of this phenomenon is the Gaussian family. One approach to bound the parametric complexity is by constraining the sample space. We calculate the SC for the Gaussian linear regression by using the NML density and consider it as a criterion for model selection. The final form of the selection criterion depends on the method for bounding the parametric complexity. As opposed to traditional fixed penalty criteria, this technique yields adaptive criteria that have demonstrated success in certain applications.

Keywords

Minimum description length, Stochastic complexity, Normalized maximum likelihood, Parametric complexity, Adaptive selection criteria.

References:

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica 14*, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics 11*, 416–431.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42(1)*, 40–47.