Generalization Theory of Two-part Code MDL Estimator

Tong Zhang
Department of Statistics
Rutgers University

I will present a finite-sample generalization analysis of two-part code MDL estimator. This method selects a model that minimizes the sum of the model description length plus the data description length given the model. It can be shown that under various conditions, optimal rate of convergence can be achieved through an extended family of two-part code MDL that over-penalize the model description length.

As an example, we apply MDL to learning sparse linear representations when the system dimension is much larger than the number of training examples. This is a problem that has attracted considerable attention in recent years. The generalization performance of a two-part code MDL estimator is calculated based on our theory, and it compares favorably to other methods such as 1-norm regularization.