Optimal model selection

Sylvain Arlot* (CNRS, LIENS, Willow Team)

We consider the model selection problem when the goal is prediction. A model selection procedure is said to be optimal when it satisfies an oracle inequality with constant tending to 1 when the sample size tends to infinity.
Classical dimensionality-based penalization procedures (such as Mallows' C_p) often are suboptimal in the least-squares regression framework, when the noise-level is not constant (Arlot, 2008c). On the contrary, cross-validation methods automatically "learn" variations of the noise-level, but V-fold cross-validation (VFCV) is also suboptimal in least-squares regression when V is fixed (Arlot, 2008a), and VFCV can be computationally untractable for large V.
We show how to use resampling or cross-validation ideas for learning the penalty, with no prior information on the noise-level (Arlot, 2008ab). In particular, V-fold penalization leads to optimal model selection among regressograms—even when V is fixed—, with the computational cost of VFCV (Arlot, 2008ab); therefore, it strictly improves VFCV. This optimality result will be illustrated by simulation experiments, showing in particular that V-fold penalties can be successful in several other frameworks than regressogram selection.

References

[1] Arlot (2008a) V-fold cross-validation improved: V-fold penalization, `arXiv:0802.0566v2`
[2] Arlot (2008b) Model selection by resampling penalization, `hal-00262478_v1`
[3] Arlot (2008c) Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression, `arXiv:0812.3141v1`