Minimum Description Length Model Selection

de Rooij, S.

Publication date
2008

Abstract

Model selection is a strange and wonderful topic in learning theory and statistics. At first glance the question seems very clear-cut: how should we decide which set of probability distributions matches the observations at hand best. This question comes up time and again in many different contexts, ranging from testing scientific hypotheses in general (which among these psychological models describes best how people behave?) to more concrete applications (what order polynomial should we use to fit the data in this regression problem? What lossy representation of this image best captures the structural properties of the original?). Thus, model selection is ubiquitous, and the one-size-fits-all criteria based on the Minimum Description Length (MDL) principle and the closely related Bayesian statistics are appreciated by many.

Upon closer inspection, many applications of model selection are not as similar as they may first appear. They can be distinguished by technical properties (are the models nested? Parametric? Countable?), but also by a priori assumptions (is the process generating the data believed to be an element of any of the considered models?), as well as the motivation for performing model selection in the first place (do we want to identify which model contains the data generating process, or do we want to identify which model we may expect to predict future data best?). The best choice of methodology in any situation often depends on such particulars, and is further determined by practical considerations such as whether or not the relevant quantities can be evaluated analytically, and whether efficient algorithms exist for their calculation. MDL/Bayesian model selection has been shown to perform quite well in many different contexts and applications; in this thesis we treat some of the puzzling problems and limitations that have also become apparent over time. We also extend the idea by linking it to other topics in machine learning and statistical inference.

To apply MDL, universal codes or distributions have to be associated with each of the considered models. The preferred code is the Normalised Maximum Likelihood (NML) or Shtarkov code. However, this code yields infinite code word

lengths for many models. This first issue with MDL model selection is investigated in Chapter 2, in which we perform computer experiments to test the performance of some of the available alternatives. One result is that the model selection criterion based on the so-called prequential plug-in code displays inferior performance. This observation seems important because the prequential plug-in code is often thought of as a convenient alternative to other universal codes such as the NML code, as it is much easier to calculate. It was thought to result in code lengths similar to those obtained for other universal codes (such as NML, 2-part codes or Bayesian mixtures), but we discovered that this is only the case if the data generating process is in the model. We show in Chapter 3 that the redundancy of the prequential plug-in code is fundamentally different from the standard set by other universal codes if the data generating process is not an element of the model, so that caution should be exercised when it is applied to model selection.

The third problem treated in this thesis is that MDL/Bayesian model selection normally does not take into account that, even in the ideal case where one of the considered models is "true" (contains the data generating process), and even if the data generating process is stationary ergodic, then still the index of the model whose associated universal code issues the best predictions of future data often changes with the sample size. Roughly put, at small sample sizes simple models often issue better predictions of future data than the more complex "true" model, i.e. the smallest model that contains the data generating distribution. When from a certain sample size onward the true model predicts best, the simpler model has already built up a lot of evidence in its favour, and a lot of additional data have to be gathered before the true model "catches up" and is finally identified by Bayesian/MDL model selection. This phenomenon is described in Chapter 5, in which we also introduce a novel model selection procedure that selects the true model almost as soon as enough data have been gathered for it to be able to issue the best predictions. The criterion is consistent: under mild conditions, the true model is selected with probability one for sufficiently large sample sizes. We also show that a prediction strategy based on this model selection criterion achieves an optimal rate of convergence: its cumulative KL-risk is as low as that of any other model selection criterion. The method is based on the *switch distribution*, which can be evaluated using an efficient expert tracking algorithm. More properties of this switch distribution are treated in Chapter 4, which also contains a survey of this and other expert tracking algorithms and shows how such algorithms can be formulated in terms of Hidden Markov Models.

Finally, in Chapter 6 we evaluate the new theory of algorithmic rate-distortion experimentally. This theory was recently proposed by Vitányi and Vereshchagin as an alternative to classical rate-distortion theory. It allows analysis of the structural properties of individual objects and does not require the specification of a probability distribution on source objects. Instead it is defined in terms of Kolmogorov complexity, which is uncomputable. To be able to test this theory in practice we have approximated the Kolmogorov complexity by the compressed

size of a general purpose data compression algorithm. This practical framework is in fact a generalisation of MDL model selection.

The perspectives offered in this thesis on many aspects of MDL/Bayesian model selection, contribute to a better understanding of the relationships between model selection and such diverse topics as universal learning, prediction with expert advice, rate distortion theory and Kolmogorov complexity.