Steven Weijs

Bayes in terms of description lengths and surprises.

When testing models against observations, the log-likelihood is a measure of model quality, which is only defined if the model output and/or the observation are defined in terms of probability, i.e. incomplete knowledge. Information theoretically, the log likelihood is proportional to how surprising the data is once we know the model, but also to how much information (bits) we need to describe (store) the data in compressed form, once the model is known. The model, however is often not known, but inferred from the same data that we try to describe. The description of the model is needed to reproduce the data and also adds to the total description length. This also includes the model parameters that take up space proportional to the required precision (e.g. digits or bits). The description length of models and parameters relates to how surprising they are, i.e. to the prior probability of the model. In absence of prior knowledge, this prior probability depends only on the description length of the model (model complexity). These analogies are formalized in (algorithmic) information theory, where the relation between description length and probability, L=-log(P), appears in many different forms. The logarithm reflects that description lengths add where probabilities multiply. The probability of a model given the data is proportional to the likelihood times the prior probability of the model, while the total description length of data is the sum of the model description plus the storage of the residuals (remaining uncertainty / missing information). A perfect simple model is the shortest possible description of data. Since description lengths can be related to probabilities, a Bayes mixture over all possible descriptions of data yields the idealized prediction, including the full model structural uncertainty. Concluding, we can state that hydrological model inference is equivalent to compression of hydrological data. Hydrological models can be used to compress data and efficient data compression algorithms will resemble hydrological models. Dialog between computer scientists and hydrologists can promote more efficient storage of hydrological data and more effective ways to distill hydrological models from increasingly large volumes of data. Some initial explorations of this idea are presented using the MOPEX data set.