Proceedings of the Conference CompStat 2002
Short Communications and Posters[1]

Choosing the order of a hidden Markov chain through cross-validated likelihood

Gilles Celeux[1] and Jean-Baptiste Durand[1]

[1] INRIA Rhône-Alpes, Montbonnot, France.
email: {Gilles.Celeux; Jean-Baptiste.Durand}@inria.fr

Keywords. Model selection, multifold cross validation, hidden Markov models

1 An information-theoretic approach

We explore the possibility of selecting the number of hidden states in a hidden Markov chain model (HMC) using cross-validated likelihood. This way of selecting a model has been considered by Smyth (2000) for independent mixture models and provided promising results. This principle is based on the estimation of the Kullback-Leibler divergence between a conceptual "true density" and the different models in competition (see Burnham and Anderson, 1998). The fact that a large amount of data will support more complex models is taken into account, thus leading to parsimonious models when only small data sets are available. Furthermore, this method obviates the theoretical difficulties in the derivation of the BIC criterion (Schwarz 1978) and the derivation of statistic test distributions. Finally the "true model" is not required to be in the set of competing models.

The so-called multifold cross validation for selecting a model from a sample of size n (see for instance Zhang 1992) consists of

1. partitioning at random this sample in M subsamples of size d,
2. deriving the maximum likelihood estimates from the learning sample without one of the subsample of size d,
3. computing the loglikelihood of the estimated parameters on the subsample which does not take part to the learning process,
4. repeating 2. and 3. M times, the resulting multifold cross-validated loglikelihood is the mean of the M loglikelihoods computed in 3.

2 Dealing with missing values in HMC

In the HMC context, there is a theoretical difficulty to use cross validation for model selection, since the Markov dependence between the hidden states is lost when data are removed at random from the original sample set. We present

a first way to deal with this difficulty. It consists of using half sampling in a particular fashion. (Half sampling is multifold cross validation with $M = 2$ or equivalently $d = n/2$.)

If the two subsamples are designed by considering respectively the odd and the even indices of the original HMC sample, the resulting processes are still hidden Markov chains with the same parameters except that the transition matrix of the chain is the square of the original transition matrix. From this theorem, it is easy to compute the cross-validated likelihood with this half sampling scheme.

In the more general situation of multifold cross validation where d data points are deleted at random from the learning sample to compute the cross-validated likelihood, we have established the formulas to adapt the equations of the forward-backward recursion (Baum et al., 1970), under the assumption that the deleted data are missing at random. We use the EM algorithm to estimate the parameters in the context of two types of missing data: the conceptual hidden states of the HMC model and the observations deleted in the subsampling phase. The new algorithm leads to a forward-backward recursion involving, basically, the powers of the transition matrix.

3 Validation on simulated and real data

Equipped with those new tools, we have performed Monte Carlo numerical experiments to assess the performance of different variants of multifold cross validation to select the order of an HMC model and to compare their performance with the BIC criterion (Schwarz 1978). Those experiments show that half sampling with the "even" and "odd" hidden Markov chains gives satisfactory results and is comparable to the BIC criterion. Our method has been used on a real data set related to software reliability. As an extension of this work, we show how the multifold cross validation of the likelihood can be used with more complex hidden structure models, as hidden Makov trees for instance.

References

Baum, L. E., Petri, G., Soules, G. and Weiss, N. (1970). A maximization technique in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**, 164-171.

Burnham, K. P. and Anderson, D. R. (1998). *Model Selection and Inference*, New-York: Springer-Verlag.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461-464.

Smyth, P. (2000). Model Selection for Probabilistic Clustering using Cross-Validated likelihood. *Statistics and Computing*, **10**, 63-72.

Zhang, P. (1993). Model Selection via Multifold Cross Validation. *Annals of Statistics*, **23**, 299-313.