Using VLMC Model and Boosting to classify

Márcio Luis Lanfredi Viola
Federal University of São Carlos

Abbreviated abstract: The classification problem is useful in statistics. The usual structure of the classification problem involves a response variable and covariates. On the other hand, it is possible to observe only samples from a Markov chain. In this case, the observed values of each sample are dependent and we can model them using the Variable Length Markov Chain model (VLMC). The samples are used to estimate a context tree wich represents the dependency structure present in the data and it is used to classify. In this work, we show how to use a Markov chain in a classification problem.

Related publications:

– J. E. García, V. A. González-López and M. L. L. Viola, *Robust model selection and the statistical classification of languages*. In XI Brazilian Meeting on Bayesian Statistics: EBEB 2012, Vol. 1490, Nº. 1, pp. 160-170. AIP Publishing (2012).

– Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, Vol. 55, Nº. 1, pp. 119–139 (1997).

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022

Problem, Data, Previous Works

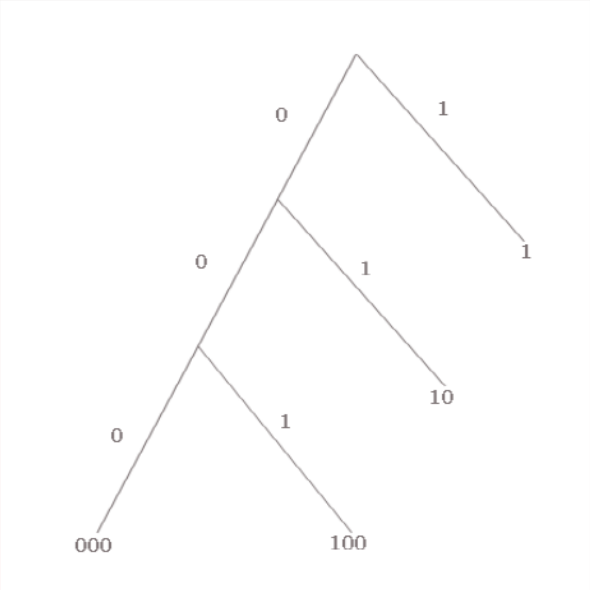
The usual structure of the classification problem involves independent samples of a response variable and covariates.

However, there are situations in which it is possible to observe only a large number of values of the response variable.

In this last case, the observed values of each sample are dependent and we can model them using Markov chain.

In particular, García, González-López and Viola (2012) used robust context tree to discriminate languages.

Methods



Example of context tree with alphabet $\mathcal{A} = \{0, 1\}$.

Pseudo-likelihood Function

$$L = \prod_{a,s} P(a|s)$$

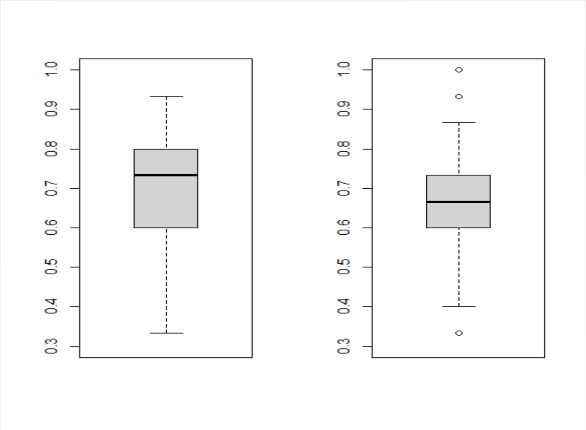
where a and s are elements of the alphabet and state space, respectively.

Maximum Likelihood Estimator

$$\hat{P}\,(a|s) = \frac{N_n(s,a)}{N_n(s)}$$

4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Results and Conclusions



We consider two Markovian Processes of order 1 with alphabet {0,1}.

- Model 1: $P(0|0) = 0.7$ and $P(1|1) = 0.4$;
- Model 2: $P(0|0) = 0.65$ and $P(1|1) = 0.45$.

For each model, we generated 60 samples of length 150. This procedure was done 50 times considering training and test samples.

We obtained a median value 73,3% and 66,6% of correct answers for models 1 and 2, respectively.

For the next step, we will investigate whether the AdaBoost algorithm improves performance of the classification.

4th Conference on
Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022