Algorithms for estimation of variable length Markov Chains and applications*

Author: David Henriques da Matta (dhmatta@mat.ufg.br) - IME/UFG

Extended Abstract

Introduction

There are many studies in the field of linguistics where the interest is to analyze the differences between Brazilian Portuguese and European Portuguese (henceforth BP and EP respectively). Both the BP on EP, have the same words set in their structure (lexicon). However, these languages have different syntaxes and different prosodies.

The key point of this process of differentiation, is related to the question of finding estimation methods. To better understand this theoretical context, we discuss here some basic concepts of variable length Markov chains, as well as a simulation study to find evidence whether to use BIC or AIC as the selection criteria of models to tune the pruning constant of the algorithm Context (Rissanen (1983); B¨uhlman and Wyner (1999)).

Probabilistic Model

First, we consider here an alphabet A as with any finite set.

Definition 1: *Let $(X_t)_{t\in\mathbb{Z}}$ an stochastic process taking values in a finite alphabet* A*. The process $(X_t)_{t\in\mathbb{Z}}$ is a Markov chain of order k, if there is an element $k \in \mathbb{N}\cup\{\infty\}$ such that for every $t \in \mathbb{Z}$ we have*

$$P(X_t = x_t|X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, X_{t-3} = x_{t-3}, ...) =$$
$$P(X_t = x_t|X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, ..., X_{t-k} = x_{t-k}) \quad (1)$$

for every sequence $x_t, x_{t-1}, ... \in$ A. The chain is also called stationary if, for all $n \in \mathbb{Z}$, we have $P(X_n = a) = \Pi(a)$, where Π is the stationary measure of the chain.

To introduce the idea of variable length memory for a Markov chain of order k, we will present the first the length function $\varrho : \mathrm{A}^k \to \{0, 1, ..., k\}$ that for each $x_{t-k}^{t-1} \in \mathrm{A}^k$

is defined as

$$\varrho(x_{t-k}^{t-1}) = \min\{\varrho \leq k | P(X_t = x_t | X_{t-k}^{t-1} = x_{t-k}^{t-1}) = \\ P(X_t = x_t | X_{t-\varrho}^{t-1} = x_{t-\varrho}^{t-1})\}. \quad (2)$$

The function $C : \mathrm{A}^k \to \cup_{m=1}^{k} \mathrm{A}^m$, given by

$$C : x_{t-k}^{t-1} \mapsto x_{t-\varrho(x_{t-k}^{t-1})}^{t-1}, \quad (3)$$

is called *context function* and the resulting vector $C(x_{t-k}^{t-1}) = (x_{t-k}, ..., x_{t-1})$ is called *context*.

Definition 2: *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary Markov chain, taking values in* A, $|\mathrm{A}| < \infty$ *and $C(.)$ be its corresponding context function defined by* (1). *Let $k \in \mathbb{N} \cup \{\infty\}$ the smallest value such that*

$$|C(x_{-\infty}^{t})| = \varrho(x_{-\infty}^{t}) \leq k \text{ to all } x_{-\infty}^{t} \in \mathrm{A}^{\infty}. \quad (4)$$

Then $(X_t)_{t\in\mathbb{Z}}$ is called variable length Markov Chain (VLMC) of order k.

The requirement of stationarity implies that the probability distribution P of a variable length Markov chain is completely specified by its the transition probabilities $P(X_0 = x_0 | C(x_{-\infty}^{-1}))$. Thus, one convenient way to represent these states, the minimal state space, is to use the tree (context trees).

Notice that the context function satisfies the property of the suffix, in other words, no context is a suffix of another context.

Algorithm Context

This algorithm was originally proposed by Rissanen (1983). Given a sample of variable length Markov chain, the goal is to estimate the context function as well as the corresponding transition probabilities.

The strategy of the algorithm is as follows. First, a maximal tree is produced. This construction considers all branches which appear at least a fixed number of times (a threshold to be determined) in the sample. Secondly, the algorithm uses a statistical procedure to proceed backwards pruning the tree.

If we interpret the pruning step of algorithm context, described in Bühlman and Wyner (1999), as a the likelihood ratio test, clearly we have that small cut points will result in largest context trees and occurrence of an overestimation.

Aiming to solve this problem, given two selection methods of models well known in the literature. The AIC and BIC criteria will be used with the objective to choose the model that best fits the data. The criteria can be described for the selection problem of context trees by the following function:

$$G(\gamma, D_n) = -2\text{log-likelihood}_{(D_n)} + \gamma(|\text{A}| - 1)|\tau_{\hat{c}_{D_n}}|, \tag{5}$$

where $\gamma = 2$ or $log(n)$ for AIC and BIC respectively, $|\tau_{\hat{c}_{D_n}}|$ is the number of free parameters (transition probabilities) from estimated tree, and finally log-likelihood$_{(D_n)}$ is the estimated value of the likelihood, where the cutoff is equal to D_n.

The purpose of these criteria to estimate the cutoff point is to minimize the divergence of Kullback-Leibler, therefore, the best model to be considered is that present lowest value in the AIC or BIC function as the choice of criteria to be used. We also, that using the BIC in the set of all possible trees, obtain a consistent estimator for the order of the chain (Csiszár, I. & Talata, Z. (2006)).

Simulations

The main objective of this work is to compare through simulation studies the behavior of the algorithm Context when applying the AIC and BIC selection criteria of models.

We will use these criteria in the following way. Change the range of the cuttoff parameter from 0.01 by 0.01, and stopping when it reaches an increase of twenty units. In the end, 2000 trees were estimated, and choose the cutoff value that provides the lowest value in the selection criterion (AIC or BIC).

First, we work with a renewal chain in the alphabet $\{0, 1\}$. The interesting characteristic of this process is that it loses memory each time it reaches the symbol 1. Therefore, this chain is composed of independent blocks of zeros delimited by 1 's.

With a renewal chain of the length $n = 3496$, we obtained the fit represented by Figures 1 and 2 when using the AIC e BIC criteria respectively. We note that

the symbol 1 branches when using the AIC criterion. Moreover, the depth of the estimated tree by the AIC criterion is greater in a unit. So, for the renewal chain, the results obtained with the BIC selection criterion were better, because we know the true structure of renewal chain, ie, symbol 1 does not branch.

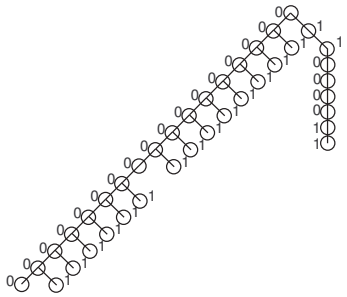

Figure 1:

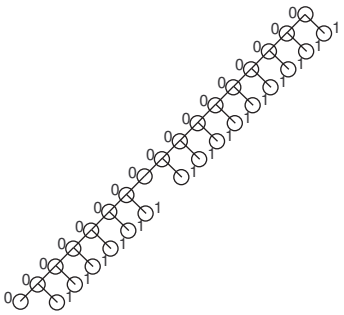

Figure 2:

The other data set which we use, refers to data with alphabet given by set A = $\{0, 1, 2, 3, 4\}$. The interest of this data set is such that not all transitions are possible, such as the symbol of 4 to 1. This data set is represented by Figure 3.

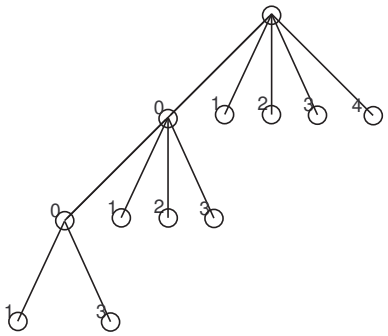

Figure 3:

From this data set, we fitted 240 trees with their respective transition probabilities, and using the function " simulate.vlmc", we simulated data sets for these 240 trees. The table below shows the fitting obtained by the algorithm Context using the AIC and BIC criteria in the simulated data. We use the following notation:

- Situation A := The estimated tree with the AIC criterion was identical to "true" tree;

- Situation B := The estimated tree with the BIC criterion was identical to "true" tree;
- Situation C := Both fits (AIC and BIC criteria) were identical to "true" tree;
- Situation D := Both fits were wrong, but estimated identical trees ;
- Situation E := Both fits were wrong, but the estimated tree of the AIC criterion was closer to "true" tree;
- Situation F := Both fits wrong, but the estimated tree of the BIC criterion was closer to "true" tree.

	Quantidade
Situation A	124
Situation B	92
Situation C	89
Situation D	68
Situation E	30
Situation F	15

Conclusions

With the renewal chain, it becomes clear that the context algorithm with BIC criterion estimates better trees. This feature can be justified by the fact that for these estimated trees the symbol 1 never branches, as it occurred with the estimation using the AIC criterion.

Moreover,in simulations performed in "simulated" data, in most cases the AIC criterion ultimately adjust better the estimated trees. This fact can be explained by following the principle that for small samples with large number of parameters to be estimated, the AIC criterion fits models with fewer parameters.

** Topic of master's degree dissertation. Advisor: Nancy Lopes Garcia - IMECC / UNICAMP.*

References

1. Bühlmann, P. & Wyner, A.J. (1999). Variable length Markov chains. *Annals of Statistics* **27**, 480-513.

2. Csiszár, I. & Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via bic and mdl, *Information Theory, IEEE Transactions on* **52**(3): 1007-1016.

3. Mächler, M., The VLMC package, 2005. Can be downloaded from http://cran.r-project.org/doc/packages/VLMC. Pdf.

4. Rissanen, J. (1983). A universal data compression system, *IEEE Trans. Inform. Theory*, **29**(5): 656-664.

5. The R Project for Statistical Computing, http://www.rproject.org.