Probabilistic Forests

An application to the rhythmic distinction between Brazilian and European Portuguese

Antonio Galves   Denis Lacerda   Florencia Leonardi

Instituto de Matemática e Estatística
Universidade de São Paulo

28th November, 2005

Remembering VLMC

A VLMC is a stochastic chain $(X_0, X_1, \ldots)$ taking values on a finite alphabet $\mathcal{A}$ and characterized by two elements:

- The set of all contexts.

  A context $X_{n-\ell}, \ldots, X_{n-1}$ is the finite portion of the past $X_0, \ldots, X_{n-1}$ which is relevant to predict the next symbol X_n.

- A family of transition probabilities associated to each context.

  Given a context, its associated transition probability gives the distribution of occurrence of the next symbol immediately after the context.

Remembering VLMC

A VLMC is a stochastic chain $(X_0, X_1, \ldots)$ taking values on a finite alphabet $\mathcal{A}$ and characterized by two elements:

- The set of all contexts.

  A context $X_{n-\ell}, \ldots, X_{n-1}$ is the finite portion of the past $X_0, \ldots, X_{n-1}$ which is relevant to predict the next symbol X_n.

- A family of transition probabilities associated to each context.

  Given a context, its associated transition probability gives the distribution of occurrence of the next symbol immediately after the context.

Remembering VLMC

A VLMC is a stochastic chain $(X_0, X_1, \ldots)$ taking values on a finite alphabet $\mathcal{A}$ and characterized by two elements:

- The set of all contexts.

  A context $X_{n-\ell}, \ldots, X_{n-1}$ is the finite portion of the past $X_0, \ldots, X_{n-1}$ which is relevant to predict the next symbol X_n.

- A family of transition probabilities associated to each context.

  Given a context, its associated transition probability gives the distribution of occurrence of the next symbol immediately after the context.

Remembering VLMC

- The set of all contexts is a *suffix code*. This means that no context is a suffix of another context.
- For this reason this set can be represented as a tree. This tree with the associated transition probabilities is called a *Probabilistic Tree*.

Remembering VLMC

- The set of all contexts is a *suffix code*. This means that no context is a suffix of another context.
- For this reason this set can be represented as a tree. This tree with the associated transition probabilities is called a *Probabilistic Tree*.

Variable Length Markov Chains (VLMC) have been used to discriminate rhythmic patterns in European and Brazilian Portuguese written texts (Galves et al. 2005).

- The written texts where transformed into sequences over the alphabet $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- The sequences where obtained using well defined rules taking into account the boundaries between prosodic words and the stressed syllabus.
- This analysis suggests a typical contexts tree for each language.

Preliminary study

Variable Length Markov Chains (VLMC) have been used to discriminate rhythmic patterns in European and Brazilian Portuguese written texts (Galves et al. 2005).

- The written texts where transformed into sequences over the alphabet $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- The sequences where obtained using well defined rules taking into account the boundaries between prosodic words and the stressed syllabus.
- This analysis suggests a typical contexts tree for each language.

Preliminary study

Variable Length Markov Chains (VLMC) have been used to discriminate rhythmic patterns in European and Brazilian Portuguese written texts (Galves et al. 2005).

- The written texts where transformed into sequences over the alphabet $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- The sequences where obtained using well defined rules taking into account the boundaries between prosodic words and the stressed syllabus.
- This analysis suggests a typical contexts tree for each language.

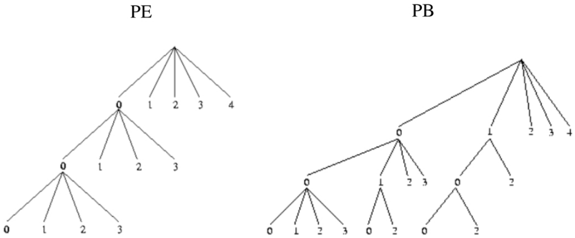
Preliminary study

Variable Length Markov Chains (VLMC) have been used to discriminate rhythmic patterns in European and Brazilian Portuguese written texts (Galves et al. 2005).

- The written texts where transformed into sequences over the alphabet $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- The sequences where obtained using well defined rules taking into account the boundaries between prosodic words and the stressed syllabus.
- This analysis suggests a typical contexts tree for each language.

Preliminary study

The typical trees for BP and EP are:



This study showed significant patterns that had been conjectured by linguists.

But it is also true that...

- The preceding results show high variability.
- The Context Algorithm used to estimate the trees is very sensitive to spurious strings.

But it is also true that...

- The preceding results show high variability.
- The Context Algorithm used to estimate the trees is very sensitive to spurious strings.

A new approach: probabilistic forests

A probabilistic forest is a mixture of probabilistic trees, that is

- A set Γ of probabilistic trees.
- A probability distribution over this set $\{p(\tau)\}_{\tau \in \Gamma}$.

A new approach: probabilistic forests

A probabilistic forest is a mixture of probabilistic trees, that is

- A set Γ of probabilistic trees.
- A probability distribution over this set $\{p(\tau)\}_{\tau \in \Gamma}$.

A new approach: probabilistic forests

A probabilistic forest is a mixture of probabilistic trees, that is

- A set Γ of probabilistic trees.
- A probability distribution over this set $\{p(\tau)\}_{\tau \in \Gamma}$.

How to define the set of trees Γ

- The set Γ depends on the problem and can be any set with a reasonable number of trees.
- For the linguistic problem, we choose the set of trees satisfying the algebraic restrictions given by the codification and with depth not bigger than 3.
- Given a sample sequence $x_1, x_2, \ldots, x_n$ (or a set of sample sequences) the probability of each tree in the set is estimated by the procedure followed in Eskin *et al.* (2002).

How to define the set of trees Γ

- The set Γ depends on the problem and can be any set with a reasonable number of trees.
- For the linguistic problem, we choose the set of trees satisfying the algebraic restrictions given by the codification and with depth not bigger than 3.
- Given a sample sequence $x_1, x_2, \ldots, x_n$ (or a set of sample sequences) the probability of each tree in the set is estimated by the procedure followed in Eskin *et al.* (2002).

How to define the set of trees Γ

- The set Γ depends on the problem and can be any set with a reasonable number of trees.
- For the linguistic problem, we choose the set of trees satisfying the algebraic restrictions given by the codification and with depth not bigger than 3.
- Given a sample sequence $x_1, x_2, \ldots, x_n$ (or a set of sample sequences) the probability of each tree in the set is estimated by the procedure followed in Eskin *et al.* (2002).

How to estimate the probability of each tree

Following Eskin *et al.* (2002):

- First, propose a *prior* weight $\omega^0(\tau)$.
- Then, for each instant of time i update this weight by the formula:

$$\omega^i(\tau) = \begin{cases} \omega^{i-1}(\tau)\frac{1}{|\mathcal{A}|}, & \text{if } i < d(\Gamma) \\ \omega^{i-1}(\tau)\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1})), & \text{if } n \geq d(\Gamma) \end{cases}$$

  where $\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1}))$ is the *Maximum Likelihood Estimate* of the transition probabilities in τ with the sample until time i, and $d(\Gamma)$ is the maximal depth of the trees in Γ.

How to estimate the probability of each tree

Following Eskin *et al.* (2002):

- First, propose a *prior* weight $\omega^0(\tau)$.
- Then, for each instant of time i update this weight by the formula:

$$\omega^i(\tau) = \begin{cases} \omega^{i-1}(\tau)\frac{1}{|\mathcal{A}|}, & \text{if } i < d(\Gamma) \\ \omega^{i-1}(\tau)\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1})), & \text{if } n \geq d(\Gamma) \end{cases}$$

where $\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1}))$ is the *Maximum Likelihood Estimate* of the transition probabilities in τ with the sample until time i, and $d(\Gamma)$ is the maximal depth of the trees in Γ.

A closed formula for $\omega^n(\tau)$

These weights $\omega^n(\tau)$ can be rewritten as:

$$\omega^n(\tau) = \omega^{n-1}(\tau)\hat{\mathbb{P}}^n_\tau(x_n|c_\tau(x_0,\ldots,x_{n-1}))$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{i=d(\Gamma)}^{n}\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1}))$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\prod_{a\in\mathcal{A}}\frac{(N_n(s,a)+1)!}{(N_n(s)+|\mathcal{A}|)!}$$

A closed formula for $\omega^n(\tau)$

These weights $\omega^n(\tau)$ can be rewritten as:

$$\omega^n(\tau) = \omega^{n-1}(\tau)\hat{\mathbb{P}}^n_\tau(x_n|c_\tau(x_0,\ldots,x_{n-1}))$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{i=d(\Gamma)}^{n}\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1}))$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\prod_{a\in\mathcal{A}}\frac{(N_n(s,a)+1)!}{(N_n(s)+|\mathcal{A}|)!}$$

A closed formula for $\omega^n(\tau)$

These weights $\omega^n(\tau)$ can be rewritten as:

$$\omega^n(\tau) = \omega^{n-1}(\tau)\hat{\mathbb{P}}^n_\tau(x_n|c_\tau(x_0,\ldots,x_{n-1}))$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{i=d(\Gamma)}^{n}\hat{\mathbb{P}}^i_\tau(x_i|c_\tau(x_0,\ldots,x_{i-1}))$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\prod_{a\in\mathcal{A}}\frac{(N_n(s,a)+1)!}{(N_n(s)+|\mathcal{A}|)!}$$

How to estimate the probability of each tree

Finally, we normalize the weights using the formula:

$$p^n(\tau) = \frac{\omega^n(\tau)}{\sum_{\tau' \in \Gamma} \omega^n(\tau')}$$

Therefore, $p^n(\tau)$ is the probability of tree τ estimated with the sample $x_1, x_2, \ldots, x_n$ by the preceding procedure.

Empirical results for EP

In the case of European Portuguese, we estimate the weights of all possible trees with depth less or equal 3. We used 31 texts of Portuguese authors.

$n = 148887$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.99$ for the following tree:

Empirical results for EP

In the case of European Portuguese, we estimate the weights of all possible trees with depth less or equal 3. We used 31 texts of Portuguese authors.

$n = 148887$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.99$ for the following tree:

Empirical results for EP

In the case of European Portuguese, we estimate the weights of all possible trees with depth less or equal 3. We used 31 texts of Portuguese authors.

$n = 148887$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.99$ for the following tree:

Empirical results for EP

In the case of European Portuguese, we estimate the weights of all possible trees with depth less or equal 3. We used 31 texts of Portuguese authors.

$n = 148887$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

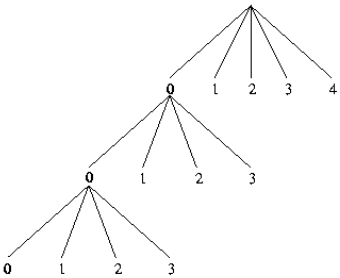
$p^n(\tau) = 0.99$ for the following tree:

Empirical results for EP

$n = 148887$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.99$ for the following tree:



Empirical results for BP

In the case of Brazilian Portuguese, we also estimate the weights of all possible trees with depth less or equal 3. We used 35 texts of Brazilian authors.

$n = 619282$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.8$ for the following tree:

In the case of Brazilian Portuguese, we also estimate the weights of all possible trees with depth less or equal 3. We used 35 texts of Brazilian authors.

$n = 619282$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.8$ for the following tree:

Empirical results for BP

In the case of Brazilian Portuguese, we also estimate the weights of all possible trees with depth less or equal 3. We used 35 texts of Brazilian authors.

$n = 619282$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.8$ for the following tree:

Empirical results for BP

In the case of Brazilian Portuguese, we also estimate the weights of all possible trees with depth less or equal 3. We used 35 texts of Brazilian authors.

$n = 619282$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

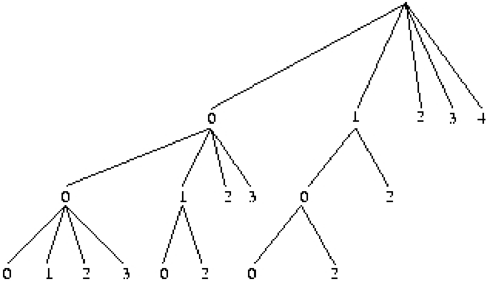
$p^n(\tau) = 0.8$ for the following tree:

Empirical results for BP

$n = 619282$.

$\omega^0(\tau) = (500n)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ.

$p^n(\tau) = 0.8$ for the following tree:



Consistency Theorem

Theorem

Let τ be a probabilistic tree. Then for almost all sequences $x_1, x_2, \ldots$ generated by τ, we have that

$$p^n(\tau) \rightarrow 1 \text{ when } n \rightarrow \infty.$$

Consistency Theorem

Idea of the proof

$$p^n(\tau) = \frac{\omega^n(\tau)}{\sum_{\tau' \in \Gamma} \omega^n(\tau')}$$

$$= \left(\sum_{\tau' \in \Gamma} \frac{\omega^n(\tau')}{\omega^n(\tau)} \right)^{-1}$$

$$= \left(1 + \sum_{\tau' \in \Gamma \setminus \{\tau\}} \frac{\omega^n(\tau')}{\omega^n(\tau)} \right)^{-1}$$

$$p^n(\tau) = \frac{\omega^n(\tau)}{\sum_{\tau' \in \Gamma} \omega^n(\tau')}$$

$$= \left(\sum_{\tau' \in \Gamma} \frac{\omega^n(\tau')}{\omega^n(\tau)} \right)^{-1}$$

$$= \left(1 + \sum_{\tau' \in \Gamma \setminus \{\tau\}} \frac{\omega^n(\tau')}{\omega^n(\tau)} \right)^{-1}$$

Consistency Theorem

Idea of the proof

$$p^n(\tau) = \frac{\omega^n(\tau)}{\sum_{\tau' \in \Gamma} \omega^n(\tau')}$$

$$= \left(\sum_{\tau' \in \Gamma} \frac{\omega^n(\tau')}{\omega^n(\tau)} \right)^{-1}$$

$$= \left(1 + \sum_{\tau' \in \Gamma \setminus \{\tau\}} \frac{\omega^n(\tau')}{\omega^n(\tau)} \right)^{-1}$$

$$\omega^n(\tau) = \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\prod_{a\in\mathcal{A}}\frac{(N_n(s,a)+1)!}{(N_n(s)+|\mathcal{A}|)!}$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\hat{\mathbb{P}}_{KT,s}(x_1^n)$$

$$\omega^n(\tau) = \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\prod_{a\in\mathcal{A}}\frac{(N_n(s,a)+1)!}{(N_n(s)+|\mathcal{A}|)!}$$

$$= \omega^0(\tau)\frac{1}{|\mathcal{A}|^{d(\Gamma)}}\prod_{s\in\tau}\hat{\mathbb{P}}_{KT,s}(x_1^n)$$

$$\frac{\omega^n(\tau')}{\omega^n(\tau)} = \frac{\omega^0(\tau')}{\omega^0(\tau)} \frac{\prod_{s' \in \tau'} \hat{\mathbb{P}}_{KT,s'}(x_1^n)}{\prod_{s \in \tau} \hat{\mathbb{P}}_{KT,s}(x_1^n)}$$

$$= \frac{\omega^0(\tau')}{\omega^0(\tau)} \prod_{s \in \tau, s \prec \tau'} \frac{\prod_{s' \in \tau', s' > s} \hat{\mathbb{P}}_{KT,s'}(x_1^n)}{\hat{\mathbb{P}}_{KT,s}(x_1^n)}$$

$$\prod_{s' \in \tau', s' \prec \tau} \frac{\hat{\mathbb{P}}_{KT,s'}(x_1^n)}{\prod_{s \in \tau, s > s'} \hat{\mathbb{P}}_{KT,s}(x_1^n)}$$

Consistency Theorem

Idea of the proof

$$\frac{\omega^n(\tau')}{\omega^n(\tau)} = \frac{\omega^0(\tau')}{\omega^0(\tau)} \frac{\prod_{s'\in\tau'} \hat{\mathbb{P}}_{KT,s'}(x_1^n)}{\prod_{s\in\tau} \hat{\mathbb{P}}_{KT,s}(x_1^n)}$$

$$= \frac{\omega^0(\tau')}{\omega^0(\tau)} \prod_{s\in\tau, s\prec\tau'} \frac{\prod_{s'\in\tau', s'>s} \hat{\mathbb{P}}_{KT,s'}(x_1^n)}{\hat{\mathbb{P}}_{KT,s}(x_1^n)}$$

$$\prod_{s'\in\tau', s'\prec\tau} \frac{\hat{\mathbb{P}}_{KT,s'}(x_1^n)}{\prod_{s\in\tau, s>s'} \hat{\mathbb{P}}_{KT,s}(x_1^n)}$$

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} = \log \frac{\omega^0(\tau')}{\omega^0(\tau)} +$$

$$+ \sum_{s \in \tau, s \prec \tau'} \left[\sum_{s' \in \tau', s' > s} \log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$+ \sum_{s' \in \tau', s' \prec \tau} \left[\log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \sum_{s \in \tau, s > s' \prec \tau} \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} = \log \frac{\omega^0(\tau')}{\omega^0(\tau)} +$$

$$+ \sum_{s\in\tau, s\prec\tau'} \left[\sum_{s'\in\tau', s'>s} \log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$+ \sum_{s'\in\tau', s'\prec\tau} \left[\log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \sum_{s\in\tau, s>s'\prec\tau} \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} = \log \frac{\omega^0(\tau')}{\omega^0(\tau)} +$$

$$+ \sum_{s \in \tau, s \prec \tau'} \left[\sum_{s' \in \tau', s' > s} \log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$+ \sum_{s' \in \tau', s' \prec \tau} \left[\log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \sum_{s \in \tau, s > s' \prec \tau} \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} = \log \frac{\omega^0(\tau')}{\omega^0(\tau)} +$$

$$+ \sum_{s \in \tau, s \prec \tau'} \left[\sum_{s' \in \tau', s' > s} \log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

$$+ \sum_{s' \in \tau', s' \prec \tau} \left[\log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \sum_{s \in \tau, s > s' \prec \tau} \log \hat{\mathbb{P}}_{KT,s}(x_1^n) \right]$$

Following the ideas of Csiszár and Talata (2005) we proved that

Lemma

Let $s \in \tau$ such that s is a proper suffix of some context in τ'. Then there exists a constant $c < 0$ such that

$$\sum_{s' \in \tau', s' > s} \log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \log \hat{\mathbb{P}}_{KT,s}(x_1^n) < c \log n,$$

eventually almost surely as $n \to \infty$.

Lemma

Let $s' \in \tau'$ such that s' is a proper suffix of some context in τ. Then there exists a constant $c < 0$ such that

$$\log \hat{\mathbb{P}}_{KT,s'}(x_1^n) - \sum_{s \in \tau, s > s'} \log \hat{\mathbb{P}}_{KT,s}(x_1^n) < cn,$$

eventually almost surely as $n \to \infty$.

Then there exist constants $C_1 \leq 0$ and $C_2 \leq 0$, not vanishing simultaneously, such that

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} < \log \frac{\omega^0(\tau')}{\omega^0(\tau)} + C_1 \log n + C_2 n$$

With $\omega^0(\tau) = (cn)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ we have that

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} \to -\infty \,,$$

when $n \to \infty$.

Then there exist constants $C_1 \leq 0$ and $C_2 \leq 0$, not vanishing simultaneously, such that

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} < \log \frac{\omega^0(\tau')}{\omega^0(\tau)} + C_1 \log n + C_2 n$$

With $\omega^0(\tau) = (cn)^{-t(\tau)}$, where $t(\tau)$ is the number of terminal nodes of τ we have that

$$\log \frac{\omega^n(\tau')}{\omega^n(\tau)} \to -\infty\,,$$

when $n \to \infty$.

Conclusion

“Only one tree represents the forest”