Árvores e Florestas Probabilísticas

e como elas ajudam a distinguir os ritmos do Português Brasileiro e do Português Europeu

Antonio Galves

Instituto de Matemática e Estatística
Universidade de São Paulo

Colaboradores

Este é o resumo de um artigo de pesquisa que estou escrevendo com

- Florencia Leonardi (minha orientanda de Doutoramento em Bioinformática)
- Denis Lacerda (meu orientando de Iniciação Científica)
- como parte do Projeto Pronex/Temático Fapesp *Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos em línguas naturais*

Colaboradores

Este é o resumo de um artigo de pesquisa que estou escrevendo com

- Florencia Leonardi (minha orientanda de Doutoramento em Bioinformática)
- Denis Lacerda (meu orientando de Iniciação Científica)
- como parte do Projeto Pronex/Temático Fapesp *Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos em línguas naturais*

Colaboradores

Este é o resumo de um artigo de pesquisa que estou escrevendo com

- Florencia Leonardi (minha orientanda de Doutoramento em Bioinformática)
- Denis Lacerda (meu orientando de Iniciação Científica)
- como parte do Projeto Pronex/Temático Fapesp *Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos em línguas naturais*

Colaboradores

Este é o resumo de um artigo de pesquisa que estou escrevendo com

- Florencia Leonardi (minha orientanda de Doutoramento em Bioinformática)
- Denis Lacerda (meu orientando de Iniciação Científica)
- como parte do Projeto Pronex/Temático Fapesp *Comportamento estocástico, fenômenos críticos e identificação de padrões rítmicos em línguas naturais*

Modelagem estocástica de seqüências portadoras de informação

Seqüências genéticas, cadeias de amino-ácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum

- um comportamento que, apesar de não ser determinístico,
- contém informações precisas a respeito do sistema que as produziu.
- No caso de cadeias linguísticas uma dessas características parece estar codificada no *ritmo*.
- Mas, o que é o *ritmo* de uma língua?

Modelagem estocástica de seqüências portadoras de informação

Seqüências genéticas, cadeias de amino-ácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum

- um comportamento que, apesar de não ser determinístico,
- contém informações precisas a respeito do sistema que as produziu.
- No caso de cadeias linguísticas uma dessas características parece estar codificada no *ritmo*.
- Mas, o que é o *ritmo* de uma língua?

Modelagem estocástica de seqüências portadoras de informação

Seqüências genéticas, cadeias de amino-ácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum

- um comportamento que, apesar de não ser determinístico,
- contém informações precisas a respeito do sistema que as produziu.
- No caso de cadeias linguísticas uma dessas características parece estar codificada no *ritmo*.
- Mas, o que é o *ritmo* de uma língua?

Modelagem estocástica de seqüências portadoras de informação

Seqüências genéticas, cadeias de amino-ácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum

- um comportamento que, apesar de não ser determinístico,
- contém informações precisas a respeito do sistema que as produziu.
- No caso de cadeias linguísticas uma dessas características parece estar codificada no *ritmo*.
- Mas, o que é o *ritmo* de uma língua?

Modelagem estocástica de seqüências portadoras de informação

Seqüências genéticas, cadeias de amino-ácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum

- um comportamento que, apesar de não ser determinístico,
- contém informações precisas a respeito do sistema que as produziu.
- No caso de cadeias linguísticas uma dessas características parece estar codificada no *ritmo*.
- Mas, o que é o *ritmo* de uma língua?

A conjectura das classes rítmicas

Lloyd James (anos 40) e Abercrombie (anos 50) conjecturam que as línguas se agrupam em classes rítmicas.
Classes rítmicas conjecturadas:

- línguas acentuais: Holandês, Inglês, Polonês, Português Europeu,...
- línguas silábicas: Catalão, Espanhol, Francês, Italiano, Português Brasileiro, ...
- línguas moraicas: Japonês, ...

A conjectura das classes rítmicas

Lloyd James (anos 40) e Abercrombie (anos 50) conjecturam que as línguas se agrupam em classes rítmicas.
Classes rítmicas conjecturadas:

- línguas acentuais: Holandês, Inglês, Polonês, Português Europeu,...
- línguas silábicas: Catalão, Espanhol, Francês, Italiano, Português Brasileiro, ...
- línguas moraicas: Japonês, ...

A conjectura das classes rítmicas

Lloyd James (anos 40) e Abercrombie (anos 50) conjecturam que as línguas se agrupam em classes rítmicas.
Classes rítmicas conjecturadas:

- línguas acentuais: Holandês, Inglês, Polonês, Português Europeu,...
- línguas silábicas: Catalão, Espanhol, Francês, Italiano, Português Brasileiro, ...
- línguas moraicas: Japonês, ...

A conjectura das classes rítmicas

Lloyd James (anos 40) e Abercrombie (anos 50) conjecturam que as línguas se agrupam em classes rítmicas.
Classes rítmicas conjecturadas:

- línguas acentuais: Holandês, Inglês, Polonês, Português Europeu,...
- línguas silábicas: Catalão, Espanhol, Francês, Italiano, Português Brasileiro, ...
- línguas moraicas: Japonês, ...

O que caracteriza uma classe rítmica?

- Até recentemente ninguém havia encontrado correlatos do ritmo no sinal acústico de fala.
- Isso acontece pela primeira vez com um artigo de 1999, assinado por um jovem doutorando Franck Ramus e co-assinado por seus orientadores Marina Nespor e Jacques Mehler.
- Numa outra palestra poderíamos falar da noção de *sonoridade* e de como ela pode ser usada para classificar amostras de fala.

O que caracteriza uma classe rítmica?

- Até recentemente ninguém havia encontrado correlatos do ritmo no sinal acústico de fala.
- Isso acontece pela primeira vez com um artigo de 1999, assinado por um jovem doutorando Franck Ramus e co-assinado por seus orientadores Marina Nespor e Jacques Mehler.
- Numa outra palestra poderíamos falar da noção de *sonoridade* e de como ela pode ser usada para classificar amostras de fala.

O que caracteriza uma classe rítmica?

- Até recentemente ninguém havia encontrado correlatos do ritmo no sinal acústico de fala.
- Isso acontece pela primeira vez com um artigo de 1999, assinado por um jovem doutorando Franck Ramus e co-assinado por seus orientadores Marina Nespor e Jacques Mehler.
- Numa outra palestra poderíamos falar da noção de *sonoridade* e de como ela pode ser usada para classificar amostras de fala.

O que caracteriza uma classe rítmica?

- Até recentemente ninguém havia encontrado correlatos do ritmo no sinal acústico de fala.
- Isso acontece pela primeira vez com um artigo de 1999, assinado por um jovem doutorando Franck Ramus e co-assinado por seus orientadores Marina Nespor e Jacques Mehler.
- Numa outra palestra poderíamos falar da noção de *sonoridade* e de como ela pode ser usada para classificar amostras de fala.

A identificação do ritmo em textos escritos

- Como extrair padrões rítmicos de textos escritos?
- Possivelmente o texto completo com todas as letras tem informação demais
- e o ritmo fica escondido numa cadeia subjacente mais simples.
- Que cadeia será essa?

A identificação do ritmo em textos escritos

- Como extrair padrões rítmicos de textos escritos?
- Possivelmente o texto completo com todas as letras tem informação demais
- e o ritmo fica escondido numa cadeia subjacente mais simples.
- Que cadeia será essa?

A identificação do ritmo em textos escritos

- Como extrair padrões rítmicos de textos escritos?
- Possivelmente o texto completo com todas as letras tem informação demais
- e o ritmo fica escondido numa cadeia subjacente mais simples.
- Que cadeia será essa?

A identificação do ritmo em textos escritos

- Como extrair padrões rítmicos de textos escritos?
- Possivelmente o texto completo com todas as letras tem informação demais
- e o ritmo fica escondido numa cadeia subjacente mais simples.
- Que cadeia será essa?

A identificação do ritmo em textos escritos

- Como extrair padrões rítmicos de textos escritos?
- Possivelmente o texto completo com todas as letras tem informação demais
- e o ritmo fica escondido numa cadeia subjacente mais simples.
- Que cadeia será essa?

Uma cadeia simbólica de marcas rítmicas

Vamos fazer uma tentativa de marcar os elementos pertinentes do ritmo indicando para cada sílaba

- se ela carrega ou não o acento principal da palavra
- se ela é ou não começo de palavra

Uma cadeia simbólica de marcas rítmicas

Vamos fazer uma tentativa de marcar os elementos pertinentes do ritmo indicando para cada sílaba

- se ela carrega ou não o acento principal da palavra
- se ela é ou não começo de palavra

Uma cadeia simbólica de marcas rítmicas

Vamos fazer uma tentativa de marcar os elementos pertinentes do ritmo indicando para cada sílaba

- se ela carrega ou não o acento principal da palavra
- se ela é ou não começo de palavra

Acentos principais

- Em Português, com exeção das preposições, artigos e outras palavras funcionais, todas as palavras têm uma sílaba marcada por um *acento principal*.
- Exemplo: a palavra *menino* tem tres sílabas: *me-ni-no*.
- O acento principal está na segunda sílaba: *ni*.

Acentos principais

- Em Português, com exeção das preposições, artigos e outras palavras funcionais, todas as palavras têm uma sílaba marcada por um *acento principal*.
- Exemplo: a palavra *menino* tem tres sílabas: *me-ni-no*.
- O acento principal está na segunda sílaba: *ni*.

Acentos principais

- Em Português, com exeção das preposições, artigos e outras palavras funcionais, todas as palavras têm uma sílaba marcada por um *acento principal*.
- Exemplo: a palavra *menino* tem tres sílabas: *me-ni-no*.
- O acento principal está na segunda sílaba: *ni*.

Acentos principais

- Em Português, com exeção das preposições, artigos e outras palavras funcionais, todas as palavras têm uma sílaba marcada por um *acento principal*.
- Exemplo: a palavra *menino* tem tres sílabas: *me-ni-no*.
- O acento principal está na segunda sílaba: *ni*.

Palavras prosódicas

Vamos chamar de palavra prosódica a seqüência formada por

- a palavra com seu acento principal e
- todas as palavras funcionais que a precedem até a encontrar a palavra anterior com acento.
- Exemplo: a frase *O menino comeu a bala* tem tres palavras prosódicas: (O menino) (comeu) (a bala).

Palavras prosódicas

Vamos chamar de palavra prosódica a seqüência formada por

- a palavra com seu acento principal e
- todas as palavras funcionais que a precedem até a encontrar a palavra anterior com acento.
- Exemplo: a frase *O menino comeu a bala* tem tres palavras prosódicas: (O menino) (comeu) (a bala).

Palavras prosódicas

Vamos chamar de palavra prosódica a seqüência formada por

- a palavra com seu acento principal e
- todas as palavras funcionais que a precedem até a encontrar a palavra anterior com acento.
- Exemplo: a frase *O menino comeu a bala* tem tres palavras prosódicas: (O menino) (comeu) (a bala).

Palavras prosódicas

Vamos chamar de palavra prosódica a seqüência formada por

- a palavra com seu acento principal e
- todas as palavras funcionais que a precedem até a encontrar a palavra anterior com acento.
- Exemplo: a frase *O menino comeu a bala* tem tres palavras prosódicas: (O menino) (comeu) (a bala).

Codificação

Isso nos leva a usar

$$\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$$

como o conjunto de símbolos onde

- o primeiro dígito indica se a sílaba é começo de palavra prosódica (0= não, 1=sim)
- o segundo dígito indica se a síliba carrega o acento principal da palavra (0=não, 1=sim).
- Em representação binária: $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ e $(1,1) = 3$.
- Vamos acrescentar o símbolo 4 para indicar o começo de frase.

Codificação

Isso nos leva a usar

$$\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$$

como o conjunto de símbolos onde

- o primeiro dígito indica se a sílaba é começo de palavra prosódica (0= não, 1=sim)
- o segundo dígito indica se a síliba carrega o acento principal da palavra (0=não, 1=sim).
- Em representação binária: $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ e $(1,1) = 3$.
- Vamos acrescentar o símbolo 4 para indicar o começo de frase.

Codificação

Isso nos leva a usar

$$\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$$

como o conjunto de símbolos onde

- o primeiro dígito indica se a sílaba é começo de palavra prosódica (0= não, 1=sim)
- o segundo dígito indica se a síliba carrega o acento principal da palavra (0=não, 1=sim).
- Em representação binária: $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ e $(1,1) = 3$.
- Vamos acrescentar o símbolo 4 para indicar o começo de frase.

Codificação

Isso nos leva a usar

$$\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$$

como o conjunto de símbolos onde

- o primeiro dígito indica se a sílaba é começo de palavra prosódica (0= não, 1=sim)
- o segundo dígito indica se a síliba carrega o acento principal da palavra (0=não, 1=sim).
- Em representação binária: $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ e $(1,1) = 3$.
- Vamos acrescentar o símbolo 4 para indicar o começo de frase.

Codificação

Isso nos leva a usar

$$\{0,1\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$$

como o conjunto de símbolos onde

- o primeiro dígito indica se a sílaba é começo de palavra prosódica (0= não, 1=sim)
- o segundo dígito indica se a síliba carrega o acento principal da palavra (0=não, 1=sim).
- Em representação binária: $(0,0) = 0$, $(0,1) = 1$, $(1,0) = 2$ e $(1,1) = 3$.
- Vamos acrescentar o símbolo 4 para indicar o começo de frase.

Exemplo

- O começo da frase: *O menino já comeu o doce*
- é codificado como

Frase	.	O	me	ni	no
Começo de palavra prosódica		1	0	0	0
Sílaba acentuada		0	0	1	0
Código	4	2	0	1	0

- A frase inteira é codificada como

.	O	me	ni	no	já	co	meu	o	do	ce
4	2	0	1	0	3	2	1	2	1	0

Exemplo

- O começo da frase: *O menino já comeu o doce*
- é codificado como

Frase	.	O	me	ni	no
Começo de palavra prosódica		1	0	0	0
Sílaba acentuada		0	0	1	0
Código	4	2	0	1	0

- A frase inteira é codificada como

.	O	me	ni	no	já	co	meu	o	do	ce
4	2	0	1	0	3	2	1	2	1	0

Exemplo

- O começo da frase: *O menino já comeu o doce*
- é codificado como

Frase	.	O	me	ni	no
Começo de palavra prosódica		1	0	0	0
Sílaba acentuada		0	0	1	0
Código	4	2	0	1	0

- A frase inteira é codificada como

.	O	me	ni	no	já	co	meu	o	do	ce
4	2	0	1	0	3	2	1	2	1	0

Cadeias simbólicas estocásticas

- Com essa codificação, cada texto é transformado numa seqüência de símbolos no alfabeto $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- Essas seqüências não apresentam padrões identificáveis a olho nu.
- Então é necessário subir um nível na abstração e tentar identificar padrões na classe de modelos probabilístico capazes de gerar seqüências desse tipo.
- Esses modelos são as *Cadeias de Markov de Alcance Variável*.

Cadeias simbólicas estocásticas

- Com essa codificação, cada texto é transformado numa seqüência de símbolos no alfabeto $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- Essas seqüências não apresentam padrões identificáveis a olho nu.
- Então é necessário subir um nível na abstração e tentar identificar padrões na classe de modelos probabilístico capazes de gerar seqüências desse tipo.
- Esses modelos são as *Cadeias de Markov de Alcance Variável*.

Cadeias simbólicas estocásticas

- Com essa codificação, cada texto é transformado numa seqüência de símbolos no alfabeto $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- Essas seqüências não apresentam padrões identificáveis a olho nu.
- Então é necessário subir um nível na abstração e tentar identificar padrões na classe de modelos probabilístico capazes de gerar seqüências desse tipo.
- Esses modelos são as *Cadeias de Markov de Alcance Variável*.

Cadeias simbólicas estocásticas

- Com essa codificação, cada texto é transformado numa seqüência de símbolos no alfabeto $\mathcal{A} = \{0, 1, 2, 3, 4\}$.
- Essas seqüências não apresentam padrões identificáveis a olho nu.
- Então é necessário subir um nível na abstração e tentar identificar padrões na classe de modelos probabilístico capazes de gerar seqüências desse tipo.
- Esses modelos são as *Cadeias de Markov de Alcance Variável*.

Cadeias de Markov de alcance variável

- Dada uma seqüência simbólica $X_1, X_2, \ldots$ assumindo valores num alfabeto finito, tenta-se *predizer* cada novo símbolo X_n como função do *passado* $X_1, \ldots, X_{n-1}$.
- Idéia básica: Apenas uma porção do passado é relevante para a determinação do próximo símbolo.
- O comprimento dessa porção relevante varia de um passado para outro (isso explica o nome cadeia de *memória variável*).

A porção relevante do passado será chamada de *contexto*.

Cadeias de Markov de alcance variável

- Dada uma seqüência simbólica $X_1, X_2, \ldots$ assumindo valores num alfabeto finito, tenta-se *predizer* cada novo símbolo X_n como função do *passado* $X_1, \ldots, X_{n-1}$.
- Idéia básica: Apenas uma porção do passado é relevante para a determinação do próximo símbolo.
- O comprimento dessa porção relevante varia de um passado para outro (isso explica o nome cadeia de *memória variável*).

A porção relevante do passado será chamada de *contexto*.

Cadeias de Markov de alcance variável

- Dada uma seqüência simbólica $X_1, X_2, \ldots$ assumindo valores num alfabeto finito, tenta-se *predizer* cada novo símbolo X_n como função do *passado* $X_1, \ldots, X_{n-1}$.
- Idéia básica: Apenas uma porção do passado é relevante para a determinação do próximo símbolo.
- O comprimento dessa porção relevante varia de um passado para outro (isso explica o nome cadeia de *memória variável*).

A porção relevante do passado será chamada de *contexto*.

Cadeias de Markov de alcance variável

- Dada uma seqüência simbólica $X_1, X_2, \ldots$ assumindo valores num alfabeto finito, tenta-se *predizer* cada novo símbolo X_n como função do *passado* $X_1, \ldots, X_{n-1}$.
- Idéia básica: Apenas uma porção do passado é relevante para a determinação do próximo símbolo.
- O comprimento dessa porção relevante varia de um passado para outro (isso explica o nome cadeia de *memória variável*).

A porção relevante do passado será chamada de *contexto*.

Árvores probabilísticas

- O conjunto de contextos tem a *propriedade do sufixo*: nenhum contexto é sufixo de outro.
- Exemplo: no alfabeto binário $\mathcal{A} = \{0, 1\}$ poderiamos ter os contextos $\{(1); (0, 0); (1, 0)\}$ (o tempo é lido da esquerda para a direita).
- Isso quer dizer que se o símbolo precedente for 1, não é necessário recuar mais no passado para poder predizer o próximo símbolo.
- Já se o símbolo precedente for 0, é necessário recuar mais um passo no passado para poder predizer o próximo símbolo.

Árvores probabilísticas

- O conjunto de contextos tem a *propriedade do sufixo*: nenhum contexto é sufixo de outro.
- Exemplo: no alfabeto binário $\mathcal{A} = \{0, 1\}$ poderiamos ter os contextos $\{(1); (0, 0); (1, 0)\}$ (o tempo é lido da esquerda para a direita).
- Isso quer dizer que se o símbolo precedente for 1, não é necessário recuar mais no passado para poder predizer o próximo símbolo.
- Já se o símbolo precedente for 0, é necessário recuar mais um passo no passado para poder predizer o próximo símbolo.

Árvores probabilísticas

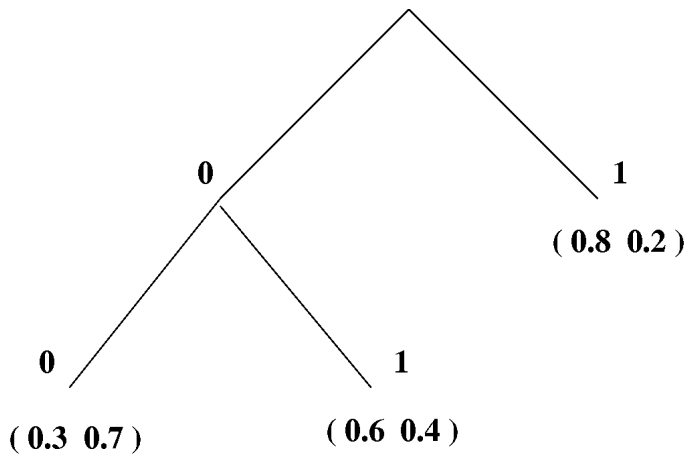
- O conjunto de contextos tem a *propriedade do sufixo*: nenhum contexto é sufixo de outro.
- Exemplo: no alfabeto binário $\mathcal{A} = \{0, 1\}$ poderiamos ter os contextos $\{(1); (0, 0); (1, 0)\}$ (o tempo é lido da esquerda para a direita).
- Isso quer dizer que se o símbolo precedente for 1, não é necessário recuar mais no passado para poder predizer o próximo símbolo.
- Já se o símbolo precedente for 0, é necessário recuar mais um passo no passado para poder predizer o próximo símbolo.

Árvores probabilísticas

- O conjunto de contextos tem a *propriedade do sufixo*: nenhum contexto é sufixo de outro.
- Exemplo: no alfabeto binário $\mathcal{A} = \{0, 1\}$ poderiamos ter os contextos $\{(1); (0, 0); (1, 0)\}$ (o tempo é lido da esquerda para a direita).
- Isso quer dizer que se o símbolo precedente for 1, não é necessário recuar mais no passado para poder predizer o próximo símbolo.
- Já se o símbolo precedente for 0, é necessário recuar mais um passo no passado para poder predizer o próximo símbolo.

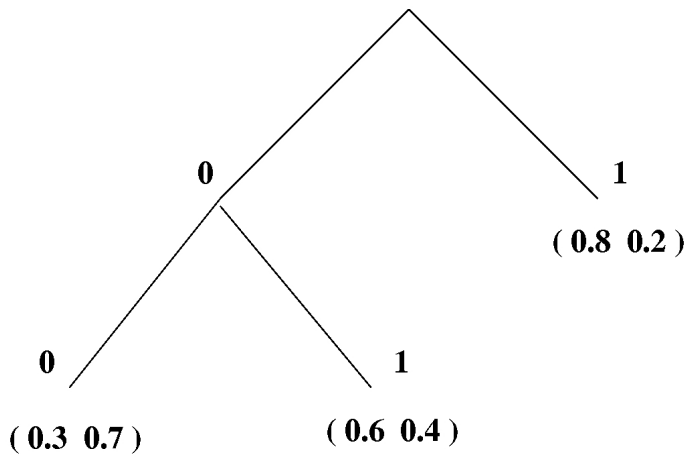
A propriedade do sufixo

- Nenhum contexto no conjunto de contextos $\{(1); (0,0); (1,0)\}$ é sufixo de outro.
- Isso torna possível representar este conjunto como uma árvore
-



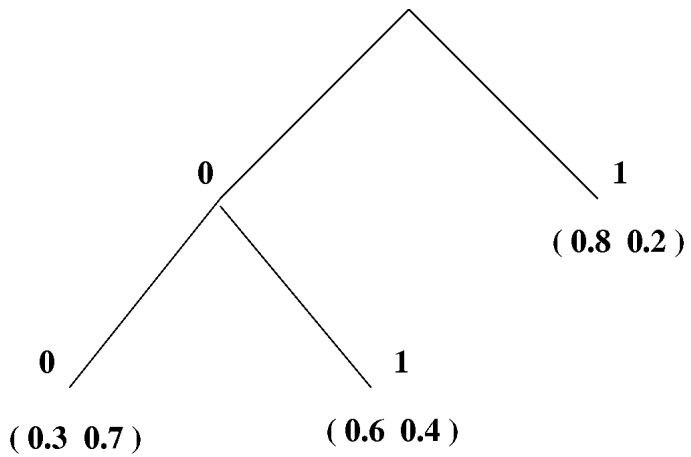
A propriedade do sufixo

- Nenhum contexto no conjunto de contextos $\{(1); (0,0); (1,0)\}$ é sufixo de outro.
- Isso torna possível representar este conjunto como uma árvore



A propriedade do sufixo

- Nenhum contexto no conjunto de contextos $\{(1); (0,0); (1,0)\}$ é sufixo de outro.
- Isso torna possível representar este conjunto como uma árvore
-



Vantagens conceituais do modelo de memória variável

- Captura **dependências de comprimento variável** entre os símbolos da seqüência.
- É capaz de identificar **características estruturais** nas seqüências.
- Aplicação do método nas cadeias codificadas com as marcas rítmicas identifica distintas para o Português Brasileiro e o Português Europeu
- Isso é feito através de Algoritmo do Contexto, introduzido por Rissanen na década do 80.

Vantagens conceituais do modelo de memória variável

- Captura **dependências de comprimento variável** entre os símbolos da seqüência.
- É capaz de identificar **características estruturais** nas seqüências.
- Aplicação do método nas cadeias codificadas com as marcas rítmicas identifica distintas para o Português Brasileiro e o Português Europeu
- Isso é feito através de Algoritmo do Contexto, introduzido por Rissanen na década do 80.

Vantagens conceituais do modelo de memória variável

- Captura **dependências de comprimento variável** entre os símbolos da seqüência.
- É capaz de identificar **características estruturais** nas seqüências.
- Aplicação do método nas cadeias codificadas com as marcas rítmicas identifica distintas para o Português Brasileiro e o Português Europeu
- Isso é feito através de Algoritmo do Contexto, introduzido por Rissanen na década do 80.

Vantagens conceituais do modelo de memória variável

- Captura **dependências de comprimento variável** entre os símbolos da seqüência.
- É capaz de identificar **características estruturais** nas seqüências.
- Aplicação do método nas cadeias codificadas com as marcas rítmicas identifica distintas para o Português Brasileiro e o Português Europeu
- Isso é feito através de Algoritmo do Contexto, introduzido por Rissanen na década do 80.

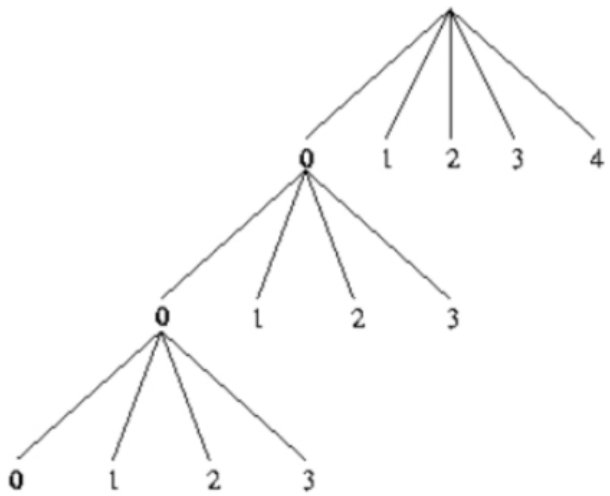
Vantagens conceituais do modelo de memória variável

- Captura **dependências de comprimento variável** entre os símbolos da seqüência.
- É capaz de identificar **características estruturais** nas seqüências.
- Aplicação do método nas cadeias codificadas com as marcas rítmicas identifica distintas para o Português Brasileiro e o Português Europeu
- Isso é feito através de Algoritmo do Contexto, introduzido por Rissanen na década do 80.

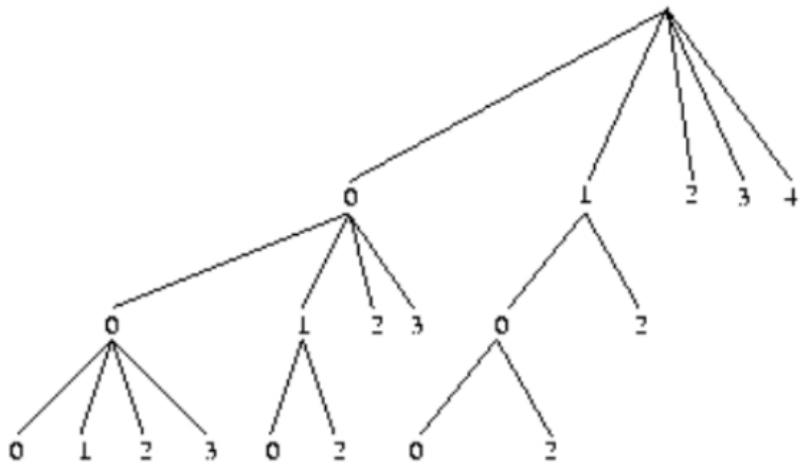
Árvores associadas ao PB e ao PE

Essas são as árvores que mais frequentemente aparecem em textos codificados de PB e PE:

PE



PB



A diferença entre as duas árvores tem um significado lingüístico.

Dificuldades com o Algoritmo do Contexto

- Ele converge lentamente (precisamos de amostras muito grandes para obter o resultado assintótico e isso é um problema com dados biológicos)
- Ele não é robusto (a contaminação com pequenas seqüências espúrias muda dramaticamente o resultado).
- Talvez por isso muitas vezes as árvores obtidas são diferentes das árvores lingüisticamente significantes.
- Ou seja, em vez de uma única árvore acabamos obtendo uma floresta de árvores. *A floresta pode ocultar a árvore lingüisticamente significativa*.

Dificuldades com o Algoritmo do Contexto

- Ele converge lentamente (precisamos de amostras muito grandes para obter o resultado assintótico e isso é um problema com dados biológicos)
- Ele não é robusto (a contaminação com pequenas seqüências espúrias muda dramaticamente o resultado).
- Talvez por isso muitas vezes as árvores obtidas são diferentes das árvores lingüisticamente significantes.
- Ou seja, em vez de uma única árvore acabamos obtendo uma floresta de árvores. *A floresta pode ocultar a árvore lingüisticamente significativa*.

Dificuldades com o Algoritmo do Contexto

- Ele converge lentamente (precisamos de amostras muito grandes para obter o resultado assintótico e isso é um problema com dados biológicos)
- Ele não é robusto (a contaminação com pequenas seqüências espúrias muda dramaticamente o resultado).
- Talvez por isso muitas vezes as árvores obtidas são diferentes das árvores lingüisticamente significantes.
- Ou seja, em vez de uma única árvore acabamos obtendo uma floresta de árvores. *A floresta pode ocultar a árvore lingüisticamente significativa*.

Dificuldades com o Algoritmo do Contexto

- Ele converge lentamente (precisamos de amostras muito grandes para obter o resultado assintótico e isso é um problema com dados biológicos)
- Ele não é robusto (a contaminação com pequenas seqüências espúrias muda dramaticamente o resultado).
- Talvez por isso muitas vezes as árvores obtidas são diferentes das árvores lingüisticamente significantes.
- Ou seja, em vez de uma única árvore acabamos obtendo uma floresta de árvores. *A floresta pode ocultar a árvore lingüisticamente significativa*.

Dificuldades com o Algoritmo do Contexto

- Ele converge lentamente (precisamos de amostras muito grandes para obter o resultado assintótico e isso é um problema com dados biológicos)
- Ele não é robusto (a contaminação com pequenas seqüências espúrias muda dramaticamente o resultado).
- Talvez por isso muitas vezes as árvores obtidas são diferentes das árvores lingüisticamente significantes.
- Ou seja, em vez de uma única árvore acabamos obtendo uma floresta de árvores. *A floresta pode ocultar a árvore lingüisticamente significativa*.

Florestas probabilísticas

- Uma cadeia de alcance variável é representada por uma árvore de contextos e por uma família de probabilidades de transição associadas a esses contextos. E isso que chamamos de árvore probabilística.
- Uma floresta probabilística é um conjunto Γ de árvores probabilísticas junto com uma distribuição de probabilidades sobre esse conjunto, $\{\omega(\tau)\}_{\tau\in\Gamma}$.
- Uma floresta probabilística pode ser interpretada como se em cada frase escolhêssemos um modelo de acordo com a distribuição ω e com esse modelo gerássemos a próxima frase.

Florestas probabilísticas

- Uma cadeia de alcance variável é representada por uma árvore de contextos e por uma família de probabilidades de transição associadas a esses contextos. E isso que chamamos de árvore probabilística.
- Uma floresta probabilística é um conjunto Γ de árvores probabilísticas junto com uma distribuição de probabilidades sobre esse conjunto, $\{\omega(\tau)\}_{\tau\in\Gamma}$.
- Uma floresta probabilística pode ser interpretada como se em cada frase escolhêssemos um modelo de acordo com a distribuição ω e com esse modelo gerássemos a próxima frase.

Florestas probabilísticas

- Uma cadeia de alcance variável é representada por uma árvore de contextos e por uma família de probabilidades de transição associadas a esses contextos. E isso que chamamos de árvore probabilística.
- Uma floresta probabilística é um conjunto Γ de árvores probabilísticas junto com uma distribuição de probabilidades sobre esse conjunto, $\{\omega(\tau)\}_{\tau\in\Gamma}$.
- Uma floresta probabilística pode ser interpretada como se em cada frase escolhêssemos um modelo de acordo com a distribuição ω e com esse modelo gerássemos a próxima frase.

Florestas probabilísticas

- Uma cadeia de alcance variável é representada por uma árvore de contextos e por uma família de probabilidades de transição associadas a esses contextos. E isso que chamamos de árvore probabilística.
- Uma floresta probabilística é um conjunto Γ de árvores probabilísticas junto com uma distribuição de probabilidades sobre esse conjunto, $\{\omega(\tau)\}_{\tau\in\Gamma}$.
- Uma floresta probabilística pode ser interpretada como se em cada frase escolhêssemos um modelo de acordo com a distribuição ω e com esse modelo gerássemos a próxima frase.

Como estimar o “peso” de cada árvore

A probabilidade *a posteriori* de cada árvore é estimada a partir de uma amostra ou um conjunto de amostras. No instante zero propomos um peso *a priori* $\omega^0(\tau)$ e em cada instante de tempo atualizamos esse peso pela fórmula:

$$\omega^{n+1}(\tau) = \omega^n(\tau) P_\tau^n(x_n | c_\tau(x_0, \ldots, x_{n-1}))$$

onde $P_\tau^n(x_n | c_\tau(x_0, \ldots, x_{n-1}))$ é o estimador de máxima verossimilhança estimado com a amostra até o tempo n.

Como implementar a proposta

- Gerar todas as árvores possíveis
- Resolver o problema do aumento da precisão
  - Guardar apenas os dígitos mais significativos
- Calcular os pesos e as probabilidades de transição de forma eficiente
  - Uso de uma árvore molde

Como implementar a proposta

- Gerar todas as árvores possíveis
- Resolver o problema do aumento da precisão
  - Guardar apenas os dígitos mais significativos
- Calcular os pesos e as probabilidades de transição de forma eficiente
  - Uso de uma árvore molde

Como implementar a proposta

- Gerar todas as árvores possíveis
- Resolver o problema do aumento da precisão
  - Guardar apenas os dígitos mais significativos
- Calcular os pesos e as probabilidades de transição de forma eficiente
  - Uso de uma árvore molde

Como implementar a proposta

- Gerar todas as árvores possíveis
- Resolver o problema do aumento da precisão
  - Guardar apenas os dígitos mais significativos
- Calcular os pesos e as probabilidades de transição de forma eficiente
  - Uso de uma árvore molde

Como implementar a proposta

- Gerar todas as árvores possíveis
- Resolver o problema do aumento da precisão
  - Guardar apenas os dígitos mais significativos
- Calcular os pesos e as probabilidades de transição de forma eficiente
  - Uso de uma árvore molde

Como definir a penalização *a priori*

- A verossimilhança da amostra aumenta quando aumenta o tamanho da árvore, portanto devemos penalizar as árvores com mais parâmetros.
- Quando aumentamos a amostra o tamanho da árvore estimada aumenta, portanto a penalização deve levar em consideração o tamanho da amostra.
- Ex: $\omega(\tau) = (ct)^{-N}$, onde N é o número de nós terminais da árvore τ e t é o tamanho da amostra.

Como definir a penalização *a priori*

- A verossimilhança da amostra aumenta quando aumenta o tamanho da árvore, portanto devemos penalizar as árvores com mais parâmetros.
- Quando aumentamos a amostra o tamanho da árvore estimada aumenta, portanto a penalização deve levar em consideração o tamanho da amostra.
- Ex: $\omega(\tau) = (ct)^{-N}$, onde N é o número de nós terminais da árvore τ e t é o tamanho da amostra.

Como definir a penalização *a priori*

- A verossimilhança da amostra aumenta quando aumenta o tamanho da árvore, portanto devemos penalizar as árvores com mais parâmetros.
- Quando aumentamos a amostra o tamanho da árvore estimada aumenta, portanto a penalização deve levar em consideração o tamanho da amostra.
- Ex: $\omega(\tau) = (ct)^{-N}$, onde N é o número de nós terminais da árvore τ e t é o tamanho da amostra.

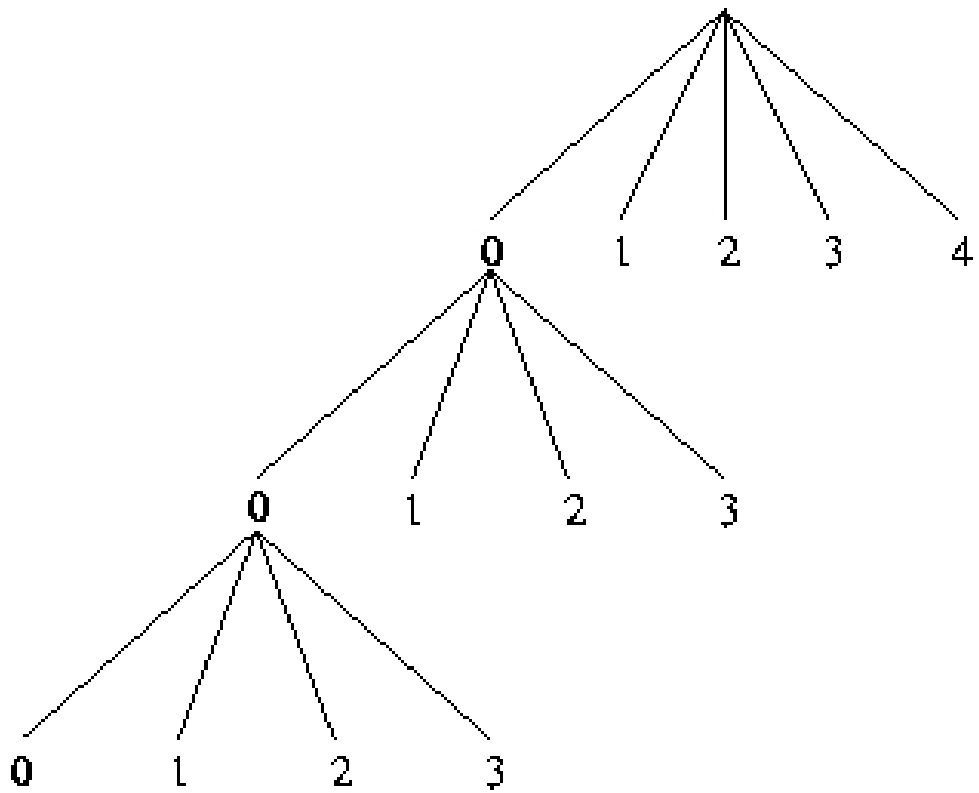
Como definir a penalização *a priori*

- A verossimilhança da amostra aumenta quando aumenta o tamanho da árvore, portanto devemos penalizar as árvores com mais parâmetros.
- Quando aumentamos a amostra o tamanho da árvore estimada aumenta, portanto a penalização deve levar em consideração o tamanho da amostra.
- Ex: $\omega(\tau) = (ct)^{-N}$, onde N é o número de nós terminais da árvore τ e t é o tamanho da amostra.

Resultados para PE

Árvore com probabilidade ≈ 1, para um tamanho de amostra de 148887 símbolos. O peso a priori foi $\omega^0(\tau) = (500t)^{-N}$, onde

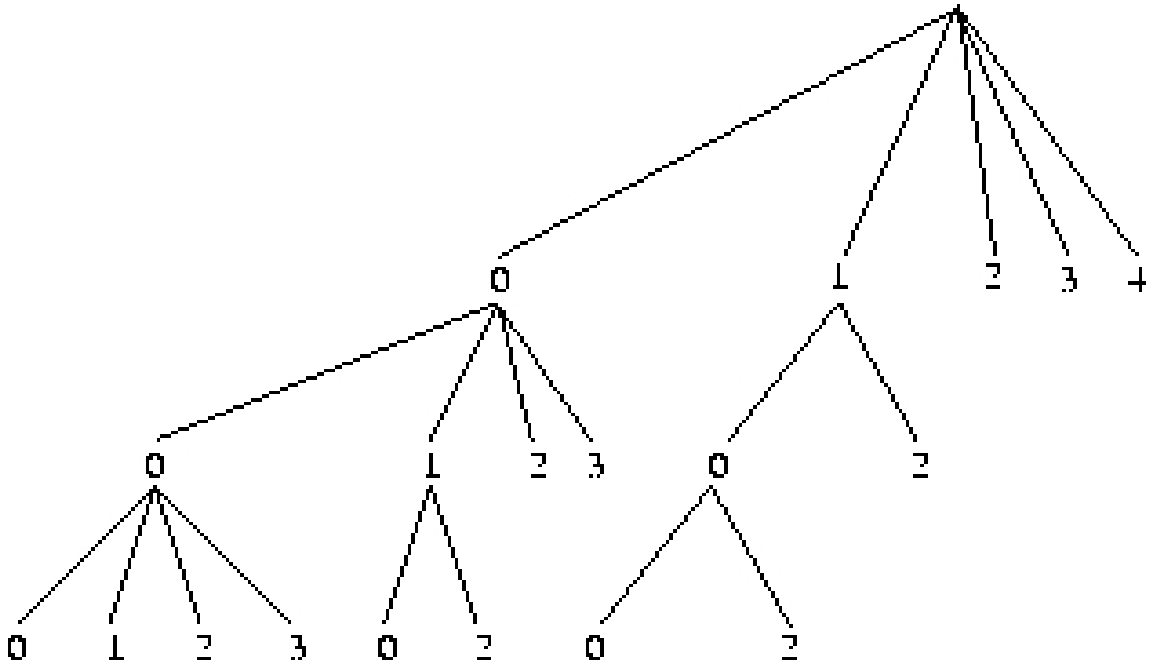
$N = \#$ de nós terminais da árvore τ.



Resultados para PB

Árvore com probabilidade ≈ 0.8, para um tamanho de amostra de 619282 símbolos. O peso a priori foi $\omega^0(\tau) = (500t)^{-N}$, onde

$N = \#$ de nós terminais da árvore τ.



Um fato notável: uma única árvore se sobresai na floresta

- É notável e surprendente que num conjunto com cerca de 6000 árvores o procedimento leve a uma medida de probabilidade concentrada numa única árvore.
- Isso nos leva à sequinte conjectura: se a amostra de textos codificados for gerada por uma única árvore, então o peso dessa árvore tenderá a 1 quando o tamanho da amostra cresce.
- Essa conjectura é verdadeira e é isso que nós demonstramos.

Um fato notável: uma única árvore se sobresai na floresta

- É notável e surprendente que num conjunto com cerca de 6000 árvores o procedimento leve a uma medida de probabilidade concentrada numa única árvore.
- Isso nos leva à sequinte conjectura: se a amostra de textos codificados for gerada por uma única árvore, então o peso dessa árvore tenderá a 1 quando o tamanho da amostra cresce.
- Essa conjectura é verdadeira e é isso que nós demonstramos.

Um fato notável: uma única árvore se sobresai na floresta

- É notável e surprendente que num conjunto com cerca de 6000 árvores o procedimento leve a uma medida de probabilidade concentrada numa única árvore.
- Isso nos leva à sequinte conjectura: se a amostra de textos codificados for gerada por uma única árvore, então o peso dessa árvore tenderá a 1 quando o tamanho da amostra cresce.
- Essa conjectura é verdadeira e é isso que nós demonstramos.

Um fato notável: uma única árvore se sobresai na floresta

- É notável e surprendente que num conjunto com cerca de 6000 árvores o procedimento leve a uma medida de probabilidade concentrada numa única árvore.
- Isso nos leva à sequinte conjectura: se a amostra de textos codificados for gerada por uma única árvore, então o peso dessa árvore tenderá a 1 quando o tamanho da amostra cresce.
- Essa conjectura é verdadeira e é isso que nós demonstramos.

Uma única árvore resume a floresta

Theorem

Let $(\tau, \mathbb{P}_\tau)$ be a probabilistic tree. Then for almost all sample $x_0, x_1, \ldots$ generated by $(\tau, \mathbb{P}_\tau)$, we have that $p^n(\tau) \to 1$ when $n \to \infty$.